提高科技信息服务效率和质量的方法

□ 孙卫 / 科学出版传媒股份有限公司 北京 100717

摘要：互联网传播时代的海量信息处理，对于教育、研究和科技工作者来说既是福音又是苦恼。一方面，当前科技文献的种类比传统图书馆能够提供的服务更加丰富；而另一方面，如何剔除随信息量增加的噪声信息，以求获取效率和质量上的平衡则是一个需要攻克的课题。图情服务机构利用图情有序组织的方法，针对海量的资源进行序化的组织和管理；结合教育、研究科技工作者的特点，利用科技信息分析的方法，在充分考虑我国国情的前提下最有效率地整合资源，并结合技术手段来提高科技信息资源使用的效率和质量。这对于中国教育、研究、科技工作者最大化利用科技文献来开展研究和工作，是最好的选择之一。

关键词：文献，知识组织，知识管理，信息处理技术
DOI：10.3772/j.issn.1673—2286.2011.08.009

1 问题的提出

上个世纪九十年代中期开始，互联网与移动通信网进入了公众的视野，在不到20年的时间里已经改变了人们的生活、工作、学习习惯而成为当今最主要的传媒方式。

尽管从1994年美国自然科学基金（NSF）开始数字图书馆的研究计划到1998年开始的第二期研究计划[1]，都期望利用数字图书馆的研究，解决互联网的信息孤岛问题、解决互联网的信息组织问题、改变互联网信息质量良莠不齐的问题，但是，效果并没有达到：①人们在网络上检索和判断有用的结果时依然花费很多时间；②由于自动摘要技术的限制和资源本身质量良莠不齐，很难保证找到的资源的准确性、有价值；③由于使用爬虫发现与获取技术，很多链接关系随时间的变化或者网站商业价值变化而成为无效链接，所以造成了检索到了但是无法获得，或者点击后才知道是无效链接等。

爱思维尔科技（Elsevier Science）公司和谷歌公司（Google）的学术搜索宗旨都是为教育和科研提供学术资源信息服务。而爱思维尔公司是以控制世界上权威的STM期刊、书籍等资源作为服务的基础，谷歌公司则主要是利用爬虫技术发现互联网的内容，以及与各个出版机构合作获取的期刊和书籍的相关元数据信息作为搜索服务的基础。

表1 爱思维尔科技和谷歌学术搜索综合对比

	爱思维尔科技[2]	谷歌学术搜索
资源	STM科技期刊和书籍	各类学术信息元数据及其链接关系
检索	有	有
写作	有	无
评价	强	弱
链接可靠	可靠	不可靠
核心影响力	高	低
使用者	最好的大学和科研机构	自由人群为主
使用效率	高	低
资源质量	高	低
使用成本	高	低

在表1中，谷歌公司的学术搜索引擎和爱思维尔科技的学术研究与服务是不同的，最主要不同是使用成本，爱思维尔科技的使用成本高。使用效率和资源质量表明爱思维尔科技提供的信息服务物有所值，这个就是今天出版机构和互联网信息服务公司依然在进行博弈的价值所在。

通过以上对比，看到两个事实：①互联网信息服务中的噪声，无法帮助教育与科研工作者进行有效率的、高质量的科研活动。②简单的书籍、刊物已经无法满足教育与科研工作的需要，STM出版资源也遇到因质高价高而变成不得已时才被使用的资源。

那么对于科技文献共享，是否有另外一个方法，既降低使用者的成本支出，又可以保证资源的质量和利用的效率呢？

2 资源的种类和组织方式

互联网信息服务系统与出版机构以外的，可以利用对文献的组织和管理的方法、经验、平台，向教育和科研机构提供服务的是图书馆和科技信息服务机构。

表2表明了图书馆科技情报服务机构既有出版机构产生的资源，也有互联网上对应的资源，还有长期的馆藏加工出来的资源。从资源优势上，可以确定其资源比通过互联网检索到的资源要集中、资源的质量优于互联网的发现的大部分资源、资源种类多于出版机构的资源。

根据表3，在图书馆与情报科学单位的绝大部分文献资源是进行过再组织、再加工的，这样的资源的质量优于互联网各个网站处理过的资源，劣于出版机构权威性编辑处理过的资源。

根据表2和表3的对比，是否可以通过图情机构为教育、科研、专业工作者提供更有效的科技信息呢？

图书馆最主要的目标是为了读者使用方便而进行文献的有序化组织和管理，另一个目标是保存与传承文化。在上个世纪90年代初期，美国开始数字图书馆项目研究的目的也是利用图书馆对于文献资源序化的方法与经验，改进互联网资源的杂乱无序的状态；利用图书馆的互操作标准和数据交换的方法与经验， 改变互联网的信息孤岛现象。所以，我们相信利用图书馆与科技信息服务机构对于资源的序化整理和一定的技术服务手段的支持是可以改进科技信息资源发现与使用的效率和提高服务质量的。

表2 资源的种类与来源

资源类型	来源	利用率
书、刊、报（含电子版）	购买、交换、共享	物理利用率低，网络电子版利用率高
中外专利数据	购买、互联网获取	分析利用高，直接利用低
学位论文	缴存、购买	物理和电子利用率高
OA资源	互联网捕获	结合主题、收藏利用率不同
科技报告	收藏、缴存	战略、战术型的报告利用率不同
外文连续出版物数据	利用馆藏加工	检索利用率很高，物理出版物利用率低，电子内容链接或者传递
机构数据	利用馆藏加工	
企业数据	利用互联网捕获、行业协会	
主题词表	利用馆藏加工	
叙词表	利用馆藏加工	

3 网络传播与资源利用的关系

《清华大学图书馆读者利用图书馆方式的调查》[3]一文中，（1）在利用图书馆网络资源上（不同时间段累计），教师占80.69%，研究生占98.7%，本科生占93.96%。（2）没有在图书馆找到资源时，首选互联网，老师占58.12%，研究生占62.73%，说明互联网资源成为图书馆资源的补充。（3）获得所需文献的相关信息途径，79.06%的老师和72.45%的研究生利用文后参考文献，然后才利用文摘索引、搜索引擎。这组调查数据清晰地说明了对于教育和科研工作，图情机构组织和提供的服务是主要的，互联网与搜索引擎是辅助的。而在资源获取上，图书馆科技服务机构的优势更强，因为这些机构是提供科技信息为教育和科研服务的公益单位之一。在学术资源这个前提下，内容丰富的互联网和图书馆处于相当的水平，而在质量可靠、首选利用上，图书馆组织和管理的资源明显占有优势；在查找方便上，图情的服务是优先的，互联网的服务是互补的。

《高校研究人员学术信息资源利用及信息查寻行为调查与分析》[5]一文中，（1）在使用过的电子资源类型中，学位论文资源占37%，电子期刊资源占91%，数据库资源占61%，电子图书资源占31%。由此可见，这四种类型的学术资源很难在互联网上免费获得。（2）对于专业学术资源的熟悉程度中，很熟悉的占28%，比较熟悉的占56%，一般占16%。由此可以看出，在利用专业资源上，针对性强是提高科技信息利用效率和质量的一个因素。（3）从获取资源的途径上看，图书馆网站、专业网站排在2和3位，而图书馆培训排在最后。这说明图情人员想按照图书情报专业的思路提高使用者的信息素养是一件很困难的事情。简单对资源序化还不能彻底提高使用效率，需要符合自然人的行为习惯与文献组织序化原理之间的收敛，才可能找到提高效率的方法。

提高效率和改进质量的关键是如何把自然人的行为习惯与图书情报专业对于资源的组织和管理的方式找到收敛和匹配的方法。

4 收敛与行为匹配的方法

4.1 用户检索词与资源关键词分析计算收敛方法

谷歌是利用词表对于资源进行切分处理，对检索

表3 资源组织与共性

资源类型	主要组织特点	处理方式	备注
书、刊、报等	书目数据（MARC）	人工	规范知识
中外专利数据	专利规范组织	人工	规范知识
学位论文	书目数据（MARC）	人工	规范知识
OA资源	简单数据（DC）	半自动	简单/规范知识
科技报告	书目数据	人工	规范知识
外文连续出版物	部分书目数据	人工	规范知识
机构数据	基于文献	人工	非规范
企业数据	基于文献、互联网、工商注册	人工	非规范、非实时
主题词表	基于一类文献统计特征词汇	半自动	非规范、非实时
叙词表	基于文献抽取、购买等	半自动	新词发现不够、规范不够

词汇与这个词表匹配进行检索和定位的。图书情报的资源组织是按照主题词表的方法进行分析处理的，而主题词表[5]的用、代、属、分、参可以把词汇进行关联。但是，传统的主题词表收集与整理的时效性是比较弱的，主题词表现的是一类文献的共性词汇。所以，可以利用主题词原理进行资源的自动分类处理，基于国家图书馆的分类主题词表已经实现的资源组织管理过程的自动分类，极大地提高了文献资源组织中知识处理的效率。每篇文献给出的关键词，代表了本篇文章的重点概念。

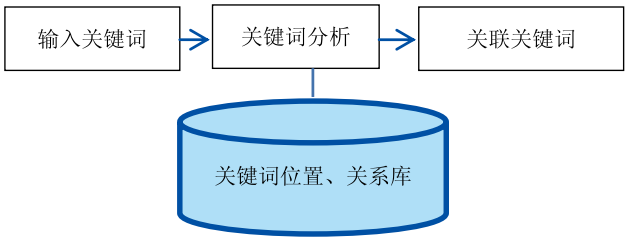

图1 建立文献资源关键词关联分析取得关键词团

在新的方法里，建立关键词位置、关系库。利用输入的检索词，根据关键词库，计算出这个检索词与其他关键词同时出现在资源中的频度，同时出现频度最高的关联关系最强，依次类推，从而根据设定的关键词群团的关键词的数量，构成关键词团。再把这组关键词团发到检索系统中去进行检索，就可以把单个关键词与多个资源关联的关键词进行资源组织关键词的收敛，然后在排序上，先处理与的关系，再处理或的关系。这样既达到组合检索的目的，又不需要使用者一定要学检索理论。这样检索命中结果按照先“与”后“或”的方式排序，可以提高符合使用者愿望的结果，按照资源组织的特点排在前面的位置上。在中国科学技术信息研究所2010年完成的国家科技基础支撑项目的科技评价分析系统中，已经实现了关键词分析，下一步将解决与检索系统的有机结合和自动处理的问题。这个收敛的方法，明显不同于完全个性化的、基于不同角度的相同词汇利用的点击统计，而是利用资源组织的关键词，利用关键词的关联分析，把绝对个性化的输入检索词与资源关键词进行关联，加速针对资源组织的收敛关系，使得检索的命中既符合读者的个性又满足资源组织的共性。

4.2 针对学术资源特点选择排序

在不同的检索结果排序中有很多算法，典型的是按照字母、偏旁部首笔画、Page Rank等。谷歌的搜索引擎采用对于组合检索中，“与”在前，“或”在后，然后是Page Rank的组合排序法。而与、或是检索方法，Page Rank是文献链接方法，没有办法证明这两个方法和学术专业的相关度更强，特别是对于研究者的兴趣的关联关系。而对研究者、科技工作者和学习者，可能关心某个领域资源的权威性，权威最主要的特点是经典性为主，而经典性主要在时间轴和被引次数累计两个维度上表现出来（传统的引文分析，只有引用率）。对于研究者和一定专业的经验者，可能关注是不是核心期刊，或者关心是不是新的为主，那么表现的特点是重要性和实时性。那么，基于研究者、科技工作者和学习者，在利用资源上，在检索结果排序原则上出现了时间、是否是核心、被引三个维度的组合。在北京万方数据的检索体系[6]里，可以让使用者选择按照时间与引用、按照引用、按照核心等条件进行排序选择。如果使用者不选择，就给一个三维加权不同的综合排序方法。其目的就是使得命中的检索结果尽可能地出现在现实的结果集的第一页上。这个算法有效地提高了用户期望的检索结果出现在第一页面的概率。

4.3 提供服务结果的服务方式

传统的信息服务商提供的，一般是检索、原文链接、点击阅读的模式。在面对海量信息的检索服务中，对于命中的结果集的分析判断是专业用户最为头疼的事情。把科技情报查新查证、分析、汇总、报告这些流程组合成资源元数据、检索工具、分类汇总、生成报告过程的工具，最后提供报告，是一种面向开题、查新查证等的科技创新辅助决策系统尝试的新服务模式，对于提高服务的效率和质量是一次有益的探索。

4.3.1 海量的资源

在中国科学技术信息研究所的“科技创新辅助决策系统[7]”中已经有2亿条以上的文摘和元数据，大部分可以连接到原始资源上。

中文期刊元数据、西文期刊元数据、中文会议元数据、外文会议元数据、中文学位元数据、中国专利

元数据、国外专利元数据、科技成果元数据、企业信息、词知识库、作者知识库等构成这个服务系统海量的科技资源相关信息。

4.3.2 五要素关联分析

基于人物、机构、项目、主题、分类五大要素对于检索结果集进行汇聚、统计分析、引文率加权的处理，实现了多种异构资源信息的统一处理的方法。根据表3可以得知，各种资源的组织模式是不同的，那么如果检索平铺，就很难聚合、统计、分析。中国科学技术信息研究所的五要素法是在原有资源组织基础之上先进行汇聚处理。科技活动中最主要的核心都是围绕这五个要素表现的，再利用情报计量学、引文分析等理论对汇聚结果进行统计，达到了发现、汇聚、统计分析、排序的目的。

4.3.3 按照报告框架构成报告内容的主题

传统的科研人员查新查证后的结果是分散的，要分别写对于各个数据库查询的结果然后汇总形成报告。而该系统在特定框架下，把汇聚分析的内容直接填入对应的部分，并给出可视化的分析图，形成主报告结构和内容，极大地方便研究者书写报告和对于主报告进行润色的工作。

该平台建立了利用各种资源（只是利用资源的摘要、元数据）、利用各种工具（检索、五要素聚合、引文分析、报告、可视化），提供信息服务的成果（报告）的崭新信息服务模式，帮助研究者、科技工作者和学习者可以高效率和高质量地进行海量科技信息的查新查证、主题分析、人物分析、项目分析、论文写作分析等。如果使用者需要，资源可以链接到原始资源供应商、图书馆、科技情报服务机构等提供原始资源的服务。

4.4 利用工具判断论文质量

研究成果的重要表现形式之一是科技论文。传统的论文主要靠编辑、专家、导师对于作者论文的审读来保证其学术质量。在没有互联网的时代，由于纸媒传播的周期长、数量有限，很少有作者能阅读到非常全面的各种科技文献。但是，在互联网时代通过搜索引擎，作者可以很方便地发现海量的科技文献。由于计算机大量使用，方便写作的工具大量涌现，可以很方便地摘抄、转译、修改并利用。同时，不少作者利用名人进行第二、第三署名，来提高论文被核心期刊录用的概率。所以，把握研究成果之一的科技论文的质量，提高科研单位的声誉，保障名人的荣誉成为打击学术不端行为的任务之一。目前中国有三种类型的相似性检测系统：清华同方知网公司和北京万方数据公司都开始为编辑部、教育与研究机构的研究生办公室提供检测服务，这两个相似性检测系统主要利用中文科技文献对于学位论文和期刊论文进行相似性检测。这两个系统的推广使用，已经减少了相当一批简单抄袭的作品发表在核心期刊上，研究生毕业论文的抄袭现象也得到控制。

另一个系统是武汉大学基于数十万互联网资源的相似性检测，对于利用互联网写作的相似性检出率很高，对于基于海量的科技文献很难检出，加上互联网测试源的快速增长，获取数百亿资源形成困难，这个系统基本无法发挥作用。

但是，利用工具来辅助提高研究成果的质量得到了编辑部和研究生办公室的认可。

中国科学技术信息研究所正在研究检测相似性论文的指标体系，期望可以更科学地对不同类型的论文、论文的不同结构、标点符号、参考文献、引文进行更规范的检测指南。同时，转译、修改等类型的相似性的算法研究也得到了社科基金的支持。

5 结束语

本文通过对问题的引出，分析了资源的类型、组织方式，分析了互联网传播与资源利用的关系，对于科技信息服务机构利用方法和技术提高科技信息资源的使用效率和保证利用的质量，通过中国科学技术信息研究所和北京万方数据、清华同方知网公司的研究、实验、提供服务的效果，已经证明，在中国这种体制下，利用图情服务机构可以为中国的教育、科学技术研究、创新等提供更有效率、质量更高的科技信息服务。

未来的中国科技信息服务需要在数据挖掘、文本挖掘分析上，提供更多的应用平台；在在线翻译、半自动翻译上为科技人员利用俄文、日文、英文科技资源提供帮助；在多角度的评价指标上帮助科技人员更好地对待科技文献资源的取舍。

参考文献

[1] NSF. Digital Libraries Inittiative –Phase 2 [EB/OL]. [2011-04-26]. http://www.nsf.gov/pubs/1998/nsf9863/nsf9863.htm.
[2] Elesvier Science [EB/OL]. [2011-04-26]. http://www.hub.sciverse.com/action/home/proceed.
[3] 杨毅. 清华大学图书馆读者利用图书馆行为方式的调查[J]. 数字图书馆论坛,2009(1):1-24.
[4] 李晓东. 高校研究人员学术信息资源利用及信息查寻行为的调查与分析——经北京大学图书馆用户调查为例[J]. 数字图书馆论坛,2009(1):25-42.
[5] 中国科学技术信息研究所,国家图书馆. 汉语主题词表[M]. 北京:中国科学技术文献出版社,1980.
[6] 北京万方数据股份有限公司[EB/OL]. [2011-04-26]. http://www.wanfangdata.com.cn/Help/index7.html.
[7] 北京万方数据股份有限公司[EB/OL]. [2011-04-26]. http://stads.wanfangdata.com.cn.

作者简介

孙卫，科学出版传媒股份有限公司数字出版技术总监，中国科学技术信息研究所高级顾问，计算机高级工程师，中国计算机学会高级会员，北京通信学会理事，美国计算机协会会员。目前主要在知识组织、知识挖掘、知识处理技术上进行研究与教学。E-mail: sunw@nlc.gov.cn

The Methodology of Improving the Efficiency and Quality of Providing Services on Information Science and Technology

Sun Wei / Digital Publishing Center, Science Press Ltd., Beijing, 100717

Abstract: At the time of internet, massive information is a gift as well as a headache to the educators, researchers and scientists. On one hand, the kinds of scientific documents available now on the internet are much more than what they can get in a traditional library; on the other hand, it is also a challenge to reduce the noise which is increased along with the quantity of information so as to achieve a balance between efficiency and quality. Library and information institutes in China use the methodology of information sequential organization to manage and organize the massive resources. The method of scientific information analysis is also adopted in accordance with the characteristics of the educators, researchers and scientists to make the most effective integration of information resources in full consideration of the Chinese situation. In addition, technology is used to raise the efficiency and quality of application of the scientific information resources. The combination of all above mentioned methods is the best option for the Chinese educators, researchers and scientists to maximize the utility of scientific document in their work and research.

Keywords: Document, Knowledge organization, Knowledge management, Information processing technology