索引研究论丛 ·葛永庆·主编

索引工作自动化

中国索引学会

索引研究论丛　　葛永庆主编

索引工作自动化

中国索引学会

责任编辑： 凌　刚
封面设计： 陶烈哉

索引研究论丛
索引工作自动化

中国索引学会出版
（上海中山北路 3663 号，华东师范大学逸夫楼）
南京市利学印刷厂印制
开本：850×1168　1/32　印张：　5.6　字数：145 千字
1994 年 11 月第 1 版　1994 年 11 月第 1 次印刷

印数：1—2000 册　　内部交流　酌收成本

中国索引学会
简　介

中国索引学会(Index Science Association of China)是从事索引研究和编纂的非营利性学术组织，也是具有法人资格的全国性社会团体。经国家新闻出版署通过资格审查，中央民政部核准登记，于 1991 年 12 月 24 日在上海成立。总部设在华东师范大学逸夫楼。学会以“真诚、求实、开拓、奉献”为办会宗旨和会员活动准则，旨在促进索引理论研究，繁荣索引编辑出版，培训索引编纂人才，加强国内外学术交流。学会设有秘书处、编纂部、研究部，下设《丛书》、《论丛》两个编辑部以及“哈佛燕京学社引得编纂史”、“中文检索与输入”、“索引技术和索引标准”等研究室。

《索引研究论丛》书目

索引的昨天、今天和明天

* * * * * *

论索引和索引法

* * * * * *

索引工作自动化

《索引研究论丛》

编　例

一、本《论丛》定名为《索引研究论丛》，不定期分辑出版。

二、《论丛》主要选载中外索引史、索引法研究、各类索引比较、编纂经验介绍、索引理论以及索引自动化和排检法问题探讨、前辈索引家传略、索引评介、索引标准化和规范化问题的研究，等等。

三、本《论丛》选载文章除少数曾经发表过但作了增补修订以外，多数是新写的。

四、入选文章不尚空谈，力求言之有物，学术性、实用性兼顾，尤其重视总结实践经验的来稿。

五、文章一经刊用，酌致薄酬。来稿请寄：上海中山北路 3663 号华东师范大学逸夫楼中国索引学会《索引研究论丛》编辑部，邮编 200062。

中国索引学会

《索引研究论丛》编辑部

一九九四年三月

目　录

电子出版物与索引编制自动化

陈光祚

一、电子出版物的特点

所谓电子出版物，是指其本身以电子文本形式出现、以磁盘或光盘为信息载体，具备相应的全文数据库检索软件、并可利用计算机对其进行高速准确地浏览与检索的出版物。

对于传统的印刷型出版物来说，电子出版物是一种体现现代信息技术的新型出版物。无论是外观或功能，电子出版物与印刷型出版物都有很大的区别。电子出版物除具有与印刷出版物相当的文献文本之外，还有如下显著功能：

1. 文献的全文检索功能。借助于全文数据库的结构与检索软件，文献中的任何人名、地名、年代、关键词均可被单项地或多项组合地检索出来。在数十万、数百万、乃至数千万字的巨著中，要查找某一事项或词句，可在很短的时间（数秒钟内）一索即得。

2. 信息计量功能。可对文献中的事件、名词、人物、地点等项目进行频率统计和排序，获得相应的统计排序表。可以借助于这种统计排序表，以了解文献内容的重点和所涉及的主题范围，判断所论述问题的焦点，了解有关人物、事件、地点的影响程度。

3. 知识项聚类的功能。文献中的各种知识项（如人物、事件、地点、名词、年代等），可以彼此组合，以获得有关信息的聚类。例如，在一部历史著作的电子出版物中，通过知识项的聚类，可以了解某

一事件涉及哪些人物、哪些地点；或者某一人物曾经与哪些人发生过关系；某一地方曾出现过哪些事件；等等。

因此，电子出版物不仅具有自动检索的功能，还具有对文献内容进行分析的功能。它要比印刷型出版物具有更高的使用价值。

电子出版物的制作与发行，无疑是出版界的一次技术革命。它为出版产品增加一个新的类型；它由于省去了传统出版物生产过程中的制版、印刷、装订等工艺环节，从而缩短了出版周期；它便于文本的增、删、改，从而易于文献内容的更新和新版的出版；它由于可以作为全文数据库通过联机情报检索系统提供广大用户利用；从而开辟了一个新的发行渠道与形式；它的制作是利用当前出版过程中激光照排（电子排版）的中间产品——电子文本为基础，依靠加工软件在计算机上自动或半自动进行的，不仅速度快，在费用上也是低廉的；由于磁盘和光盘的信息存贮密度高，因而电子出版物的体积小，便于邮寄与携带；电子出版物可以随时进行拷贝，社会需要多少份就能拷贝多少份，因而不会造成出版物的积压或脱销；由于电子出版物的文本是计算机可以识别的数字编码，因而便于电讯线路的高速传输，从而能满足对出版物的某些紧迫需求。

在国外，一些发达国家的电子出版物已达到相当普及的程度。它与印刷版出版物的比例，正在不断上升。许多百科全书、手册、词典等著作，以光盘形式发行，或以联机情报检索方式向用户提供服务。而在我国，电子出版物的制作与发行则处于起步阶段。

我们武汉大学图书情报学院“全文数据库研究”课题从1986年以来开始从事全文数据库的研究开发工作。作为国家教委文科博士点基金项目，经4年多时间的研究，于1990年10月制作成功《湖北省地方志·大事记》、《中国人民解放军大事记》（解放战争时期）两部著作的全文数据库，探索和实现了全文数据库的结构、标引规则、建库和检索软件以及对检索结果的再次编辑与输出的技术。1991年7月，与武汉大学出版社合作，制作成《国共两党关系通史》（150余万字）的电子出版物，正式由出版社向社会公开发

行。中央及地方的报刊和电视台对此曾作了报道，并给予了高度评价。该出版物已发行至北京及香港等地。我们经过技术上的改进，1993 年又由该社发行了《中国名胜诗词词典》(70 余万字)电子出版物。近来，国内出版界对电子出版物给予了高度重视。国家新闻出版署已将它列为正式的出版物类型，承认其版权，召开了电子出版研讨会，成立了全国电子出版研究会；一些专门制作发行电子出版物的公司与出版社也已出现。

二、电子出版物将是图书馆藏书的一个新品种

电子出版物的出版发行不仅是出版业的深刻变革，而且对图书馆与情报资料单位以及广大读者用户都会产生深刻的影响。

首先，电子出版物使图书馆与情报资料单位面临一个新的收藏品种。从古到今，图书馆收藏各种载体的知识记录。刻有文字的甲骨，写有文字的竹简木板或贝叶、泥版，都曾经作为古代图书形式而被古代图书馆所收藏；纸发明以后，各种写本更是图书馆的藏品；印刷术发明之后，印刷型出版物成为图书馆藏书的主体；缩微品的出现，又扩大了图书馆收藏的范围；录音带、录像带出现之后，音像出版物也开始进入了图书馆的藏书。今天，电子出版物之成为图书馆藏书的一个新类型，应该是一种很自然的事情。这也应该是图书馆现代化的一个表现。电子出版物作为一种藏书，有其独特的优点，也提出了某些特殊的问题。优点是，以磁盘或光盘形式出现的电子出版物体积小，所占库容小；对于使用频繁的工具书或珍贵的图书来说，如果有相应的电子版存在，就会避免图书的破损、污染，从而有利于保存文献原件或真迹。当然，并不需要对任何图书都购买其电子版，目前主要是应收藏下列类型的电子版图书：大部头的的工具书(如大百科全书)，具有经典性或历史意义的巨著(如二十五史)、珍本书等。一些儿童读物或外语学习等电子出版物，寓文字、图形、声音于一体，既给人以知识又给人以美的享受，因而十分吸引儿童读者。与此相似的是，一些能突出反映美术、舞蹈、服

装、名胜景点风光等文献内容的电子出版物(如多媒体光盘),也宜于选择收藏。因为它们的独特功能是印刷版图书所不能取代的。

要将电子出版物作为图书馆的一个收藏品种,有一系列问题需要研究解决。例如费用问题,分类、编目、保管和流通方式的问题,相应的服务设备与环境要求问题等等。

关于费用问题。电子出版物的价格目前比印刷书刊昂贵。这一事实,可能造成图书馆不堪承受的负担。此外,图书馆要向读者提供电子出版物的浏览与检索服务,还需配置微型计算机和光盘驱动器。这也是新增加的经费负担。因此,电子出版物要成为图书馆的收藏对象,目前有不少困难。

但是,也有不少因素推动电子出版物进入图书馆藏书的范围。

第一,电子出版物的功能是印刷版书刊所不能取代的。特别是处于信息时代的条件下,迅速准确地获得信息,将带来巨大的效益。信息本身就是一种潜在的财富,它能够转化成物质财富。从用户获取信息、并将其转化为生产力、转化为经济技术竞争的能力这一角度来看,他们支付检索电子出版物的投入费用,与他们可能从中获取的产出效果相比,有时是微不足道的。在我国条件下,从宏观来看,图书情报工作手段的现代化,需要支付比传统工作方式要多得多的费用,但是它们的效益比传统工作方式所能取得的效益更大更多,要比投入更大更多。否则现代化是不可思议的。图书馆以有偿服务的方式提供电子出版物的服务,用户是能够理解和接受的。而这种有偿服务可以使电子出版物收藏的负担由读者用户来分担。

第二,随着现代信息技术的迅速发展,电子出版物的成本将不断下降。这与印刷版出版物的连年涨价造成显明的对比。在发达国家,已经出现这种趋势:印刷版书刊品种与发行量日益萎缩,而电子版所占比重在不断上升。越来越多的书刊只出电子版而不出印刷版。这个现象不能不说有其经济上的原因。这是因为:信息技术的进步使信息的存贮密度越来越高,单位存贮费用就相应下降;

电子版制作软件越来越成熟，制作效率的提高也有利于降低电子出版物的成本；电子出版物将日益赢得更多的用户，它的发行量的增多也必将使其成本降低。

第三，电子出版物的邮寄费用低，占用书库的体积少，倒架方便，等等，这也是经费问题考虑的一个因素。

此外，图书馆除了收藏电子出版物外，还可以通过联机情报检索方式利用联机情报检索系统中所装载的电子出版物。这也是图书馆拥有电子出版物使用权的一种方式。在这种服务方式中，图书馆或用户不必为整个电子出版物付费，只为实际的检索量付费。对于使用频率较低的电子出版物，用这种方式进行利用是合算的。

电子出版物的整理、保存、流通等工作，也给图书馆提出了新的问题。这主要是电子出版物这种新型文献载体，以及其独特的使用方法所引起的。但是音像资料的整理、保存、流通问题，可作为借鉴。当然两者不完全相同。磁盘形式的电子出版物更需要注意数据的安全，不致被有意无意的操作所破坏，同时要注意防止计算机病毒的入侵。电子出版物的利用服务，还需配置相应的打印机，以便应读者用户的要求将其检索结果打印出来。从事这方面流通服务的馆员还需向使用者普及计算机数据库的初步知识和使用方法。

三、电子出版物是参考咨询的有力工具

对图书情报单位来说，电子出版物使它们获得强有力的参考咨询工具，即可以进行自动检索的全文数据库。一部著作，尽管作者是按照一定的逻辑次序、分章分节撰写的，比如按照由浅入深的次序，或历史发展的顺序，或由一般到具体，或由分析到综合，或从理论到应用等等的顺序进行阐述的，从而使有关的知识得以有序地组织。但是这种有序化是单维的。读者可以按照作者的逻辑顺序来理解和查检该著作，而更多的情况是，读者可能从多种角度和顺序来探索该著作，寻找其中的有关的事实和知识片断，进行错综

复杂的联系。也就是说，著作中的逻辑顺序不能完全满足读者检索的需要。这就是原著中的单维次序和读者的多维检索要求之间的矛盾。作为全文数据库的电子出版物却从根本上解决了这个矛盾。在全文数据库中，设立了多种索引项目，从而提供了著作中多种事项的检索点，便于读者随机检索，而不必顺序浏览。即使著作中的某些事项未作为索引项目，也可发挥计算机高速顺序扫描的能力进行检索。因此，作为全文数据库的电子出版物，它的全部文本都可由计算机进行控制与操作。一般供阅读的书也被赋予了多种检索的能力，从某种意义上来说，一般阅读书和工具书的界限正在消失。电子出版物的出现，使得有关的文献信息从出版这一发源环节就开始得到了计算机化的控制，提供了多维的、有序化的、可操作的功能。电子出版物的实质是全文数据库，可以借助于这种全文数据库对其中的数据进行多种多样的情报加工，除检索之外，还可进行有目的地抽取、排序、重新组合，从而产生新的情报产品。因此，电子出版物是进行其他情报工作的良好基础和手段。

即使是对工具书这类出版物来说，它的电子版与印刷版之间也存在功能上的差别。尽管象书目、索引、文摘、年鉴、手册、名录、辞书、百科全书之类的工具书在编撰时已将资料与数据进行条目化，各条目之间以便于查检的次序编排，并有多种索引附于书后，以提供不同于正编内容结构的检索途径。但是印刷版工具书所能提供的检索功能仍然是静态的，而不能提供动态的逻辑组配功能。举例来说，印刷版的书目索引，虽然按主题词（或分类号）排列，在每个主题词或分类号之下著录书目款目或文献编号，用户如果仅仅只查一个主题词或分类号，则可很方便地获得该主题词或分类号之下的文献线索，但是如果用户所要检索的目标是包含多个主题或分类号在内的复合概念，或者甚至是多种系列标目（如主题词是一种系列标目，分类号是另一系列标目，同样，著者姓名、出版时间、著作文种、发表出处等等都可以分别是系列标目）之间的组合，那么印刷版的工具书本身就难于提供这些复合的检索点了，而必

须依靠用户人工地来进行组配，即对比多个标目之下的文献号码，取其在各标目下都存在的相同号码，以获得所需文献，从而实现了“逻辑乘”的组配功能。当然，有的印刷版工具书为了提高检索性能，采用多级主题的方法，或多个主题轮排以互相说明与限定其含义的方法，或增加文献条目说明语的方法等等，其目的是在一个检索标目之下扩充其信息含量，提高检索的分辨能力，在一定程度上提供复合概念的表达与检索能力。但从根本来说，这仍然是静态的检索功能。与此相反，电子版工具书所能提供的多元检索标目的组配都是随意的、动态的，并且这种动态组配的能力，除了逻辑乘之外，还能实现逻辑或、逻辑非，或各种不同逻辑运算的组合组配。检索的角度、广度和深度可以随不同的用户、不同的检索目的而改变。这种动态组配是随用户的检索式而响应的，灵活的、真正多维的。确切地说，这种动态组配的能力是计算机情报检索软件所赋予的。显然，这是电子出版物之优势所在。

因此，电子出版物可以成为图书馆的文献参考咨询的强有力的工具。凡是著作中任何信息单元，如人名、地名、年代、关键词、片言只字，均可借助于作为电子出版物的全文数据库进行检索。上述信息单元，凡已被标引作为检索项的，可以通过数据库中的索引直接检索；没有被标引的，虽然索引中无此检索项，也可通过全文数据库的文中扫描办法进行检索。因此，文献中任何隐藏着的信息都被置于计算机的控制之下，随时可供查询。这样，原来无检索工具职能的文献，也都具有了检索能力。这是电子出版物的一个十分重要的特点。计算机全文数据库的强大检索功能，远远超过传统参考咨询人员的“博览群书”、“强闻博记”。在“记忆”方面，电脑是超过人脑的。可以设想，当许多经典性著作和资料价值极高的著作成为电子出版物时，就会使图书馆拥有丰富的参考咨询的信息资源。那时，传统的参考咨询工作就可以利用各类数据库（如书目数据库、事实与数值数据库，以及全文数据库）进行系统全面、高速准确的查找，从而使参考咨询工作的流程发生新的变化，步入数据库时

代。当然，并不要求每个图书馆都拥有包括全文数据库在内的各种数据库，而可以是借助于联机情报检索系统的终端，调用检索系统中所装有的数据库。也正因为如此，使传统的参考咨询工作由各馆各自为战的方式变成社会化的情报服务。这种社会化的情报服务方式，确实改变了各个图书馆以馆为单位、依靠本馆收藏和本馆参考馆员的那种手工作坊式的服务，而可以借助通讯网络，依靠广泛得多的参考信息资源，利用包括全文检索手段在内的各种自动化的处理手段，使图书馆参考服务水平登上了新的台阶。与此相适应的是，对参考服务人员的素质要求也产生了新的变化，即不仅要求他们"博览群书"，熟悉文献，掌握各种检索工具书及其手工检索方法，而且要求他们了解各类数据库，熟悉计算机检索的技能，并且善于把计算机检索和手工检索结合起来。这就要求他们学习新的知识和技能。

四、电子出版物使文献计量学获得新的应用领域

电子出版物使文献计量学的分析得以深化。文献内容的各个知识项，甚至字频统计的计量分析为文献计量学开拓了新的应用领域。

文献计量学以文献单元（如一篇文献，一部图书等）为对象，通过数量的统计分析，包括文献的绝对数量、相对数量等的分析，来揭示文献的增长、分布、老化等规律，从而获得文献领域的各个总体性概况，有助于核心著者和著作、核心期刊等的测定，并有助于通过文献现象，进而了解一个国家、一个地区、一个知识部门的知识生产、传布、吸收以及学术界整体素质等的更普遍性的情景。然而，文献计量学毕竟是以文献单元（包括其中的标引项）作为统计单位的，而不是以文献中的知识单元为单位的，因此，这种计量分析仍然显得较为粗泛。电子出版物作为一种全文数据库，它的文本中的任何知识单元甚至每个单字都是可以检索和统计的。这样，就有可能使文献计量学的计量单元从一篇篇文献而深化到文献中的

各个知识单元，甚至单字一级。与此相适应的是，基于这种文献内容细节的计量分析，可以获得一般文献计量分析所达不到的深度。

作为全文数据库的电子出版物的计量分析可以有如下一些角度：

1. 词频统计。具有全文检索功能的电子出版物，一般在对词进行标引时同时也指明词的属性（如关键词、人名、地名、年代等等）。不同属性的词可以分别统计分析。例如全书的关键词频率统计，可以据此了解著作论述的中心论题与边缘论题；如果对连续出版物的电子版分时期进行这种统计，可以据此了解新出现的名词术语，从而了解学科的新发展，并了解一些有关概念的分化或组合。对于著作中的人名进行统计，则可以根据其出现频率，判断各个人物的影响和著名度。

词频统计可以在不同的字段中进行。在电子出版物中，一般来说，为了便于数据库的建立和检索，对文本是设立记录或字段的。例如，在我们研制的《中国名胜诗词词典》电子版（1993 年武汉大学出版社出版）中，每一首诗（或词）作为一个记录。而在一条记录中，分设诗（词）文、诗（词）名、作者姓名、作者介绍或写作背景、诗（词）文中难点的注释、所歌颂的景点、该诗（词）在印刷版图书中的相应页码等字段。词频统计（也包括字频统计）可以在不同的字段中进行。比如，如果确定诗（词）文这一字段作为统计的对象，则可以发现许多有用、有趣的事实。在这部《中国名胜诗词词典》中，收集了全国几百个景点有关的 3000 多首诗词。在诗（词）文这个字段中，若进行字频统计，可以发现："红"、"橙"、"黄"、"绿"、"青"、"蓝"、"紫"等 7 个字，其中"青"字出现频率最高；"喜"、"怒"、"哀"、"乐"、"悲"5 个字，其中"悲"字出现频率最高；"大"字的频率比"小"字高；"山"字的出现频率比"河"、"江"、"海"、"潭"、"湖"等字高；在"风"、"花"、"雪"、"月"、"人"等字中，"人"字的出现频率最高。从这些字频统计中，有助于从整体上了解中国名胜诗词中若干风格与情调。"大"字比"小"字多，主要是因为名胜诗词歌颂的是祖

国的名山大川，诗意磅礴，意境辽阔；“山”字频率高，是由于祖国名山众多，高山奇峰特别能激发诗人的情感；“悲”字出现频率高，也显示了历代诗人面对景点而勾引起对历史上的战乱离异及时过境迁的伤感情绪，以及面对大自然而深感人生短暂的悲伤心情；“人”字的频率较“风”、“花”、“雪”、“月”都高，说明这些诗词并未脱离人生现实。

这种字频统计还可以做得更深入细致一些。例如将这种字频统计限定在某一时期的作品或某一作家的作品的范围之内，并将不同时期和作家的作品字频统计结果进行比较，以分析其不同的风格和特色。

全书的字频统计也具有重要意义。这种字频统计可以表明各个汉字的存在生命力，可以了解哪些汉字是常用字，哪些汉字是一般字，哪些汉字是罕见字。这对于汉语教学、汉字编码等有直接的指导意义。凡是电子出版物的篇幅越大，所得出的统计越有典型意义。当然，不同学科内容的电子出版物的字频统计会出现差异，这是可以理解的。正是这种差异，可以使汉字频率的统计按学科范畴进行，使统计做得更为深入，更有针对性。

与字频统计相似的是，利用电子出版物也容易进行对文句平均长度统计、段落平均长度的统计、各种标点符号使用频率的统计。这些统计对于某些特定的研究也是有帮助的。例如，句子平均长度的统计，有助于了解作家的写作风格是朴实无华还是华丽冗长的。因为后者使用各种形容词较多，而前者则相反。又例如书引号的多少，可以了解作者引用文献多寡。文本中脚注符号使用频率的高低，也可借以了解其注释是否丰富，或引文的多少，从而有助于了解作者写作时掌握有关文献的广度和写作的严肃性。

2. 不同属性词的关联统计综合分析。这是将两个或两个以上属性的词进行有目的联系起来统计分析。比如，在《国共两党关系通史》电子版中，可以将“西安事变”和与其有关的人名进行这种综合，以了解与该事变有关的人物名单。这种名单可以是按频率从高