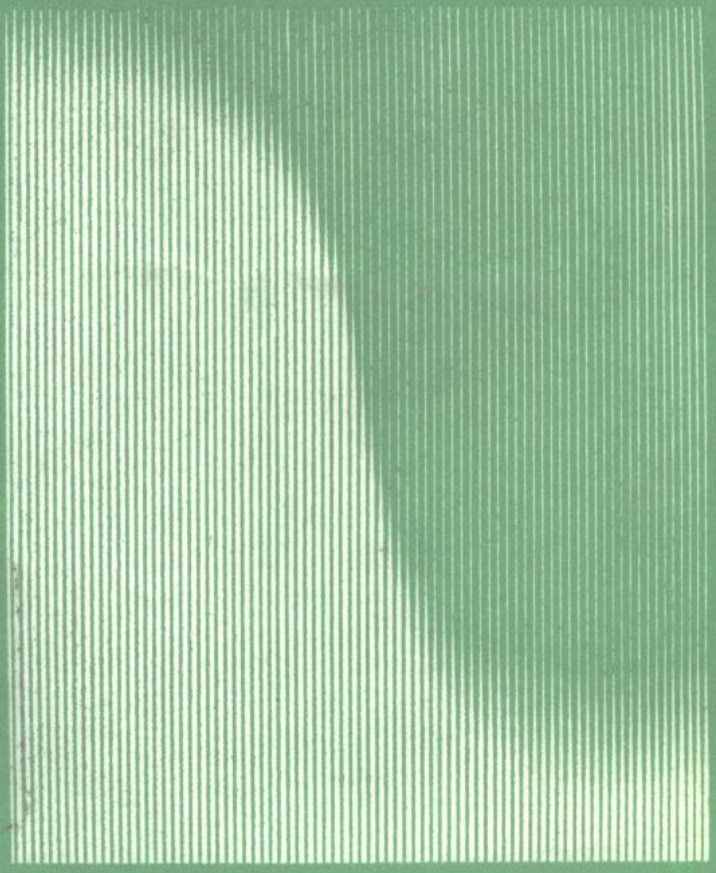
文献计量学基础

丁学东　编著



北京大学出版社

文献计量学基础

丁学东　编著

北京大学出版社

新登字(京)159号

书　　名:文献计量学基础
著　　作
责任者:丁学东　编著
标准书号:ISBN 7-301-02151-8/G·177
出 版 者:北京大学出版社
地　　址:北京大学校内
邮政编码:100871
排 印 者:北京大学印刷厂
发 行 者:北京大学出版社
经 销 者:新华书店
版本纪录:850×1168毫米　32开本　12印张　300千字
　　1993年9月第一版　1993年9月第一次印刷
　　印数:0001—4000册
定　　价:11.00元

序

文献计量学是信息科学领域中的一个重要分支学科，目前在许多学科中日益得到广泛的应用，是情报学研究最活跃、发展最迅速的专业领域之一。

十几年来，文献计量学研究在我国已取得了长足的进步，关心的人日渐增多，并已获得了一些可喜的成果：或在理论或应用方面进行深入研究，推动了学科的进步；或在学科知识系统化方面做了不少工作，推动了学科知识的普及。

丁学东同志在读研究生时就开始了对文献计量学的研究，留校任教后又承担了这门课程的教学任务。多年来他非常注意国内外的有关研究活动、新的动向和新的发展，并扎扎实实、默默地开展了一些有意义的研究工作。《文献计量学基础》简明易读，层次清楚，论述准确。该书不仅使文献计量学知识条理化，而且也融会了作者自己的研究成果。

本书的出版对于文献计量学的教学及学科的发展都是有益的。

王万宗

1992 年 9 月 26 日

前　言

文献计量学是采用定量手段以各类文献为对象研究文献交流过程中存在的各类数学规律的一门新兴学科。确切地说，就是以数学与统计学的方法来探讨文献交流的规律，从而加强情报科学的理论体系建设，并借以提高文献情报管理的科学水平，同时也将有助于揭示科学技术发展的一般规律。

具有现代意义的文献计量研究，可以认为发端于本世纪的二十年代。短短的几十年时间里，特别是自六十年代以来，该学科在理论与应用两方面均取得了长足的进步。目前，虽然该学科的内容尚不够成熟，体系也不够完备，但是由于它在情报学、图书馆学等学科的研究中所取得的显著成绩和展现出的诱人前景以及与诸多社会科学学科构成的共同统计背景，日益受到包括情报学图书馆学在内的各有关学科学者们的普遍关注。近年来，国内外一些高等学校陆续设置了文献计量学这门课程，并相继有教材问世。

北京大学图书馆学情报学系自 1987 年在本科生和研究生的教学中正式开设了这门课程，于今已历四载，本书就是在教学和科研的基础上对原教材《文献计量学讲义》进行增删而成的。希望读者通过本书能对文献计量学的研究内容、方法及有关应用有一个概括了解。全书共分八章，介绍的内容主要有：文献增长与文献老化的理论、文献计量学的三条基本定律和一般原理以及引文分析等等。每章后面均附有有关的参考文献。内容较为专深的章节皆以※符号标出，略去不读并不会影响读者对全书基本内容的理解和掌握。为了达到通俗易懂的目的，本书并不涉及较高深的数学知识，读者仅需具备初等微积分、线性代数和概率数理统计的一般知

识，即可通晓全书的主要内容。

本书在编写、修订和出版过程中，始终得到我系周文骏教授、朱天俊教授、王万宗教授和情报学理论教研室各位教师，特别是徐万丽编辑的热情支持和帮助，在此，谨向上述所有同志以及书中被引文献的作者表示衷心的感谢。

作者限于水平和时间，虽黾勉从事，数易其稿，但谬误与疏漏之处，在所难免。诚望专家和读者给以批评指正。

丁学东

1992 年 5 月于北京大学

目　录

第一章 绪 论

对于任何学科都是一样，数学成份的多寡决定了它在多大程度上够得上是一门科学。

——康德

第一节 文献计量学概述

1948年，印度图书馆学家阮冈纳赞(S. R. Ranganathan)曾指出："由于图书馆工作与服务涉及到大量的与数有关的问题，图书馆学工作者必须得仿效生物计量学、经济计量学和心理计量学来发展图书馆学的计量技术。"当时图书馆学界对这番话的感觉可能只不过是他给图书馆工作的前途重重地抹上了一笔漂亮的油彩罢了。到了六十年代，当美国科学学、科学史专家普赖斯(D. Price)令人信服地阐明了科技文献按指数增长规律时，人们似乎已经相信，在图书馆书架上一堆堆薄厚不均、参差错落的书刊中的确可能存在着与优美简单的数学曲线有关的东西，这实在是足以令人振奋的事情。而在今天，当英国情报学家布鲁克斯(B. C. Brookes)多次向学术界表明，在包括情报交流过程在内的有人参与的各种社会现象中存在着与经典高斯分布平等地位的布拉德福-齐夫分布时，人们已经切切实实地感觉到，文献计量学可能已经不仅仅是一种对文献传统定性描述的有益补充，而且是一门有自己理论基础、与其它学科并立的新兴学科了。

一、文献计量学的产生

科学技术的迅猛发展是自文艺复兴时代以来，人类获得的最值得骄傲的伟大成果。在不足五百年的时间里，人类对于世界的认识和使其发生的变化，是过去几千年文明史所创造的全部价值难以望其项背的。同时，人们也切实地觉察到，当今科学技术每向前迈进一步所付出的代价也是过去任何时代所无法比拟的。究其原因，大致有二：(1)不仅在于我们对尚未涉足的未知世界更难登堂睹奥，板薄之处所余无几，钻之弥艰。(2)同时，人类活动本身也给这个世界造成了一系列新的问题。

对于从事情报学、图书馆学以及一切与情报文献(特别是科技文献)工作有关的人们说来，所面临的主要问题显然是后者。经验告诉我们，解决这类新问题的难度是相当大的。我们不仅必须应付由于科学技术高速度发展造成的科技文献绝对数目的急剧增加，以及由此而产生的科技文献种类繁多、质量不一的混乱局面，还必须要构造独立的文献交流理论体系，即这种理论体系不应该也不可能是其它学科理论体系的派生产物。

自本世纪六十年代以来，由于文献交流系统逐步采用了电子计算机等一系列新技术，无疑使人们应付文献量增长的能力有了极大的提高，但计算机等新技术的出现，没有而且也不可能为我们自然而然地建立本领域真正的独立理论体系，况且计算机等先进技术所能发挥作用的大小最终仍将取决于我们对文献交流规律的认识深度。

众所周知，迄今为止的文献交流基础理论(主要是包括传统的图书馆学理论和正在形成的情报学理论)基本上仍是通过定性手段得出的结论。但是自本世纪二、三十年代以来，人们陆续开始利用数学的方法对各类文献群体内部以及各类文献间的相互联系进行了定量研究。尤其在二次世界大战后，文献量的急剧增长和文献

价格的大幅度上涨与广泛存在于图书情报部门的有限资金、有限存储空间和整理能力之间的矛盾日趋突出，同时，由于科学技术的迅猛发展，面对浩如烟海的各类文献资源，科学工作者都希望能迅速、准确地占有他们所需要的资料，从而对文献交流的速度和质量都提出了更高的要求。总之，这一切都促使人们对各类文献群体进行数学、统计学的研究来帮助解决上述问题。

请看下面的例子：

某图书情报部门所获得的期刊采购资金量为 B；经调查，若欲满足读者的需求，须采购 N 种学科或专业的期刊。试问在上述条件下（当然，条件的类型与多寡可因部门的具体情况而定。这里为说明问题，仅选取上述两种条件。），最优的采购策略是什么？显然只有采用定量手段此类问题才能得到满意的解答。

我们设：

1. 用 U 来表示某种采购方案实施后所获得之效益；

2. U_i 为第 i 种学科或专业由于实施上述方案而获得的期刊论文数量；

3. a_i 为表征上述期刊论文相对重要性之权重系数（定义权重系数的方法甚多，例如，可以用第 i 种学科或专业期刊在给定时间内的流通量(circulations)等来表示）；

4. X_i 为刊登上述论文的期刊数量；

5. P_i 为上述期刊的平均价格。

在此例中，最佳采购策略显然要涉及如下两个问题：

1. 不同专业期刊的数量分配；

2. 同一专业期刊的数量分配。

为此，我们令

$$U = \sum_{i=1}^{N} a_i U_i$$

于是有：

$$\begin{cases} B = \sum_{i=1}^{N} P_i X_i \\ U = \sum_{i=1}^{N} a_i U_i \end{cases} \tag{1-1}$$

为了简化说明问题，在这里仅令 $N = 2$，则上式变为：

$$\begin{cases} B = P_1 X_1 + P_2 X_2 \\ U = a_1 U_1 + a_2 U_2 \end{cases} \tag{1-2}$$

为了求得 U 的最优值(在本问题中，U 的最优值即为 U 的最大值)，可采用拉格朗日乘子法，即设函数：

$$L = U + \lambda\varphi \tag{1-3}$$

其中：λ 为不为零之常数(即拉格朗日乘子)，并令：

$$\varphi = B - (P_1 X_1 + P_2 X_2) \tag{1-4}$$

为了求得 U 的最大值，令

$$\begin{cases} \dfrac{\partial L}{\partial X_1} = 0 \\ \dfrac{\partial L}{\partial X_2} = 0 \end{cases} \tag{1-5}$$

由此得到：

$$\begin{cases} a_1 \dfrac{\partial U_1}{\partial X_1} - \lambda P_1 = 0 \\ a_2 \dfrac{\partial U_2}{\partial X_2} - \lambda P_2 = 0 \end{cases} \tag{1-6}$$

消去 λ 后有：

$$\frac{a_1}{P_1} \cdot \frac{\partial U_1}{\partial X_1} = \frac{a_2}{P_2} \cdot \frac{\partial U_2}{\partial X_2} \tag{1-7}$$

为了将上式与资金限制条件 $P_1 X_1 + P_2 X_2 = B$ 联立以求得 X_1 和 X_2 的值，必须要得到 $U_1(X_1)$ 和 $U_2(X_2)$ 的具体表达形式，用数理统计的语言来说就是要获得论文量关于期刊的累积分布函数。这种数量关系的确立正是文献计量学所要研究的问题之一。我

们将在本书第四章中对此予以详细讨论。

六十年代以来，与此类似的定量研究得到了迅速的发展，并逐渐形成了较为系统的文献定量研究学科——文献计量学。它的出现不仅大大加深了人们对各类情报产生、传递及吸收过程的认识，而且还取得了一些根本不可能用传统定性手段获得的重要成果。本书的目的即是要将本学科中的主要理论、研究方法及其应用介绍给读者，希望大家能对这门新兴学科的主要内容有初步概括的了解。

另外，在这里尚应补充提及为大家所熟知的学科发展规律，即定性研究发展到一定的阶段必然会迎来采用定量手段研究的时代，如生物计量学、经济计量学、定量语言学、社会计量学及历史计量学等都是诞生于传统的定性学科之内。因此不难想象，图书馆学，情报学等也概莫能外。当然还应指出，自本世纪五十年代以来，申农(C. Shannon)的“通讯的数学理论”(the mathematical theory of communication)与维纳(N. Wiener)的“控制论”(Cybernetics)等对于文献交流的定量研究都起了重要的作用。

二、文献计量学的定义

文献计量学(bibliometrics)可以认为是由“统计书目学”(Statistical bibliography)一类名称演变而来，后者是由英国伦敦专利局的图书馆员休姆(E. W. Hulme)于 1922 年提出的[1]，当时主要是指通过简单文献计数并用常规的统计方法来揭示“人类文明进程的定量研究手段”，由于种种原因，在图书馆学、情报学界没有广泛地采用这一术语。

1969 年英国计算中心的普里查德(Alan Pritchard)首先提出用“文献计量学”这一新名称来代替“统计书目学”一词[2]。次年出版的《图书馆文献》(Library Literature)和《图书馆学情报学文摘》(LISA)就已将其作为标目使用。在此后的 20 年时间里，该术语已

得到了图书馆学情报学界的普遍认可，并在《图书馆学情报学百科全书》第七本补编(1987)(A Supplement to the Encyclopaedia of Library and Information Science)中有详细的阐述。文献计量学这一名称的提出不是偶然的，除了A.普里查德认为"统计书目学"这一名称存在的一些诸如表达不够明确等不足之处之外，恐怕更主要的原因乃是文献的定量研究水平，范围已非昔日可比，用"统计书目学"这一名称是不能概其全貌的。的确，40年后的文献定量研究水平已经不再是简单统计计数和用简单百分数来表示结果的初等统计定量阶段，而是逐渐被以数理统计以及其它数学工具为研究手段，以计算机为计算工具的高水平研究工作所取代。显然，原始意义上的统计这一术语已不敷用，必须代之以能反映水平更高，范围更广的定量研究术语。还应指出的是，当时已出现了生物计量学(biometrics)、经济计量学(econometrics)和定量语言学(quantitative linguistics)等这样新兴的定量研究学科。特别是在前苏联和东欧还出现了与文献定量研究内容有密切关系的"科学计量学"(scientometrics)(为科学学(science of science)的一个新兴分支学科)。该名称来自俄语"наукометрия"，其含意为对科学技术进展的计量研究。它的计量单位可有多种，其中有许多与情报计量单位相同，其研究方法也相似。毫无疑问，这一类新术语的问世必然使普里查德得到了启发。因此用书目(bibli-)和计量学(metrics)构成一个新的术语文献计量学(biblio-metrics)，看来已是十分自然的事情。

关于文献计量学的定义，由于本学科仍处在迅速发展之中，很多方面尚不够成熟，致使人们对其认识不尽相同，故至今仍众说纷纭，或繁或简莫衷一是，就笔者所见，大概也不会下于20余种。舍繁就简，本书根据普里查德的原始提法"将数学和统计学的方法运用于图书及其它交流介质的研究"[3]，稍加修改变为"将数学和统计学的方法运用于图书及其它交流介质研究的一门学科"作为文

献计量学的定义。请注意，当时普里查德仅将文献计量学解释为一种对文献进行定量研究的手段，但事隔三年[4]，他已经将文献计量学进一步解释为“情报传递过程中的计量科学。其研究目的乃是对该过程进行分析和控制”。在这里普里查德已经将其定义为一门计量科学。

值得注意的是，这里的“图书及其它交流介质”的含意非常广泛。它不仅包括我们所熟悉的各类具体的文献（如书、刊、科技报告、会议论文、专利说明书等等），而且还包括一切与文献有关的可以计量的各类指标（如作者数量、词频统计、引文数量、流通数量、复制数量等等）。很明显，文献计量学这一术语中的“文献”二字必须要做广义的理解，决不仅仅指通常意义上的具体文献。

八十年代以来，随着定量研究的不断深入，情报学界逐渐感觉到“文献计量学”这一名称又已不能完全包括日益扩大的研究范围。布鲁克斯就曾在1988年指出过[5]，文献计量学这一名称将我们过窄地限制在图书馆以及传统的文献范围之内。因此文献计量学这一名称今后仅应专指对图书馆和书目的定量研究。而东欧和前苏联使用的科学计量学这一名称又过于偏向于对科技发展中实际问题的研究。同时布鲁克斯尚考虑到由于新技术（主要指电子计算机）出现而不断产生的知识记录及传播的非传统文献形式，于是建议应该采用前西德人纳克（Otto Nacke）于1979年提出的情报计量学（informetrics）这一名称来取代使用仅二十几年的文献计量学。目前情报学界对是否采取这一新名称仍持有异议，但估计在不久的将来这一取代必将成为现实。情报计量学显然将基本包括方法相似、目标各异的文献计量学与科学计量学这两门学科的研究内容。

三、文献计量学的研究对象和目的

根据计量数据的来源，文献计量学的研究对象可大致分为如

下四类：

1. 出版物

各个学科和技术领域所发表和被引用的各类文献是文献计量学的主要对象。上述文献载有大量可资分析的信息，是数理统计施展其威力的绝好用武之地。在进行这类研究时，应充分利用它们的控制工具，如各类检索工具：文摘、索引、年鉴、手册、百科全书、指南等，例如，美国科学情报社出版的《科学引文索引》(SCI)就是当今进行引文分析不可缺少的工具。目前较好的这类工具书对专业文献的收集一般都比较完备，而且所记录的文献特征也很全面。因此充分利用上述工具可以很方便地获得大量的有关信息，从而对各类文献出版的数量和时间、文献出版类型、文献出版使用的语言和文献出版地域分布特征进行各种计量研究。

2. 著者指标

各类出版物，特别是期刊论文的数量与其作者数量之间关系是文献计量学研究的重要内容之一。一般讲，获得这类数据是比较容易的。例如各类检索工具均附有累积索引的作者部分，甚至许多期刊的年终本也都附有有关作者的累积索引。收集著者指标的专门工具书为《科学论文著者名录》(Who is Publishing in Science)，它罗列了《近期目次》(Current Contents)所报导的全部论文的主要著者的姓名及地址。可以利用该名录来了解各国各学科科学家的人数与相应出版物的分布情况。这里应指出的是，在大科学时代的今天，合作者、团体机构作者日趋增多，这一现象也是文献计量学当前有关研究中的重要内容，因此应充分利用各类工具书的多重揭示、互设参见等手段来进行此类数据的收集。

3. 服务指标

对图书情报单位的各类登记记录以及读者问卷进行统计分析可以获得大量可资研究的数据和资料，如阅览数、借阅数、资料复制数、读者类型分布、使用文献类型和情报需求特点等等。特鲁斯

威尔(R. W. Trueswell)得出的著名80/20规律就是通过资料流通数量的统计得出了文献资源利用的重要规律。

4. 词语指标

众所周知,词语在情报检索中有重要的作用,因此对各类词语的研究一直是图书馆学情报学界关注的问题之一。近年来定量方法已成了检索语言研究的重要手段,例如利用词频分布的规律来选择标引词等等。各类词表(如词频词表和叙词表等)也是这方面研究经常使用的辅助工具。

文献计量学在文献情报系统中的研究目的是非常明确的,首先它试图将长期建立在定性基础上的传统图书馆学情报学提高到定量水平的高度,从而进一步揭示文献情报体系的结构和其中存在的数学规律,这必将使图书馆学情报学的理论更加完备、更加科学。同时,它也将会为图书情报部门各项工作提供可资参考的数量根据,以便使文献资源的利用处于最佳的状态。此外,文献计量学也是科学学和科学史等学科的理论研究及实际应用的重要手段。近年来通过文献计量手段所获得的各种科学指标已成为许多国家判定科技政策和预测未来的重要依据。在这方面,匈牙利和荷兰学者的贡献最为突出。

四、文献计量学研究的意义和作用

从文献计量学的研究目的可知,文献计量学必然对下面的各项工作具有重要意义和作用。

1. 在情报学理论的研究中,可以在原有定性的基础上,采用定量手段对社会信息流通状态的规律、文献信息选择与整序的理论方法以及对情报用户的需求、需求类型、解决需求的行为和用户需求内容分析等进行研究,这也是文献计量学理论研究的核心部分。

2. 在图书情报工作中,可以根据诸如文献增长、老化的数学