词频分析

邓珞华

词的频率，近些年来在国内外一些领域中逐渐引起人们的重视。在文字学、机器翻译、汉字编码中固然少不了词频的统计，在情报学和图书馆学中对它的研究也逐渐兴起，大量的词频统计数据在情报检索、情报基础理论、书目计量学、情报分析研究、类表词表编制的文章中出现，引起人们瞩目。

词是具有单独意义、单独作用的字或字母的组合，但在情报学、尤其是情报检索中，常常把"term"(名词、术语)译成"词"，这是因为情报专家对名词术语的兴趣远远大于对其它词类的兴趣，只有名词术语才最能代表文献的内容。在很多外文文献中，term 被赋于更广泛的意义，美国著名情报检索专家 G·索顿曾用 term 来代表主题词、名词、标引词、情报项，甚至用来代表分类号，因为分类号实际上是类名名词的一种符号表达形式。①在不同的场合下，term 几乎可以表示一切文献著录、标引、检索的信息单元。本文所说的"词"，正是这种广义上的 term，具体来讲，是指主题词、关键词、叙词、著者、分类号、篇名刊名、汉字、字母等，本文所说的词频分析，也就是对以上这些"词"在标题中、文章中、目录中、文献库中的出现频率进行的分析。

本文着重从以下两方面总结已有的词的统计规律及其实用价值。

一、词频的一般特性

在不同的环境中，比如说在一篇文献中，在一个数据库中，在不同取法的标目中，各类词（主题词、著者、类号等）的分布频率是不同的，但在很多情况下，它们都呈现出规律性的分布。

1．单词在文章中的分布服从双曲线分布。

1948年，K·齐夫根据他的最小努力原则的指导，统计了各种文体的文章，发现单词在任何一篇文章中的出现频率都服从一个规律，即如果把每个词在一篇文章中的出现频率按照递减的顺序排列起来，并用自然数从小到大给词频的倒排序予等级，频率最高的词的等级为 1，频率次之的等级为 2，……频率最小的词等级为 N（N为单词个数），若R表示等级，F 表示频率，C为常数，那么就有 $R \times F = C$。齐夫揭示的这种词频规律，后来被人们称为齐夫定律。

学术界的进一步研究表明，齐夫定律的适用范围有一定的局限性，尤其是频率特别高的词(如英文中的 the)和频率特别低的词，不能完全满足 $R \times F = C$的关系式。②1952年，B·芒代尔布罗运用信息论原理，重新解释了齐夫定律并得到齐夫定律的修正式：

$$f = C(r-a)^{-s}$$

式中 f 是词的频率，r 是词的等级序号，a 和 s 是修正齐夫定律的两个常数。a 的值很小，作用是使低频率词与公式相符，s 的值接近 1，目的是使高频率词与公式相符。③

汉语中，词在文章中的频率分布也有类似情况，但还没有足够的统计资料作依据，这是因为汉语文章中词与词之间没有间隔，计算机识别困难，因而统计也困难；而手工统计工作量太大，使这一统计难以详尽。但汉字在文章中的分布频率是有较可靠的数据的。前几年，武汉大学语言自动处理研究组曾对《骆驼祥子》一书的字频进行过详细的统计，全书总字数107360个，由2413个不同单字组成，"的"字的频率最高，占

4.1198%；“组”字频率最低，占0.0009%，“的”的C值是1×4.1198%=0.041198，“组”的C值是2413×0.0009%=0.021717，但中间的约2000个字的C值都在0.01左右，这个结果说明汉字在文章中的分布是符合齐夫定律的。

很多统计表明，虽然C值随着统计样本的不同而异，但只要样本足够大，总是围绕着0.1上下波动。最早的一个较全面的文章词频统计表是由杜威（Dewey）编成的，他对一篇有10万个单词的文章进行了词频统计，得到的C值如下（每个数据下的脚标为相应词的级别序号）：

0.73_1，0.80_2，0.98_3，0.117_4，0.106_5，0.116_{10}，

0.124_{50}，0.111_{100}，0.102_{200}，0.100_{400}，0.108_{600}④

库塞拉（Kucera）和弗朗西斯（Francis）统计了一个由不同文章组成的总长为100万个单词的样本以及1965年化学题录（Chemical Title）磁带、第58期MARC磁带以及气相层析数据库的词频，统计数据证明，出现在足够大的英文样本中的词，大至可以满足关系式R×F=C，而且C值较固定（0.1左右）；而在较小的样本中却不尽然；对于高频词和低频词，则要应用芒代尔布罗修正式加以修正。⑤

S·希普斯（S.Heaps）证明，当词的个数N等于500，10000，50000和100000时，C值分别为0.11，0.10，0.09和0.08，C的平均值在0.1左右。⑥

C值不仅随样本的长度不同而略有差异，而且随词性的不同也呈现较大的波动，希普斯统计MARC磁带58期上的60778个主题词和53387个出版者名称的C值，发现主题词的C值和出版者的C值差距很大，前者在0.038到0.1之间波动，后者在0.53和0.254之间波动；他还对气相层析磁带上的化合物名称、物质名称、混合叙词短语进行了统计，发现他们的C值也有较大差异。

2．标引词在文献库中的分布服从正态分布。

文章被标引后，每篇文章被形式化，成为几个主题标引词的组合形式，这样我们研究单篇文献中词频的分布已没有什么意义，只有从整个文献库的角度研究标引词的频率才有实际意义——例如可以估计索引倒排档的容量以统筹安排计算机存贮空间。好在这种标引词的频率很容易统计，我们只要统计文献库后的主题索引，每个主题词下的篇数即是该词的频数。

很多统计数据都表明，标引词在文献库中的分布是服从正态分布的。G·索顿指出：“研究许多文献的存取手段，如对指南、索引、目录、题录的研究表明，在每种情况下，数量的分布（就题录中的词数量、书后索引的词条数量而言）都服从正态分布式：

$$1(x)=\frac{1}{xs\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\frac{\log x-m}{S}\right]^2\right\}$$

这里m是平均数，S是分布的标准离差。让m和S趋于无穷大，则$\frac{m}{S^2}$趋近一个参数k，那么对数分布就能转换成齐夫定律的式子。”⑦

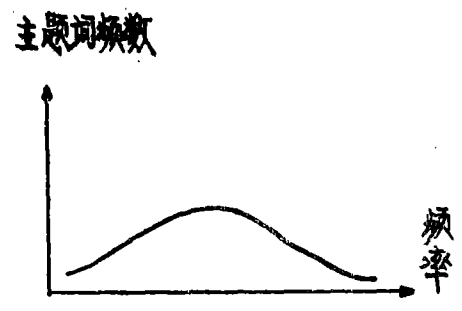

图 1

我们用计算机对1983年INSPEC磁带上的自由词项进行了统计，结果如图1，图象表明自由词在文献库中的分布是正态的；我们还对部分美国《生物文摘》（BA）概念组配索引，美国《合金索引》进行了手工的统计，统计结果也支持了上面的结论。

我们要强调的是，正态分布是图书馆工作和情报工作中各种随机现象中的最为广泛的频率分布，不光是主题词、自由词、标题词在文献库中的频率分布可以用它来描述，用户的查寻行为（或说读者到馆率、借阅频率）可以用它来描述，⑧我们在工作中还发现每类书的购进频率，甚至著者和分类号在文献库中的频率也经常呈现正态分布（目前还不敢断然下“总是呈正态分布”的结论）。这种现象决不是偶然的。我们知道，具有正态分布的随机事件往往具有这样的特性：它们都可以看作许多微小的、独立的随机因素作用的总结果，而每一个因素的影响都很小。根据概率论的中心极限定理，可知具有

这种特点的随机变数一般都可以认为近似地服从正态分布，这正是正态分布在理论上和实践上都极其重要的原因。图书情报工作中的许多随机现象，例如读者到馆、主题词的选取、著者个人文章的入选等，都具有以上特点，这使得正态分布绝非偶然地大量出现在图书情报领域中。

3．分类号在大型文献库中的频率分布符合双曲线分布。

由于分类号是类名词的代号，所以分类号的频率分布实际上也是一种词频分布，即类名的频率分布。笔者曾调查了武汉地区几所大型图书馆的分类目录和历年进书的分类统计，发现无论是在按《中图法》进行分类的图书馆还是在按《科图法》进行分类的图书馆中，各大类号在文献库中的分布频率都大至符合双曲线分布，其中武汉大学图书馆理科教师参考室尤其明显。最引人注目的是，统计了79—84年每年的进书后，发现除了个别类号的等级序号略有变动外，整个频率曲线的走向也大致符合双曲线的走向。类似走向的数据在近几年的图书情报刊物中已多次出现过。

是由于齐夫的"最小努力原则"在起作用，还是由于别的什么原因，使得双曲线分布在词频分布中不时地显示出来，甚至大量出现的正态分布的词频也转换成了双曲线分布，这一定有某种原因和某种机制在起作用。有兴趣的同志若能采用一种好的分析方法进行更为深入细致的研究，可能会发现这种机制，这种发现可能会对情报学产生比较重大的影响。

4．文献库中，词汇的增长速度越来越小。

在文献库建库初期，随着每一篇文献的增加，几乎都有新的主题词出现，主题词的增加速度大大高于文献篇数的增加速度；随着文献的大量增长，后面文献使用的主题词重复前面文献使用的主题词的机会越来越大，主题词的增加速度就会越来越小，最终要低于文献篇数的增长速度；当主题词表的所有主题词都至少在文献库中出现了一次时，主题词的增加速度为0，这时文献的增加不会带来主题词的增加，除非主题词表再收进新词。

有人注意到，当一个专业文献库的文献达到3000篇左右时，标引用的主题词的增加Δt_1，将明显小于文献的增加Δt_2，且$\lim_{i\to\infty}\frac{\Delta t_1}{\Delta t_2}\approx 0$，$i$是文献篇数。如果这个结论带有普遍性，那就告诉我们：当存入的文献超过3000篇时，文献倒排档（或说主题索引）需要着重考虑的是每个主题词标目下容纳文献顺序号的空间，而不是标目增加所需要的空间，也就是说要着重考虑主题词在文献库中出现的频率，对频率高的主题词在相应标目下留下较大的空间，对频率低的主题词则正好相反。

在一篇文章中，词汇随着文章长度的增加，其增加速度也逐渐减小，这其中的道理和前面一样，即词表的词总是有限的，而文献的长度却可以无止境地增长。H·S·希普斯认为，以一篇含2000个单词的文本为起点，当文本长度增加时，出现的不同词的数目D与文本的长度（总单词数）N的关系满足下式：

$D=KN^{\beta}$ 或 $\log D=\beta\log N+\log K$

所以logD是logN的线性函数。⑨这里K和β是视具体文本而定的常数。在化学题录（CT）磁带中，K=53，β=0.5；在MARC磁带中，若D和N是标题用词的话，K=15.5，β=0.6；若D和N是主题词的话，K=9，β=0.66。

词汇的速度增长规律主要用于计算倒排文档的体积以合理安排计算机存贮空间。

5．著者在文献库中的分布服从"洛特卡定律"。

洛特卡定律也称"倒数平方定律"，它是文献计量学的重要定律之一。早在1926年，美国人J·洛特卡（J·Lotka）发表了《科技生产率的频率分布》一文，论述了化学与物理学领域中作者数量与其论文数量的关系。⑩

洛特卡分析了他的统计数据，提出：写了x篇论文的作者频率与其所写的论文之间存在着以下关系：

$X^nY=C$

式中C为常数。在CA统计分析中，C=0.5669；在《物理学史一览表》统计分析中，C=0.6079，在洛特卡的数据中，n=2，这就是"倒数平方"的由来。但在其它领域中，n不一定等于2。在情报学领域中，n=3.5，这时上式虽成立，但不能称之为"倒数平方"定律了。

洛特卡之后，许多学者继续研究作者分布规律问题，如Dresden对美国数学家进行了研究，Huber对

加拿大数学家进行了研究，Leavens对计量经济学家进行了研究；以后美国人Murphy又证明了洛特卡定律在人文科学中也适用；紧接着，Schorr把这个定律应用于地图图书馆工作中，也获得一定的成功；1979年，加拿大的学者又尝试着把洛特卡定律引进应用科学和工程领域，分析了计算机科学的文献，分析的数据与洛特卡定律有一定偏差，他们认为这正好说明了理论科学与技术科学的不同。⑪

需要说明的是，洛特卡定律是一条统计定律，随着时代条件的不同，统计的特征值可能会产生一些变化，例如按洛特卡定律，写一篇文章的作者应占发表文章作者总数的60%，而据我们对我国这些年的图书情报刊物的粗略统计，作品在二篇以上的"多产作者"远不止40%。国外有些人在探索修正定律的可能性，不过未见有什么大的进展。

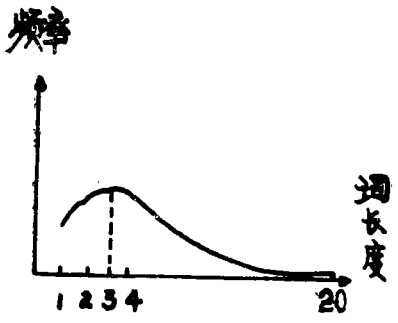

图 2

6。词长度的分布大致服从泊松分布。

词长度的分布由于样本条件的不同，其分布有所不同，但总的来说，它们都趋于泊松分布。当样本的随机性越大时，这种分布就愈明显；当样本的随机性减少，即条件概率的成分增大时，这种分布就不太明显。

库塞拉和弗朗西斯通过大量统计证明，在一般的英文文章中，词长度的平均值约为4.7个字母。他们在统计了一个有1000000个词的文章中的不同长度词的频率后，得出以下频率分布图（图2），这只是标准泊松分布图象的一部分。⑫短词的频率与长词的频率不相对称的主要原因是文章中许多短词如a、the大量重复的使用，语法的局限，人们用词习惯等因素使得"文章"这样的样本中条件概率的成分大大增加，从而使得词长度的分布偏离标准的泊松分布。

当条件概率的成分大大减少时，如在较少顾及语法和人们用词习惯的标题和主题索引中，统计标题中的主题词和索引中用作标目的主题词，词长度的分布就比较明显地趋于"泊松分布。"库塞拉和弗朗西斯统计了CT63—65、MARC01—58、CAIN、CANDICT这些数据库里各种长度标题词的出现频率，其图象（图3）就趋于比较标准的泊松分布，这是因为只取标题中的词，而这些词的出现受作者习惯、语法限制等条件的局限相对来说较少；当我们使词的出现受外在条件的影响最小时，例如当我们只取标题中的主题词而舍弃冠词、形容词、代词、副词时，各种长度的词的出现频率就会趋于更加标准的泊松分布，即图象的对称性更强。

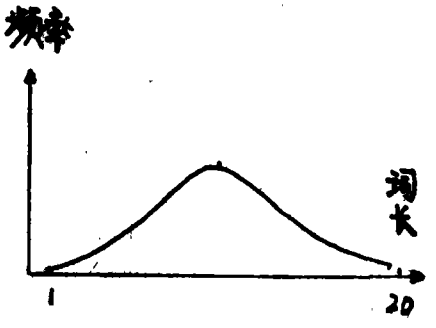

图 3

当词存贮在词典里时，每个词只出现一次，不象文章中大量短词重复出现，因此平均长度增加，为8.1个字符。⑬

泊松分布常常被应用于稀有事件（即每次试验中事件出现的概率很小）在贝努里试验次数很大时发生的频数分布，⑭主题词频率、词长度的频率很大程度上符合泊松分布的概率模型：第一，它们的出现只有"出现"和"不出现"两种结果；第二，它们是稀有事件，每个词的出现概率尽管有差别，但都很小；第三，试验的次数很大，统计样本通常是几万、十几万甚至上百万的词。唯一不符泊松分布条件的，是词的出现不是独立事件，而是受条件影响，即它们不是"贝努里"型的，所以一旦词的出现受外界条件的影响越少，泊松分布的趋向就越明显。

7。字符的频率分布呈"直线分布"。

字符是组成字、词的最基本的单位，如字母、各种笔划笔形等。一个基本字符集中的各种字符的出现频率由于它们出现的场合不同是有很大差异的，例如英文字母"S"在牛津词典中的出现频率是13.9%，居所有英文字母之首；而在杜威选择的一个包括100000个词的文章中，S的出现频率只有6.3%，居第六位。H·S·希普斯在《情报检索计算与理论引论》一书中，列表对比了五种情况，即杜威选择的英文通俗文章、牛津词典、加拿大词典、科学美国人、杜威选的一篇长达100000个词的文章中英文字母的频率。除了在两种词典中

各字母有相近的频率外，其余情况下各字母出现的频率均有较大差异，但我们按照齐夫定律中的等级排序方法画出各字母的频率分布图象，会发现以上五种情况下的字符频率分布图象都近似一条直线。

笔者通过一些统计，认为汉语拼音字母的声母也有同样的分布规律。同英文字母一样：1）汉语拼音各字母在不同的场合下的出现频率是不一样的，例如统计字典中的声母，会发现Z、J、Y、L、S等字母的频率较高，但一般都未超过10%；在图书馆目录的标目中统计声母，会发现L、D、Z、S的频率较高，最大值也不超过10%，在文章中由于加进了元音字母，字母的分布频率更不同一些（例如A的频率大大增加，频率最大值下降）。2）尽管有以上不同，但同英文字母一样，当我们按等级排序方法画出各汉语拼音字母图象，会发现各种情况下字母频率的分布图象也都近似一条直线，下面是我们粗略统计新华字典中23个汉语拼音声母的出现频率（括号中是频率值）：

A(0.9%)	B(5.1%)	C(5.7%)	D(5.7%)
E(0.6%)	F(3.4%)	G(5.2%)	H(4.9%)
J(7.9%)	K(2.7%)	L(6.9%)	M(4.1%)
N(2%)	O(0.2%)	P(3.2%)	Q(4.4%)
R(1.4%)	S(6.8%)	T(4.5%)	W(2.6%)
X(5.4%)	Y(7.6%)	Z(8.9%)	

它的等级排列频率分布图象近似一条直线（图4）。

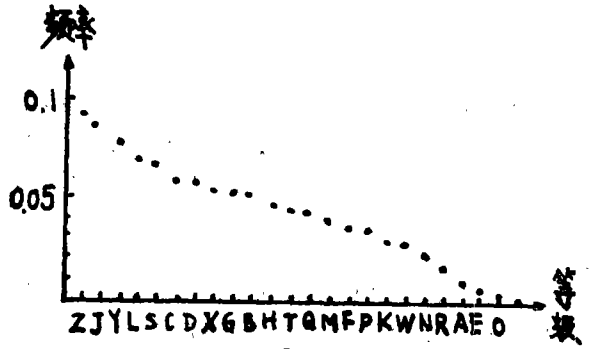

图 4

研究字符频率的作用我们将在后面谈到。

以上我们谈了词频的七个方面的特性。近些年来由于计算机界注重于研究汉字信息处理，出现了大量的汉字频率统计数据，如各个汉字的频率，各种偏旁部首的频率，各种基本笔划的频率，……由于统计这些频率的原理和分析的方法与前面所述并无新鲜之处，更由于笔者尚未来得及对这些数据进行分析归纳以总结出规律性的东西，因此不在此介绍了。

二、词频分析的作用

一些对情报学、图书馆学、计算机文字处理中的一些进展不太感兴趣的人，一些对此虽感兴趣但缺乏数理知识的人，对于词频的统计分析于情报学、图书馆学有何作用往往茫然无知。我想用这样一句话来说明词频分析的作用不知是否恰当：如果有一件工作在情报图书工作和情报学图书馆学理论研究中需要付出最艰巨的劳动，需要最深入细致、一丝不苟的工作态度，需要收集最广泛领域的素材，需要利用最现代化的统计工具，而又能对图书情报工作定量化和情报学图书馆学理论研究的定量化起巨大的推动作用，那么这项工作就是词频的统计和分析。

词频分析对推动情报学图书馆学从定性分析的低级阶段向定量分析的高级阶段发展具有举足轻重的作用。对词频的统计以及词之间关系的研究是情报学图书馆学理论研究的核心问题之一。文字信息处理和编码、检索语言和标引语言、类表词表的编制原则和方法、情报用户查寻行为模型这些领域中的理论研究固

然少不了词频的统计分析，但最能说明问题的是情报学的五大定律，即齐夫定律、布拉福德定律、洛特卡定律、文献增长定律和文献老化定律中有两个是以"词"为单位进行统计的，还有两个与词的统计也有关，完全以文献为单位进行统计的只有文献增长定律。

词频分析的作用还在于它作为一种崭新的情报分析方法使得传统的情报分析从原则到方法都起了微妙的变化。词的频率和一定的社会现象、自然现象、情报现象是有紧密联系的，这是用词频进行情报分析的理论依据。有人利用词频统计分析来预测学科和行业的发展，预测科学奖金的候选人，评论一篇文献或一种杂志的优劣，测定核心期刊和核心作者，甚至进行古文真伪的鉴定。……很多预测结果与里手行家的结论是一致的，1983届和1985届的武汉大学情报检索方向的研究生毕业设计中就有二项是为检索系统配备情报辅助分析软件的，而这些分析的基本方法就是词频的统计。1985年的《情报科学》杂志上有人撰文怀疑情报学五大定律的实用性，我认为这是缺乏深刻分析，缺乏远见的，词频统计在情报分析中的应用在今后的发展中将会证明这种论调是错误的。

词频分析也是图书情报工作自动化的基础工作之一。研究文献的自动标引和自动分类的最新手段，就是按一定原则从文献中读出一些词并计算出它们的频率，通过机器的自动甄别从而确定文献是属于哪一类或者应该用哪些词标引。在马隆（Maron）和霍格尔(Hogle)用这种方法进行的自动分类实验中，准确率分别达到95%和87%。⑮索顿在著名的SMART系统中，根据它的一种建立于词频基础上的自动标引理论——判别值加权标引理论，对三个文库CRAN、MED、TIME进行了抽样检索，分析的结果表明，判别值加权标引的使用在检全率不变的情况下，大大提高了检准率。⑯此外，我们还可以根据词的频率来推算机器检索的时间长短以及估计倒排文档的容量。

下面我们举例谈谈词频分析的应用。

1. 利用词频进行情报分析。"书目信息的统计规律与社会现象和情报现象之间有着内在的联系，一定的社会现象和情报现象要引起一定的书目信息现象"，"而书目信息的数量变化往往反映社会现象和情报现象的动态变化"，⑰而产生这种现象的原因，如米哈依诺夫所说，是由于"人类社会具有对科学情报进行评价的某种机制"。⑱

1980年，日本小森隆作了一个极其有趣的统计，他统计了1978年4月至1979年12月的日本《科技文献速报》中主要塑料名词的出现次数。分析了这些统计数据后，小森隆得出结论：热塑性塑料仍占塑料产品的主要地位；聚氯乙烯仍不失"热门"；聚烯烃、丙烯树脂、聚苯乙烯仍是令人瞩目的产品。⑲令人惊奇的是，他的结论与日本的化工专家的预测结果是一致的。

1983年，武汉大学情报检索专业研究生汪建平对1979—1981年的英国INSPEC磁带上情报学的三个主题词进行了统计，列出了与这三个词关系最大的外文期刊。结果发现：三个词出现频率最高的10种外文期刊，恰恰就是我国情报机构和图书馆订户最多的外文期刊。我曾请教过一些学科的专家，没人对"以刊登某主题文献最多的期刊作为该主题领域的核心期刊"这一点表示怀疑。因此，我们只要每隔几年用计算机对大型数据库中的词频进行一次分类统计，就可大致确定各主题领域的核心期刊，这无疑是期刊采购人员翘首以望的。

给情报分析增添一件新武器，这可以说是词频分析的一个功劳吧！

2. 利用词频论证科技劳动集体性的加强。揭示科学家劳动的某些规律，在西方一直是理论界的热门课题，洛特卡的登载倒数平方定律的《科学生产率的频率分布》一文论述的就是科学家劳动成果的分布问题。由词频（作者频率）统计得出的洛特卡定律可以用于证明科学发展的集团化趋势，在情报界已属公认。

证明科学劳动的集团化趋势，还可以通过统计历年来一篇文章的平均著者数量来证明。D·普顿斯对《化学文摘》进行的统计表明：在1910年，80%的化学论文是由一个作者写的；1963年，单个作者发表的论文数只占总数的32%，而每篇文章的平均作者数是2.15。数学方面也有类似趋势。他将历年统计数据绘成图象，发现每篇文献的平均作者数在迅速增加，从而推断：科学劳动的集体性正在加强。⑳

用书目信息的定量分析证明社会学的规律，这恐怕是词频分析的另一功劳吧。

3. 利用词频合理安排码号的长度。图书情报工作中，常常会碰到号码和符号的长度难以确定的问题：分类表中，类号既不能太短（类目不能太大）也不能无限细分下去，长（短）到什么程度最合适？目录的标目

太短了会产生太多重复标目，太长了又可能造成不必要的浪费，应该取多长可以使绝大多数标目不重复？用于图书情报机构的计算机汉字编码方案中，每个字、每个部首、每个汉语拼音字母……是否要安排相同长度的代码？……这些问题，都涉及到词频的统计分析。

大家知道，图书分类法中对于类号的使用，一般都给使用者一定的机动性：一个图书馆可以根据自己馆藏的情况，将分类法中的某些类目的细分程度扩大或缩小，而一个类目是否需要继续细分主要是根据这个类目下的书的多少来决定。我们只要统计每个分类号在目录中的出现频率，对使用频率高于某一经验值的类号可以继续将其细分，对使用频率低于这一经验值的类号可以合并到它的上位类中。这样，分类表既可达到类分图书的目的，又不至于浪费类号以致加重编目人员负担和浪费用户的查找时间。

笔者在《关于书名目录的一个命题的数学证明》(载于《高校图书馆工作》1981年3期) 一文中，用统计汉语拼音声母的频率的方法证明了这样一个命题：藏书100万种左右的图书馆在按书名的汉语拼音字头组织书名目录时，只要收上五个拼音字头，即可基本上把各种书分开。这也是利用词频(字母频率)安排号码长度的一个很好的实例。

4．利用词频估计倒排档容量，推算检索时间。当我们已有了一台计算机，要建立一个文献库；或者已确定了一个文献库，要选择恰当的计算机时，文献库容量的推算是十分重要的。数据库中正文的体积往往是事先固定的或容易推算的，而倒排档往往要在机内建立，所以文献库容量的推算往往归结到倒排档容量的推算。前面我们曾提到，当文献不超过3000篇时，着重要考虑迅速增加的主题词标目的位置，而后却要着重考虑每个标目下文献顺序号的位置。前面我们还谈到，标引词在文献库中的频率分布服从正态分布。在倒排档(主题索引)中，标引词的个数就是倒排档检索标目的个数，每个标引词在文献库中的频率正好就是相应标目下文献顺序号的个数，因此，只要知道文献库中标引词的个数，按照正态分布公式进行积分计算，就可得出文献库倒排档中所有文献顺序号的总数n，n乘上每个顺序号占的存贮空间单元数，即是该倒排档占的存贮容量。

前面我们还提及，在机内主题词典里，由于舍弃了冠词、形容词、代词、副词，各种长度的词的出现频率会趋向更加标准的泊松分布，而平均长度则有所增加，达8.1个字符。假如已知比较一个字母的时间为t，那么每查到一个词之前平均要花多少时间来比较呢？

H·S·希普斯举出了一个实例来计算这个答案：根据他的计算，在CAIN数据库里，平均要经过8次比较，即平均经过8t的时间，才能查到主题表里的词。㉑

5．根据词频挑选标引词和索引词。自动选词一直是计算机标引和计算机检索中的重要课题，标引词和索引词在外文中都是index term，这是因为标引词虽用于标引，索引词虽用于主题索引，但二者往往是一致的，例如一个主题词既可用于标引又可用于检索时的选词。

根据词频自动选择文章中的词进行标引，虽然目前仍有一些问题，但迄今为止，还没发现有更好的办法来代替它。人们通过统计分析发现，频率太高的词和频率太低的词都不宜抽出来作为标引词，㉒频率太高的词多是虚词，而频率太低的词一般不能反映文章的主题内容，所以都不适于标引，自动标引时首先要排除这类词。

计算机检索用的主题词表中正式主题词的入选标准往往也是词的频率。出现次数太少的词，往往代表刚刚兴起的学科和趋于成熟无所进展的学科内容，其文献不多，检索价值不大，全部入选会无谓地增大主题词表的体积，影响词表的使用．所以很多文献库的主题词表都只收录出现频率达到一定数量的词，如英国INSPEC数据库的词表声称它的收词标准就是词的出现频率。㉓有些图书分类法列出的类目达到哪一级往往也是根据该类书的频率来决定的。

6．利用词频进行相关性加权。文献的自动加权标引一般的依据都是标引词在文献中的某种特征值，将这种特征值按某种公式加以计算就得出该标引词的加权值，而这种特征值在很多实验中，都是词的频率。一般来讲，加权标引的检索效果都要优于简单词匹配的检索效果。㉔

七十年代后期，英国罗伯特逊(Robertson)和斯帕克·约恩斯(Karen Sparck Jones)进行了一次有趣的实验：在一、二篇相关文献中取出相关信息作为文献库标引词的加权值，看看这种加权方法是否能改善检索效率。实验结果证明，即使用这样小的相关信息进行加权也能改善检索效果。他们在实验报告中列出了14项

实验数据，其中有3项就是词频统计数据：每个提问的提问词平均数；文献中的词干数；每篇文献的平均词干数。在他们的加权公式中有4项参数，其中就有2项是词频。㉕

在我们见到的文献中，自动加权的计算或是直接地，或是间接地，几乎全都与词频有关，词频在自动加权标引中的地位可想而知。

词频应用的例子还有很多，有些为大家所熟知。如用于汉字信息处理和编码；有的则牵涉到比较复杂的数学推导，如布什(Booth)由齐夫表达式 $\sum_{r=1}^{D} P_r = 1$ 推导出的一个公式 $I_n/I_1 = 2/[n(n+1)]$ 可用于预测一般英文文章中低频率词出现的频率，㉘在这里都不一一介绍了。

词频的统计与分析，作为一种方法，早已应用于情报学和图书馆学，但是很深刻很普遍地认识到它对于情报学图书馆学的意义，很清楚地看到它解决困难问题的巨大潜力，很熟练地掌握它的运用方法，还需要经过一番艰苦的努力。作为情报学中最有前途的学科，有人推崇文献计量学及引文分析；作为情报学中最有前途的分析方法，我极力推崇词频分析方法，因为它涉及了科学交流中知识及其传播者的最基本单元——主题词、类名类号、著者、字母等等，它使用的方法是一门学科发展到中、高级阶段时必须使用的定量分析方法；它依据的原理是社会现象、情报现象与词频现象之间有某种内在的联系。词频分析的种种长处，使我们不得不对它推动情报学图书馆学向高级阶段发展的潜力寄于莫大希望。

参 考 文 献

①⑦⑯ G·Salton：《Mathematis and information retrieval》，载《Journal of documentation，1979年第1期，第1、7、10页。

②㉒ 王洵：《最小努力原则与齐夫定律》，载《情报科学》，1981年第2期，第33、32页。

③ 邓珞华等：《图书情报数学》，东北师范大学，1983年，第190页。

④⑤⑥⑨⑫⑬⑮㉑㉖ H. S. Heaps：《情报检索计算与理论引论》，武汉大学图书情报学院，1982年，第131、135、135、148、141、142、224、143、136页。

⑧ 卢绍君：《用户情报查寻行为的数学模型》，载《情报科学》，1985年第1期，第23页。

⑩⑪ 王洵：《洛特卡定律》，载《情报科学》，1981年第6期，第74、77—78页。

⑭ 中山大学数学力学系：《概率论及数理统计》，人民教育出版社，1980年，第85页。

⑰ 邓珞华等：《书目信息统计与情报分析》，载《情报科学》，1984年第2期，第10页。

⑱⑳ 米哈依诺夫：《科学交流与情报学》，科技文献出版社，1980年，第421、10页。

⑲ 王洵：《计量文献与预测情报》，载《情报科学》，1980年第3期。

㉓ The Institution of Electrical Engineers：《INSPEC Thesaurus》，The Gresham Press，第1页。

㉔㉕ K. S. Jones：《Search term relevance weighting given little relevance information》，载《J. Doc》，1979年第1期，第31、33页。