中国图书馆学报(季刊)1992 年第 2 期
ZHONGGUO TUSHUGUANXUEBAO

●孙清兰

高频词与低频词的界分及词频估算法

齐夫第二定律揭示了低频词的分布规律,给出:

$$I_n / I_1 = 2 / n(n+1) \cdots\cdots \quad (1)$$

式中，I_n 代表文中出现 n 次的词汇数量。比值与文章长度无关〔1〕。

高频词与低频词分界有个临界值，这是 Dono hue, J · C · 于 1973 年提出的〔2〕。其计算公式为:

$$n = \frac{1}{2}(-1+\sqrt{1+8I_1}) \cdots\cdots \quad (2)$$

可见，公式 (2) 依赖于文中出现一次的词数。

本文也给出以下一条高频词、低频词分界临界值的计算公式:

$$n = \frac{1}{2}(-1+\sqrt{1-4D}) \cdots\cdots \quad (3)$$

公式 (3) 与公式 (2) 的不同点是，(3) 式取决于文章中的不同词数 D，而 D 显然比 I_1 容易得到。进一步，我们可以给出公式 (3) 的以下简化形式:

$$n = \sqrt{D} \cdots\cdots \quad (4)$$

利用这一简化公式界分高频词、低频词，不仅与 Dono hue 公式的效果一致，而且计算简捷，使用方便。

本文通过验证，得到了不同数量同频词的词频估计公式，从而进一步丰富了词频分布规律的内容。

一、高频词与低频词的界分方法

(一) 词的等级的确定。

在研究词频的分布规律时，对词的等级有几种不同的确定方法。我们这里采用的是最大值法，即把文章中的词按出现次数由高至低顺序排列，遇到同频词任意排序。词的等级取作同频词词序的最大值。显然，不伴有同频词的词 (同频词总数为 1)，其等级就是它的词序值。

表 1 以文献〔3〕为依据。其中，文章长度 T=3907，不同词数 D=813。

表 2 以文献〔4〕为依据。其中，总词数 T=6958，不同词数 D=1170。

(二) 高频词与低频词的界分公式。

从表 1、表 2 可见，最后一个词的等级恰好等

(二) 含有概念排除的文献主题检索。

实际检索中，有时读者可能需要查找除某一个或几个方面之外的某主题文献。如检索主题为“除喷雾淬火之外的各类模具的各种淬火”、“除金钢镗床和非标准设计之外的其它各类镗床的各种设计”的有关文献。这种要排除的概念，计算机工作时，是逻辑非运算。进行检索句式标引，必须要用逻辑非符号表征出它的特殊含义。上面两个主题用句式标识为:

1.模具[喷雾]淬火

2. (金钢) 镗床[非标准]设计

填写机检单时，要将排除概念连同它的逻辑非符号一并写入相应词类项。

综上所述，句式法的优点有以下几个方面:

1.以汉语为基础的检索语言形式，表意直观、明确，便于掌握，能准确标识各种错综复杂的文献主题，能进行多途径检索。

2.具有专指性和一定的系统性检索功能。

3.使用语法职能符号，保证了检索的准确性，避免了检索中的“噪音”。

4.提高了主题的存贮和检索效率：概念单元取代了词表，极大地减轻了存、检的工作量；标有语法职能符号的标识单元，有效地节省了计算机运行时间。

5.不仅适于联机检索，而且能实现微机的独立存贮和检索。

(来稿时间：1991.6.编发者：刘喜申。)

表 1 文献〔3〕的词频分布

同频词数	词序	等级	频次	词等级与频次乘积	同频词数	词序	等级	频次	词等级与频次乘积
I_n		r_n	n	$r_n \cdot n$	I_n		r_n	n	$r_n \cdot n$
1	1	1	342	342	1	24	24	27	648
1	2	2	134	268	2	25～26	26	25	650
1	3	3	118	354	3	27～32	32	22	704
1	4	4	117	468	3	33～35	35	19	665
1	5	5	110	550	3	36～38	38	17	646
1	6	6	88	528	3	39～41	41	15	615
1	7～8	8	68	544	3	42～44	44	14	616
1	9	9	54	486	5	45～49	49	13	637
1	10	10	53	530	5	50～54	54	12	648
1	11	11	51	561	5	55～59	59	11	649
1	12	12	48	576	6	60～65	65	10	650
2	13～14	14	47	658	9	66～74	74	9	666
1	15	15	46	690	12	75～86	86	8	888
1	16	16	40	640	14	87～100	100	7	700
1	17	17	37	629	21	101～121	121	6	726
1	18	18	36	648	24	122～155	155	5	775
1	19	19	35	665	35	156～190	190	4	760
1	20	20	31	620	68	191～258	258	3	774
1	21	21	30	630	140	259～398	398	2	796
2	22～23	23	29	667	415	399～813	813	1	813

表 2 文献〔4〕的词频分布

同频词数	词序	等级	频次	等级与频次乘积	同频词数	词序	等级	频次	等级与频次乘积
I_n		r_n	n	$r_n \cdot n$	I_n		r_n	n	$r_n \cdot n$
1	1	1	512	512	2	38～39	39	28	1092
1	2	2	312	624	3	40～42	42	27	1134
1	3	3	199	597	2	43～44	44	26	1144
1	4	4	177	708	3	45～47	47	25	1175
2	5～6	6	157	942	3	48～50	50	23	1150
1	7	7	137	959	3	51～53	53	22	1166
1	8	8	121	968	5	54～58	58	21	1218
1	9	9	120	1080	6	59～64	64	20	1280
1	10	10	114	1140	5	65～69	69	19	1311
1	11	11	105	1155	2	70～71	71	18	1278
1	12	12	104	1248	8	72～79	79	17	1343
1	13	13	87	1131	5	80～84	84	16	1344
1	14	14	84	1176	4	85～88	88	15	1320
1	15	15	75	1125	6	89～94	94	14	1316
2	16～17	17	62	1054	5	95～99	99	13	1287
1	18	18	55	990	6	100～105	105	12	1260
3	19～21	21	45	945	11	106～116	116	11	1276
1	22	22	44	968	10	117～126	126	10	1260
3	23～25	25	43	1075	11	127～137	137	9	1233
3	26～28	28	42	1176	13	138～150	150	8	1200
1	29	29	41	1189	20	151～170	170	7	1190
1	30	30	40	1200	28	171～198	198	6	1188
1	31	31	38	1178	38	199～236	236	5	1180
1	32	32	37	1184	63	237～299	299	4	1196
1	33	33	35	1155	88	300～387	387	3	1161
1	34	34	33	1122	175	388～562	562	2	1124
2	35～36	36	32	1152	608	563～1170	1170	1	1170
1	37	37	29	1073					

于文章的词数 D。而且，除少数高频词外，词汇的齐夫分布特征较为明显：词频与词等级的乘积围绕 D 值上下波动。根据齐夫第一定律：

$$r_n \cdot n = D \cdots\cdots \quad (5)$$

进一步，词频为n的词数为：

$$I_n = r_n - r_{n+1} \cdots\cdots \quad (6)$$

比如，表 1 中 $I_2 = r_2 - r_3 = 398 - 258 = 140$；表 2 中 $I_3 = r_3 - r_4 = 387 - 299 = 88$。

把公式 (5) 代入公式 (6) 式，则：

$$I_n = r_n - r_{n+1}$$
$$= D / n(n=1) \cdots\cdots \quad (7)$$

公式 (7) 可看作词数为 n 的词数量计算公式。当知道一篇文章包括词数 D，利用该式便可计算出 1 次、2 次、…n 次的词各有多少。当 $n=1$ 时，则：

$$I_1 = D / 2 \cdots\cdots \quad (8)$$

公式 (8) 和公式 (7) 相比，即可得到齐夫第二定律表达式。

大量统计研究表明，词频越高，同频词在文章中出现的越少。如果把不伴有同频词的词视为高频词（同频词数为 1），由公式 (7) 可以得到：

1. $D / n(n+1) = 1$

2. $n = \frac{1}{2}(-1 + \sqrt{1 + 4D})$

由该式计算出来的是高频词中的最低频次值，因此可称作高频词与低频词的词频临界值。若把公式 (8) 式代入公式 (7)，则：

$$n = \frac{1}{2}(-1 + \sqrt{1 + 4D})$$
$$= \frac{1}{2}(-1 + \sqrt{1 + 8I_1})$$

这就是 *Dono hue* 给出的公式 (2)。于是，从理论上讲，公式 (7) 与公式 (2) 的计算结果必然是一致的。

为了计算方便，忽略 D 的微小误差，给出公式 (7) 的简化形式 $n = \sqrt{D}$

(三) 界分公式的可靠性检验。

从高频词、低频词界分公式 (3)、简化公式 (4) 和 Dono hue 的公式 (2) 计算表 1、表 2 数据的结果与实际值对照（见表 3）可以看出，公式 (3)、公式 (4) 的计算精确值与公式 (2) 的十分接近。若取整数，结果基本一致。这表明公式 (3) 和公式 (4) 理论上是可靠的。公式的计算值与实际值对照结果又显示了计算公式在实际中的可信性及适用性，这里的实际值是按以下原则确定的：以同频词数连续（一般多于 3 个）出现 1 值的最小词频值定为高频词、低频词分界值。于是表 1 的实际值取作 30，表 2 实际值取作 33。从表 1、表 2 中都可以看出，高频词部分基本体现了同频词个数为 1 的规律。

表 3 文献〔3〕、文献〔4〕的高频词、低频词界分计算值与实际值比较

表	D	I_1	计算值			实际值
			数值类型	用公式 3 计算	用公式 4 计算	用公式 2 计算	
表 1	813	415	精确值	28.018	28.51	28.314	30
			整数值	28	29	28	
表 2	1170	608	精确值	33.71	34.21	34.37	33
			整数值	34	34	34	

二、词频估计方法

如果由公式 (7) $I_n = D / n(n+1)$求解 n，则可得：

$$n = \frac{1}{2}(-1 + \sqrt{1 + 4D / I_n}) \cdots\cdots \quad (9)$$

公式 (9) 即为同频词数等于 I_n 的词的最低频次计算公式。显然公式 (3) 就是公式 (9) 中 I_n 等于 1 的情况。

相应地，数量为 I_n 的同频词词频取值区间为：

$$\frac{1}{2}(-1 + \sqrt{1 + 4D / I_{n+1}}) > n$$
$$\geqslant \frac{1}{2}(-1 + \sqrt{1 + 4D / I_n}) \cdots\cdots \quad (10)$$

因为 n 代表词频，计算值只取整数。

一般地，I_n 值较大时，公式 (10) 的区间值小于 1，词频趋于唯一的整数值，用公式 (9) 确定词频即可。I_n 值较小时，用公式 (10) 确定词频取区间值。

现将表 1、表 2 数据用公式 (9) ($I_n \geqslant 6$) 和公式 (10) ($I_n < 5$) 计算，词频估计值与实测值对照结果列于表 4、表 5。

在表 4 中，同频词数 9～415 的低频词词频计算值与实测值完全一致，同频词数为 6 的词词频估计值与实测值仅相差 1。显而易见，用公式 (9)

估计低频词词频值效果相当好。对于同频词数小于6的词，公式（10）的词频区间估计值与实测区间亦比较接近。特别是高频词、低频词的分界值计算更为准确。

在表5中，同频词数10～608的词频计算值与实测值非常一致，高频词与低频词分界值计算也很准确。只是同频词数4～8这部分中频词词频计算值与实测值吻合情况欠佳。其原因在于表2中词频值12～21这个范围内同频词数分布比较混乱。

表4 文献〔3〕的词频估计值与实测值对照

同频词数(I_n)			415	140	68	35	24	21	14	12	9	6
词频	估计值	计算精确值	0.986	1.961	2.99	4.34	5.34	5.74	7.14	7.75	9.02	11.1
		整数值	1	2	3	4	5	6	7	8	9	11
	实测值		1	2	3	4	5	6	7	8	9	10
同频词数(I_n)			5		4		3		2		1	
词频	估计值	计算精确值	(12.26,13.76)		(13.76,15.96)		(15.96,19.67)		(19.67,28.02)		28.02–	
		整数值	12–13		14–15		16–19		20–28		29–	
	实测值		11–13		–		14–19		25–29		30–	

表5 文献〔4〕的词频估计值与实测值对照

同频词数(I_n)			608	175	88	63	38	28	20	13	11	10	9
词频	估计值	计算精确值	0.975	2.13	3.18	3.84	5.07	5.98	7.16	9	9.82	10.32	10.9
		整数值	1	2	3	4	5	6	7	9	10	10	11
	实测值		1	2	3	4	5	6	7	8	9	10	–
同频词数(I_n)			8	7	6	5		4	3		2		1
词频	估计值	计算精确值	11.6	12.4	13.5	(14.8,16.6)		(16.6,19.25)	(19.25,23.69)		(23.69,33.71)		33.71–
		整数值	12	12	14	15–16		17–19	20–23		24–33		34–
	实测值		17	–	12–14	16,19		15	22–25		26–32		33–

综上，理论分析与实例验证一致表明，公式(9)和公式（10）揭示了词频与同频词数量的内在联系以及它们与文章中不同词个数的依赖关系。文章中词出现的频次、同频词的个数均与不同词数有关，与文章长短无关。这一客观规律的发现，对于语言学和文献学的研究有着广泛的学术意义。

参考文献

〔1〕Booth, A.D.: A Law of Occurrences for Words of Low Frequency, Information and Control, Vol. 10, No. 4, 386–393, 1967

〔2〕Donohue, J.C.: Understanding scientific Literature: A Bibliographic Approach, The MIT press, cambridge, 1973

〔3〕The Impact on Libraries, Special Libraries, Vol. 78, No. 1, Winter, 7–14, 1987

〔4〕王崇德等.汉语文集的齐夫分布.情报科学，1989，10（4）：1～8

（作者单位：东北师大图书情报学系。来稿时间：1991.7.编发者：丘峰。）

An Interpretation of Library Management / Zhao Chengshan / / Bulletin of the Library Science in China / China Society of the Library Science. -1992, 18(2).69～71

The definition of library management can be summed up as follows: It is a kind of activities conducted by the library administrative personnel for the purpose of achieving the optimun goal of a certain activity of library systems, who, by way of several processes of planning, organizing, controlling, etc., optimizes the combination of all the library resources to reach the hight of improving the social benefit of library systems. 6 references.

Theory of systems —— Applications
Library management —— Studies
Library undertaking —— organization and management
G251

Time Management and the Director of the Library / Chen Shu / / Bulletin of the Library Science in China / China Society of the Library Science. -1992, 18(2).71～73

Accompanied with the radical increase of information and wealth of man, the library world is now confronted with the challenge of time. A director of the library will consequently be far behind if he handles his work merely by the virtue of his good intentions and experiences. He has to know well the theory and technique of time management, besides. The chief items of time management are: the 4D law, the exception law, the negative law, etc. as well as those techniques derived from these laws. In short, time management is none other than the knowledge of approaching ways on eliminating waste of time, thereby, it will be able to attain the goal of running a library. So far as the significance of the modern management is concerned, the wing for the Chinese libraries to soar is exactly the time. 2 illus. 6 references.

Library undertaking —— Scientific management
Directors of libraries —— Personal qualities
Time management —— Theories
G251-36

A Preliminary Probe into Subject Indexing Retrieval of Chinese Scientific and Technical Documents / Subject Indexing Research Group, Library of Hebei Mechanical and Electrical College / / Bulletin of the Library Science in China / China Society of the Library Science. -1992, 18(2).74～78

The abbreviation of Chinese sentence-mode subject indexing is the "sentence-mode method" — a new method for Chinese scientific and technical document subject indexing retrieval. The Chinese sentence-mode being a form of the retrieval language, is compatible with some characteristics of subject indexing and classification. The article also makes an approach to a new may for the standardization of mark unit. In each particular subject, there exists objectively a kind of "concept unit" which one may follow to use. The "concept unit" is not like the unit word of the basic word method, nor is it like the subject word derived from a thesaurus artificially standardized. It is an objective, intrinsic concept unit separated out from a particular subject, i.e. a kind of standard subject word of a special form without a thesaurus. The method has already been retrieved and tested by a computer. 1 table.

Subject word method —— Approaches
Sentence-mode method —— Reviews
G254.2

The Demarcation of High - and Low - frcquency Terms and Ways of Estimating the Frequency of Terms / Sun Qinglan / / Bulletin of the Library Science in China / China Society of the Library Science. -

1992,18(2).78～81

Mr. J.C. Donohue, in 1973, believe that there was a critical value in the demarcation of high – and low – frequency terms used in articles, and at the same time he proposed formulas of calculation. On the basis of adopting maximum value to determine grades of terms, this article brings forth and introduces a new formula of demarcating the high – and low–frequency terms which is not only simple and direct in the process of calculation but also more practical as compared with that of Mr. Donohue. Having made a study of ways of calculating for various kinds of terms with similar frequency, the article also gives a formula of calculation and reveals the inherent law governing the number of frequency of terms and terms with similar frequency. The author holds that the frequency of terms appeared in the article and the number of terms with similar frequency have something to do with the number of different terms but have nothing to do with the length of the articles. 10 formulas. 5 tables. 4 references

Zepf's Law —— Studies
Frequency of terms —— Calculations
Document metrology —— Theories
G256

Basis, Adjustment and Control of Developing Library Work of Town and Township Libraries / Yan nan / / Bulletin of the Library Science in China / China Society of the Library Science. –1992,18(2).82～84

The author has put forward an mathematical formula for defining the basis of developing library work of town and township libraries. The formula shows that the cultural demand coefficient of towns and townships are determined by the annual social total output, the per capita, per annum income, the total number of population, the rate of non–illiteracy, the land area and the amendment value. In order to show an object scene of the cultural demand of various towns and townships and the conditions of library work coordination, the author brings forward the cultural demand coefficient and the distribution illustration of the sample variable probability of the library holdings, thus the cultural demand curve is comparable to that of the existing holdings. 1 illus.

Public library work —— Organizations and managements
Town and township libraries —— Studies of development
G259.252.3

A New Probe Made by Heilongjiang Provincial Library to Serve the Economic Construction / Yang Xuemei, Wang Aijuan, and Liu Lingzhi / / Bulletin of the Library Science in China / China Society of the Library Science. –1992,18(2).85～86,90

The Heilongjiang Provincial Library has not only probed into a new way of developing information resources and serving the four modernization to provide information service for the leading bodies of the provincial party committee and provincial government to give a proof of making a macro economic decision, but also has led all the municipal and county public libraries in the province to make a joint effort to run the "Spark information newsflash". Thus it not only plays the key role of a provincial library, but also strengthens library co–operation and makes contributions to the development of the middle–sized and small–sized enterprises and the rural economy.

Provincial public libraries —— Heilongjiang Province
Reader services —— Effects
G252

(频萍译校)