中国图书馆学报(双月刊)2000 年第 1 期
ZHONGGUO TUSHUGUANXUEBAO

●刘静一 张云新 曹 东 傅 亮

美国联机目录中检索点的多余性和唯一性研究概述

——许红和兰开斯特的合作研究

摘 要 联机目录可大大提高主题检索能力,人们早已达成共识。F·W·兰开斯特等人于 1991 年提出了篇名与分类号更多的是与著作固有的主题发生重合,而不是提供更深层次的检索点。据此,许红和兰开斯特从OCLC 的联机目录中进行提样研究,认为联机目录在检索方面比在提供每本书详尽的主题表达式方面具有更高的区别能力。表 5。图 1。参考文献 6。

关键词 美国 联机目录 检索点 研究概述
分类号 G354.2

ABSTRACT It is generally agreed that online catalogs can enhance the ability of subject search. In 1991, F. W. Lancaster and other people analyzed some characteristics of title, class number and subject searches On this basis, Xu Hong and F. W. Lancaster made a sampling research of the OCLC Online Union Catalog and got some findings. 5 tabs. 1 fig. 6 refs.

KEY WORDS U. S. A. Online catalog Access point
CLASS NUMBER G354.2

1 问题的提出

对于联机目录可大大提高主题检索能力这一点,人们早已形成共识。甚至即使不增加常规记录(如来自目次页或其它信息源的款目),也可仅靠设置其它可检字段来达到增加主题检索点个数(在此,“检索点”一词指的是书目记录中任何一个可表示当前文献主题的项目,如主题号,分类号,或出现在篇名、主题标目和其它位置中的字词)。显然,题名字段和分类号字段中都包含有在主题检索中大有用途的项目。然而,对于一些文献来说,也很可能其篇名及分类号不能增加由主标题无法提供的检索点。举一个绝对假设的例子:如果一个篇名为“鸟”的著作只有一个主标题“鸟”和一个杜威分类号 598,那么它的篇名与分类号就不能增加由主标题无法提供的检索点。

F·W·兰开斯特等人曾提出,篇名与分类号似乎更多的是与著作固有的主标题发生了重合,而不是提供了更深层次的检索点[1]。

从事这一研究的目的是确定在一个典型目录记录中篇名与分类号提供的检索点和主标题已经提供的检索点之间的差异度。引发这一研究的两个主要设想是:首先,在典型目录记录中,有分类号、篇名和主标题提供的检索点呈现出相当大的重复性。其次,在典型目录记录中,分类号、篇名与主标题字段与它们所提供的“唯一”的检索点的个数是完全不同的。在这里,“唯一”的含义是检索点仅出现在 3 个字段其中的一个里。

2 研究的过程和方法

许红和兰开斯特的研究工作是从联机计算机图书馆中心OCLC 的联机联合目录中针对杜威十进分类法的 300 类目(社会科学)、500 类目(自然科学)、600 类目(技术)和 700 类目(艺术)抽取的一组记录样例开始的[2]。

至 1994 年 8 月 10 日,OCLC 的联机联合目录中适合此项研究(即基于所选定的 4 个杜威类目进行的专题研究)的记录已有 734 000 条。考虑到一些随机抽取的记录可能不合研究之用,因此,为保证 4 大类每类中至少抽取到 44 条有效记录,人们在某种程度上按层扩展了样例(即在各类中成比例抽取)。在排除掉不符合本研究参数设置的

记录后(如1990年以前出版的书籍、无主标题及非英语语种书籍),最终的样例包含有205条记录: 300类目58条记录, 500类目46条记录, 600类目46条记录, 700类目55条记录。为简化分析,非英语语种书目记录被删除; 为避免出现编目政策方面存在明显差异的可能性,设置了出版日期范围限定(1990～1994年间出版物),这些变化的编目政策包括较长一段时间内有明显不同的主标题与分类号。

这样就形成了一个3 * 4的因素层面设计格式,它可以用来确定不同主题类目中检索点的重复程度。这可能是由于3个主题字段(篇名、主标题、分类号)与上述4个主题类目有一定的相关性。

此研究思路主要来源于文献[3]和文献[4],其作者一致认为,对于一个3 * 4的涉及重复方法的方差分析(ANOVA),样例的范围可小一些: 4个主要单元中各抽44个记录,共176条。

这一研究还包括对记录样例的篇名、分类号及主标题字段内容的比较。在篇名字段(MARC 245字段)中,还包括其它题名信息; 在主标题字段(MARC 6××字段),如遇到标题比主标题能更好表明出版物类型及形式这一特殊情况,副标题也可作为主题标目。处理分类号字段(MARC 082字段)相对复杂一些,因为分类号需被译成文字(如327・73应译为"外国关系"和"美国"),翻译规则参见DDC的凡例、注释及复分表。

这一研究与以前较早一些时期作者[5,6]的研究所不同的是,许红和兰开斯特(以下简称许氏)不是在字面上而是在含义上进行比较。例如,当涉及到"美国"时,"外国政策"、"外务"与"国际关系"等词的相互关系是如此紧密,以至于不可能极为准确地区分它们。于是为了一切实际目的,将其划作"同义词"。

在"含义"上对检索点进行处理区分,很显然地带有主观色彩,又因为,由于一些实际原因,所有"同义"方面的决定最终将由许氏做出,那么确定她的决定是否能得到大多数人的支持就非常必要了。为此,需要一套健全的程序。

这一有效性测试是建立在许氏已为之进行了同义鉴定和为其3个字段指定了唯一检索点的30条记录的基础之上的。30位来自伊利诺斯大学图书情报学研究生院的研究生自愿参加了此次有效性测试,30条记录与30名学生被随机地编为6个组,这样保证1名学生可检测5条记录,同时1条记录可得到5个人的检测。

每一项目中,1名学生都将获得: A)在OCLC记录中3个字段的内容; B)许氏在这些字段中所做的"同义鉴定"与"唯一检索点"指定; C)许氏在做"同义鉴定"时所遵循的规则(见表1)。学生们要表明他们是否同意这些指定并解释不同意的原因。

表1　用于确定主题检索点同义词的规则

序号	规　　则
1	忽略反映表现方法的词而不是主题内容(如"报告"、"探讨"、"研讨会")。
2	在各种会议记录中,只考虑表明主题的词语,忽略标明会议位置及频率的词语。
3	一个专指的主题概念包含具有直接上位泛指概念。如: 一个题名中含"教育"一词,而主题标目中表明"小学教育",该例中,题名与主题标目中都含有"教育"一词,但主题标目中多含一词"小学"(即"一个特定的年龄群")。
4	在以下情况中,与主题检索点相关的词或短语被看作同义词(因此也可看作相等): (1) 完全一致(如主标目为"鸟",题名也为"鸟"); (2) 缩写(如USA = United States); (3) 通俗用法与正规用法(如stamp collecting= philately); (4) 虽然在单个词层面上为非同义词,但在短语层面上为同义词(如,church music= sacred music= liturgical music,即使church与sacred, liturgical并不同义); (5) 隐含同义(如题名中的"America"与主题中的"USA"相等); (6) 单复数具对等性(如mouse= mice); (7) 标准语与俚语对等(如high fidelity= hi fi); (8) 不同形式的拼写(如catalog= catalogue, online= on-line); (9) 英语本身的不同用法(如railroad= railway); (10) 同字根词(如electroluminescence= electroluminescent, Egypt= Egyptian); (11) 同一历史时期的不同表达方式(如1930's= 1930- 1939= the thirties)。

于是, 我们就可确定对于每组 5 个记录, 该组的学生与许氏所做的决定的一致程度。例如, 对一组 5 条记录, 许氏指定 27 个检索点。所有组的 5 名学生同意其中的 21 个, 5 名中的 4 人同意另外的 6 个检索点("同意"指学生们认为这一检索点与任何一个它所在的记录中的检索点都不同义), 并非期望所有学生都同意许氏的所有认定。

然而, 她的决定有 78% 获得了所有人的支持, 另外有 18% 获得了 80% 的学生的支持。这样的"一致程度"已足以使许氏继续做出此类决定而不必再进一步做这样的验证, 主要是因为验证过程中出现的"不一致意见"源于对许氏所建立的规则的误解。如一个一般性概念或明显或含蓄地存在于一个专指概念中时, 它不能算作具有唯一性, 比如"住宅建筑"中的"建筑"一词, "氰化钠"中的"氰化物"一词。

其实, 即使许氏的决定真的带有主观性也未必不可取, 可以借助适当的工具书(百科全书、词典、同义语词典、术语汇编等)加以转换, 或者如有必要的话, 可以请教本校其他比较熟悉本学科的专业人士。

需要指出的是, 检索点的同义词是由书目记录本身的款目内容决定的, 而不是由款目出处(如分类表)的内容决定。因此, "美国+ 外国关系"、"美国+ 国际关系"、"美国+ 外交关系"在概念含义上非常相近以至于我们不能清晰地区分它们。在某种意义上, 一本关于美国外交政策的书肯定会谈及美国的外交(国际)关系, 尽管这些概念在分类表及主标题中被划定为不同的含义。

3 研究结果

表 2 举例说明了研究获得的结论。该例中抽取了 4 个检索点, 主标题中全包括, 题名字段包含其中的 3 个, 分类号中包含其中的 1 个。

宏观上, 图 1 表明了本研究的主要结论。由于跨主题范围的对比不是这些数据的重点, 所以这里未给出此方面结论。总体来看, 844 个检索点分布在 205 个款目中, 平均每个记录 4. 12 个, 其中, 210 个检索点(24. 88%)同时出现在 3 个字段中; 另外, 有 414 个检索点(49. 05%)仅出现在 1 个字段中; 主标题字段中 209 个(24. 76%); 题名字段 118 个(13. 98%); 分类号字段 87 个(10. 31%)。

表 2 从单一样例记录中获得的结果统计样例

题 名	Efficient masonry house building
主标题	Masonry-Great Britain
	House construction-Great Britain
分类号	693, construction in specific
	Types of materials and for specific purpose
唯一主题词	Masonry
	Houses
	Construction
	Great Britain
题名中出现的主题词	masonry, houses, construction (= building)
主标题中出现的主题词	4 个全有
分类号中出现的主题词	Construction(只有 1 个)

如图 1 所示, 主标题字段单独提供 634 个(75. 12%)检索点, 而题名字段提供了 458 个(54. 27%), 分类号字段提供 406 个(48. 1%)。

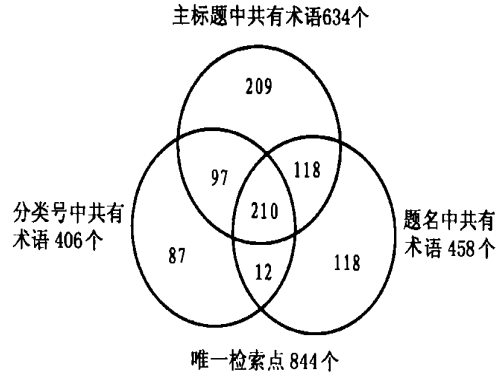

图1 3个主题字段中主题词的重复

处理这些数据的另外一个角度是分析每个字段提供的唯一检索点比例。上文提及, 单一字段提供的唯一检索点共 414 个, 其中, 主标题字段 209 个(50. 48%), 题名字段 118 个(28. 50%), 分类号字段 87 个(21. 02%)。

从图 1 可看出, 主标题 SH 字段与题名 TI 字段重复率最高, 其次是 SH 与分类号 CN 字段。TI 与 CN 字段重复率最低的主要原因是题名提供的检索点专指度高。

表 3 列出了两字段间检索点重复率的方差分析结果。由于每两个字段之间"重复的检索点"的平均数量差异很大($F = 38.30, P < 0.001$), 这使得

我们的第 1 个假设——每一对字段中重复检索点的平均数量差别不大——被排除了。第 2 个假设——3 个字段分别所能提供的检索点数量明显不同——通过对 3 个字段中的唯一检索点进行比较得到验证。

表 3　对两个字段中重复的检索点的单因子的方差分析

来源	自由度	平方总和	平方的平均数	F 值
字段	2	29.08	14.54	38.30
错误	402	152.60	0.38	

注: 这里"错误"是指抽样误差, 自由度是指 $X^2(n)$ (卡方) 分布中自由度 n 的取值, 平方总和是指抽样的平方和, 平方的平均数是指抽样数据的平方和的平均, F 值是指抽样数据的检验数, p 是指显著性水平。

表 4 显示了对 3 个字段提供的唯一检索进行方差分析的结果: 3 个字段提供的检索点个数明显不同($F = 24.29$, $P < 0.001$), 所以无效假设——3 个字段提供数目相近的唯一检索点——被排除。换句话说, 我们可以下这样一个结论: 3 个字段所能提供的唯一检索点数目有显著区别。

表 4　3 个主题字段提供的唯一检索点的单因子的方差分析

来源	自由度	平方总和	平方的平均数	F 值
字段	2	42.13	21.07	24.29
错误	402	384.67	0.78	

表 5 表明每个字段所能提供的检索点数目都与另外两个字段明显不同。

表 5　3 个字段中唯一检索点的两两比较

字　段	重要性层次
CN-TI	$p < 0.05$
CN-SH	$p < 0.001$
TI-SH	$p < 0.001$

4　研究结论

显而易见, 在典型的联机公共目录中唯一的检索点为数不多——稍多于 4 个(本例中为 4.12 个)。本研究中, 联机目录在提供唯一检索点的数目方面并不比卡片目录更能满足图书馆员的期望, 至少在本例中的 4 个主题范围内可如此定论。如果每个记录所能提供的检索点数目可被看作"检索率"的量度的话, 题名仅仅是对主标题检索点的补充, 分类号提供的不同于其它字段的检索点则为数甚少。然而, 在主题检索中, 联机目录比卡片目录的优越之处在于: 联机目录在检索方面(合并不同字段的款目词; 题名提供更大的专指性; 检索有时可用日期、语种或其它标准进行限定)比在提供每本书详尽的主题表达式方面具有更高的区别力。换句话说, 查准率的潜在发展能力大于查全率。

参考文献

1 Lancaster, F. W., et al. Identifing barriers to effective subject access in library catalogs. Library resources & technical services, 1991, 35: 377～391

2 Xu, hong, and Lancaster, F. W., Redundancy and uniqueness of subject access points in online catalogs. Library resources & technical services, 1998, 42: 61～66

3 Cohen, J. Statistical power analysis for the behaviour sciences. New York: Academic Press, 1988

4 Stevens, J. Intermediate statistics: A modern approach. Hillsdale, N. J.: Lawrence Erbaum Associates, 1990

5 Markey, K, and K. Calhoun. Unique words contributed by MARC records with summary and/or contents notes. Proceedings of the American Society for Information Science, 1987, 24: 153～162

6 Frost, C. O. Title words as entry vocabulary to LCSH: Correlation between assigned LCSH terms and derived terms from titles in bibliographic records with implications for subject access in online catalogs. Classification and quarterly 10, nos, 1989, 1&2: 165～179

刘静一　南京政治学院上海分院信息管理系工作。通讯地址: 上海市。邮编 201600。

张云新　曹东　傅亮　同上。

(来稿时间: 1999-07-13。编发者: 翟凤岐)

中国图书馆学报(双月刊)2000 年第 1 期
ZHONGGUO TUSHUGUANXUEBAO

●许儒敬 陶宗宝

继往开来 再铸辉煌

——中国科学院文献情报系统发展概述与展望

摘 要 概述了中国科学院文献情报系统的建立和发展历程,特别是 1978 年以来改革创新和图书馆自动化与网络化的成果,并对今后的发展提出了若干建议。参考文献 6。

关键词 中科院文献情报系统 文献情报事业 专业图书馆建设
分类号 G258.5

ABSTRACT The authors describe the establishment and the development of the system of information services of the Chinese Academy of Sciences, especially the reform, computerization and networking since 1978, and then propose some recommendations for future planning. 6 refs.

KEY WORDS Chinese Academy of Sciences Information services Special libraries
CLASS NUMBER G258.5

1 中国科学院文献情报系统的建立与发展

1950 年建院初期,根据全国自然科学工作者代表会议的建议,中科院于同年 4 月成立了图书管理处,1951 年 2 月图书管理处改名为中国科学院图书馆,由陶孟和副院长兼任馆长。从此科学院的文献情报工作一直受到郭沫若院长等历届院领导的关心和重视。院图书馆成立以后,陆续在上海、兰州、成都、武汉建立了地区图书馆,各研究所也相继建立了所图书馆。而后,科学院成立了编译出版委员会,以加强对图书馆工作的领导。1958 年 9 月,召开了全院第一次图书馆工作会议,会上规定了图书馆的任务,即"为无产阶级政治服务,为科学研究服务,为生产建设服务",开创了为科学研究服务的新局面。1959 年冬,在大连召开了全院第二次图书馆工作会议,会议的中心议题是:如何开展文献参考和书目情报工作。会议促进了图书馆工作向深度和广度发展。此时,全院的图书馆已发展到 200 多个,初步形成了院图书馆、分院图书馆和研究所图书馆的三级图书馆体系,院馆对分院馆和所馆的领导关系改为业务指导关系。十年动乱,百废待兴。1978 年,全国科学大会召开后,科学院图书馆系统乘着大会的东风,走上了图书情报一体化的道路。就在这一年,科学院召开了全院第一次图书情报工作会议,它是科学院文献情报系统发展史上一个重要的里程碑。这次会议充分肯定了图书馆工作在"文化大革命"前 17 年取得的成绩,进一步明确了图书情报工作的性质、方向和任务,提出了"图书情报工作是科学研究工作的一部分,图书情报人员是科学研究人员的一部分",并在全院实行了"图书情报一体化"的管理体制。这些政策观点和发展模式的创新和突破,不但确立了科学院文献情报工作的地位和作用,而且在中国图书情报学术发展史和图书情报事业发展史上有其独特的历史地位。80 年代初起,为了发挥全院文献情报系统的整体优势,加强单位之间的横向联合与合作,先后建立了 22 个文献情报协作网,它们在建立全国基础学科文献情报检索体系和推进全院文献数据库建设中起了重要作用。1985 年以后,院图书馆和各地区图书馆相继改名为文献情报中心,逐步形成了具有科学院特色的图书情报一体化的体系结构。同时,科学院在文献情报人员中实行职称评定和聘任制,评定情报研究成果,调动了文献情报人员的积极性和创造性。1986 年全院第二次文献情报工作会议召开,会议提出了坚持改革,努力创新,加快实现文献情报手段现代化,更有效地为科学研究和国民经济建设服务的指