LIBRARY AND INFORMATION SERVICE
Vol. 49, No. 4, April, 2005

基于 Web 的数字图书馆跨库检索系统的比较研究

张　秋
中国科学院文献情报中心　北京 100080
中国科学院研究生院　北京 100039

〔摘要〕以 Searchlight、NLM Gateway 等 6 个国内外数字图书馆跨库检索系统为对象，从检索性能、检索结果处理、检索效率、资源整合、用户服务、使用管理 6 项功能指标对其进行比较分析，得出现有跨库检索系统大多还处于实践和试用阶段，距离完善的跨库检索系统还有一定距离以及由于用户习惯等众多因素的限制，跨库检索系统并不是对单个检索系统的取代，而是其有益有补充的结论。

〔关键词〕跨库检索系统　数字图书馆　互操作
〔分类号〕G250.76

A Comparative Study of Web-based Digital Libraries' Cross-database Integrated Search Systems

Zhang Qiu
Library of Chinese Academy of Sciences, Beijing 100080
Graduate School of Chinese Academy of Sciences, Beijing 100039

〔**Abstract**〕This paper analyzes six web - based cross - databases searching systems in and outside of China. Through the compare of the above six systems in the function of search function, search results disposal, search efficiency, information integration, patron service and management function, the author concludes that most of the cross - database search system are still at the trial stage, there still a lot of way to go to the ideal phrase. Besides, the cross - databases searching system is the complement rather than the substitute of the single system.

〔**Keywords**〕cross-database search system　digital library　interoperation

1　引　言

随着数字图书馆建设的蓬勃发展，图书馆信息资源的种类和数量越来越多。然而，由于不同数据库往往拥有不同的检索界面和使用方式，用户在进入各个数据库时需要进行不同的身份认证。频繁的登陆和退出，加之因对不同系统的陌生而带来的恐惧，使得即使是那些对于图书馆资源十分熟悉的用户，在面对数字图书馆这个信息海洋时也望而却步。因此，图书馆大量的信息资源并没有像图书馆员所预想的那样得到充分和有效的利用。国外的一项调查报告认为，虽然大部分信息用户已经习惯于通过网络来检索信息资源，但是越来越多的信息用户倾向于将搜索引擎而不是图书馆作为研究工作的起点[1-2]。因此，如何准确选择数据库，减轻用户学习与操作的负担，以及如何有效利用多个数据库的集成资源与检索能力，从而保证数字图书馆已经拥有的分布式、异构型的信息资源得到充分和有效的利用，已成为数字图书馆建设进一步优化和发展的既重要又关键的问题。正是为了解决这一问题，数字图书馆建设者设计和开发了基于 Web 的跨库检索系统。

数字图书馆基于 Web 的跨库检索系统的建设实践起源于 20 世纪 90 年代末。1998 年，加利福尼亚大学圣地亚哥分校开发了可以同时检索 25 个科学数据库的跨库检索系统，名为 Database Advisor[3]，现在已升级为加利福尼亚数字图书馆(CDL)的 Searchlight 系统。除此之外，国内外其他许多图书馆都开发运行了基于 Web 的跨库检索系统，国外的如美国国家医学图书馆(The National Library of Medicine, NLM)的 Gateway、洛斯阿拉莫斯国家实验室研究图书馆的 Flashpoint、亚历桑娜大学图书馆的 Multi - SEARCH、沃里克大学图书馆(University of Warwick Library)的 Cross - database search[4]；国内的如 CSDL 跨库集成检索系统(以下简称"CSDL 系统")、CNKI 源数据库跨库检索系统(以下简称"CNKI 系统")以及清华大学图书馆跨库检索系统、西安交通大学图书

收稿日期：2004 - 05 - 27

馆跨库检索系统(以下简称"交大系统")等。本文拟对Searchlight、NLM Gateway、Flashpoint、CSDL系统、CNKI系统以及交大系统这六种跨库检索系统进行比较分析,从而深入地了解国内外对此主题的研究进展。

2 跨库检索系统的功能指标

一个理想的跨库检索系统能够完全整合和替代多个单个数据库的检索系统,其功能可从以下几个方面来衡量:

2.1 检索性能

检索性能分为5个方面:①检索方式,包括提问检索(简单检索/高级检索)、浏览查找(可按出版物名称/受控词表/分类法/出版商/著者单位)、自然语言检索和受控语言检索等;②检索字段,包括著者、著者单位、文献篇名、出版物名称、文摘、主题词/关键词、分类号、ISSN/ISBN和材料识别号等;③检索技术,包括布尔检索、组配检索、截词检索、词根检索、位置算符与优先算符检索、相关检索以及引文检索等;④检索限定,包括对出版时间、文献类型、语言文字、处理类型等的限定;⑤检索界面,包括帮助信息、工具条、界面设计和检索步骤等。

2.2 检索结果处理

检索结果分为3个方面:①检索结果调整,包括检索提问修改、二次检索、合并检索和相关检索等;②检索结果显示,包括排序方式、显示格式以及每屏显示量等;③检索结果输出,包括检索结果标记方式、最大输出量以及输出方式等。

2.3 检索效率

检索效率指跨库检索系统具有卓越的检全率和检准率、较短的响应时间、较快的检索速度等。

2.4 资源整合

资源整合指跨库检索系统可以:①资源整合,即对同一公司的不同数据库、不同公司的不同数据库、外部数据库等资源进行整合;②检索结果去重,即对来自分布、异构数据库的检索结果进行去重处理;③超链接,即对同一数据库的不同记录、同一平台的不同数据库、因特网资源、全文数据库、馆藏OPAC以及文献传递服务等的超链接。

2.5 用户服务

用户服务指跨库检索选题具有用户界面定制、创建个人账户、邮件定题服务、个人期刊列表、存储检索式和检索记录等服务功能。

2.6 使用管理

使用管理指跨库检索系统具有使用管理、用户管理、统计报表、数据更新等管理功能。

3 国内外6种跨库检索系统比较研究

3.1 基本情况

- Searchlight[5]:前身是加利福尼亚大学圣地亚哥分校于1998年开发的Database Advisor,是加利福尼亚数字图书馆(CDL)的重点建设项目之一,于2000年1月开始试运行。
- NLM Gateway[6]:由美国生物医学资讯中心(LHNCBC)和美国国家医学图书馆(NLM)共同创建,是一个基于Web的、可由用户对NLM所拥有的信息资源(或数据库)进行一站式检索的系统。
- Flashpoint[7-8]:属于美国洛斯阿拉莫斯国家实验室研究图书馆(The Research Library of the Los Alamos National Laboratory)"无墙图书馆"项目的一个子项目。由于受知识产权的限制,Flashpoint有用户权限限制,需要有用户名和密码才能进入。
- CSDL系统[9]:是中国国家科学数字图书馆(CSDL)项目的子项目,由CSDL自行开发,2003年3月投入试用。
- CNKI系统[10]:是CNKI工程的子项目,由CNKI开发建设,2003年投入试用。
- 交大系统[11]:由西安交通大学组织专人开发,2003年5月投入试用,通过IP进行权限控制。

需要说明的是,由于Flashpoint和西安交通大学图书馆的跨库检索系统有用户权限限制,笔者无法对这两者的一些功能指标进行深入考察,只能根据有关的资料进行研究。

3.2 功能指标的比较

3.2.1 检索性能 Searchlight系统默认的检索字段为关键词检索,该系统的设计目标为只要用户输入关键词,系统即自动在每种资源的题名、文摘及主题中对该词进行检索。但事实上,这只是一种理想化的状态。因为CDL所整合的大多数网络信息资源的检索方式是对每个字段进行单独检索,而不是对同一主题进行多重检索。因此,每一个检索词只需要在一个字段中进行检索而不需要在所有的字段中进行检索,即可对检索进行限定(refine your search)。

NLM Gateway的检索分为简单检索和限制检索两种方式,默认方式为简单检索。在限制检索中,用户可对检索资源类型(期刊索引、图书/连续出版物/视听资料、消费者健康、会议文摘以及其他资源)、资源内容(AIDS、生物伦理学、医学史、1996年以前的MEDLINE、隔离生命科学)、每页显示结果数目、语言、出版日期等进行限定。

Flashpoint提供快速检索和高级检索两种检索方式。快速检索方式只提供一个输入框,用户可以从作者、题名/主题/摘要、来源/期刊三种途径进行检索,时间预先设定为近两年,每次检索出50条记录。高级检索方式提供多个输入框,可用布尔逻辑符and、or和not、OR和NOT等进行限定。相对快速检索而言,高级检索增加了会议、机构和报告序号等字段,检索结果的数目也有所增加。

CSDL系统提供篇名、作者、文摘以及全字段检索,其中全字段检索是指根据不同的数据库或指全文检索、或指全标

引字段检索,系统默认检索途径是"全字段"检索。同时,CSDL跨库集成检索系统还支持多词检索和短语检索:①多词检索方面,单词间用空格分隔。检索时①凡包含全部检索单词的文献即为命中文献。②短语检索方面,对于由双引号引起来的多个单词,系统将作为短语进行匹配,如:"digital library"。

CNKI系统提供题名、作者、机构、关键词、摘要、全文、引文等字段检索,可进行二次检索。

交大系统提供题名、作者、摘要、关键词、刊名、ISBN、主题、全文、机构等字段的检索,提供简单检索和高级检索,高级检索中不同的检索词之间采用的是and、or或not等布尔逻辑算符。

在检索帮助方面,Searchlight、NLM Gateway、CNKI等设有检索帮助。

3.2.2 *检索结果处理* Searchlight的检索结果依据资源类型排序,包括图书、期刊索引、电子期刊、电子文本和文件、参考资源、网络资源等。结果页面的右栏显示的是数据库名称,左栏为检索结果数量。右栏数据库名称分为两种:一种有符号"*",一种没有符号"*"。有符号"*"意味着通过链接进入后,用户需要对该数据库重新进行检索;没有符号"*"意味着用户可以直接浏览结果,不必对数据库重新检索。点击左栏结果数量,即可打开新窗口浏览检索结果。

NLM把资源分为期刊索引、图书/连续出版物/视听资料、消费者健康、会议文摘及其他,检索结果按资源类型进行排序,用户可以下载或打印检索结果。

CSDL可按相关度、题名、作者排序,用户可以直接查看检索结果。

CNKI系统检索结果的排列方式有相关度排序、更新日期排序和无排序三种。其中,相关度排序是按照检索式和检索项的相关程度由高到低排序;更新日期排序是按照上网的时间顺序排列检索结果;在概览区上方有数据库切换功能,可以实现在不同数据库检索结果之间的切换。交大系统有按日期最近和按相关度两种排序方式。

3.2.3 *检索效率* 在检索等待时间上,Searchlight的用户可对检索等待时间(timeout settings)进行限定,时间值为0.5分钟到5分钟之间,可以是0.5分、1分、1.5分、2分、2.5分、3分、4分、5分,默认值为1分。

CSDL系统的用户可对单库检索时间进行限定。

3.2.4 *资源整合功能* CNKI系统是对同一公司(清华同方)的不同数据库进行跨库检索,其检索源包括《中国期刊全文数据库》、《中国优秀博硕士学位论文全文数据库》、《中国重要会议论文全文数据库》和《中国重要报纸全文数据库》等。

其他的几个跨库检索系统都是对不同公司的不同数据库和外部数据库进行跨库检索。Searchlight提供包括文摘/索引数据库、图书馆目录、网站和其他类型资源的一站式检索,并把这些资源分为自然科学和工程技术、社会科学和人文科学两大类。NLM Gateway将资源分为期刊索引、图书/连续出版物/视听资料、消费者健康、会议文摘及其他,资源整合功能很强,用户点击相应类型的源,即可进入。

CSDL系统可以跨库检索的数据源包括购买的9个全文数据库、文摘索引数据库以及众多图书馆的OPAC资源等;交大系统提供对它所购买的全文数据库、专题数据库、二次文献数据库、特种文献数据库4类数据库的跨库检索。在这些跨库检索系统中,均有对于数据库及相应服务的超链接,但基本上还没有"检索结果去重"这一功能。

3.2.5 *用户服务* 在这6个跨库检索系统中,尤以NLM Gateway和CNKI系统的服务功能较有特色,例如NLM Gateway系统具有以下功能:

- 术语发现(find terms):为使检索更为精确,用户可以使用"Find Terms"功能查找术语的定义、同义词以及相关术语,该功能是通过依托NLM的医学图书馆标题表(NLM's MeSH)和UMLS元分类词表(UMLS Metathesaurus)查找术语的定义以及相关术语而实现的。
- 检索历史(history):可对用户最近的25个检索词予以保留,用户可对这些检索历史进行回溯,也可对这些检索进行合并或修改。
- 检索锁定(locker):用于存储用户想要打印、下载或者预定的检索项目。使用该功能时,首先选中想要锁定的条目,再点击选中条目旁边的复选框,最后点击"Locker"按钮,即完成了锁定过程。存储条目最多为500条。
- 检索定制(preference)功能:用户可对NML Gateway的使用界面。例如检索限制、结果显示方式、下载选择及检索锁定等进行定制。用户也可修改密码将想要看到的结果记录元素进行定制。这些定制既可应用于正在进行的检索过程,也可应用于将来的检索进程。需要注意的是,检索锁定和检索定制功能只有那些拥有用户名和密码的用户才可以使用。对于检索结果,用户可采取下载、打印、电子邮件发送等方式予以保存。

CNKI系统特有的是"导航"功能,它将CNKI源数据库分为9大专辑、126个专题数据库,可以通过导航进行分类检索,逐级细化,直到专题层。也可以细化导航,缩小检索范围,提高查准的比例。

3.2.6 *管理功能* Searchlight有用户调查(searchlight survey)功能,可通过匿名调查的方式对用户了解获知Searchlight的途径、使用效果(新资源发现、二次检索效果、检索结果处理)及用户检索行为进行调查,以便进一步改进Searchlight系统。

有使用权限限制的是Flashpoint和交大系统。Flashpoint从2000年4月推出以来,经历了多个版本的演变和升级(见

表1)。

表1 Flashpoint的版本沿革

版本	时间	内 容
V1.0	2000年4月	快速检索模式,可跨8个数据库检索
V1.1	2000年6月	增加高级检索模式
V1.2	2000年8月	添加Nuclear Science Abstracts数据库,增加数据库说明和提供使用检索技巧
V1.3	2000年10月	增加外部用户检索入口,改进了作者字段识别,增加了临近检索("near" operator)
V1.4	2001年1月	添加了图书馆目录,高级检索中增加题名检索,扩展了作者帮助(author help)
V1.5	2001年9月	重写面向对象的代码
V2.0	2001年12月	添加MathSciNet,这是第一个通过Flashpoint可以检索到的非本地存取的数据库;2002年9月添加PubMed
V3.0	2002年12月	增加数据库选择界面,LANL过滤器;改善了用户界面;2003年8月添加ISI Proceedings数据库

4 几点思考

通过对上述大多处于试用或实验阶段的6个跨库检索系统的比较,我们不难得出以下结论:

4.1 从跨库检索系统的产生来看

基于Web的跨库检索系统的出现是数字图书馆建设发展到一定阶段的产物。数字图书馆的跨库检索系统出现在信息资源建设之后,是在数字图书馆已经拥有多种信息资源的情况下产生的。由于国外的数字图书馆建设早于国内的数字图书馆建设,所以在跨库检索系统方面国外起步也较早。

4.2 从跨库检索系统的效果看

虽然国内外关于跨库检索系统的实践正在蓬勃开展,但从现有的跨库检索系统来看,距离理想的跨库检索系统还有一定的距离。

首先,目前基于Web的数字图书馆跨库检索系统大多还处未定型,于实践和试用阶段,依旧存在这样或那样的问题。例如Searchlight系统就有以下问题:①如果一个资源改变它的底层结构或者是检索界面,对其检索策略将失效;②有时将检索请求发至检索资源会不成功;③如果资源给出检索时间过长的警告信息,Searchlight将不再继续反应,检索终止;④如果检索词过于宽泛,或者各检索资源采取的字段不同,会出现大量的不相关结果(extraneous/irrelevant results)。对一些资源进行关键词检索、全文检索也会产生大量的不相关资源。

其次,目前已有的跨库检索系统实现跨库检索基本上都是基于Z39.50进行跨库检索的模式,例如Searchlight中的大多数数据库都是通过Z39.50界面进行检索的,包括ABI/Inform等29个数据库等。相较于其他的跨库检索模式,即基于元搜索引擎的跨库检索、基于XML的跨库检索和基于WebService的跨库检索,基于Z39.50的跨库检索模式查询准确度较高,因此运用基于这种模式建立跨库检索系统是数字图书馆的优势。但是其他3种检索模式也各具优点和特色,尤其是基于Webserivce的跨库检索目前被认为是实现跨库检索的最完美的一种模式[12]。因此,借鉴其他模式的优点,也是数字图书馆构建跨库检索系统必需注意的。

4.3 对跨库检索系统作用的再认识

我们已经知道,一个理想的跨库检索系统最终要达到的功能就是将众多的数据库整合在一起,完全替代单一的数据库检索系统。尽管数字图书馆跨库检索系统具有便利、有效、易用等优势,但是由于图书馆信息用户认为使用单个的数据库检索系统检索更加有效以及用户检索习惯的约束等原因,用户更加倾向于使用单个数据库检索,而不是跨库系统[13]。正如元搜索引擎的出现和发展并未替代源搜索引擎一样,基于Web的数字图书馆跨库检索系统也不可能完全取代单个的数据库检索系统。因此,数字图书馆跨库检索系统不是对单个检索系统的取代,而是对其的有益补充。在实际使用中,跨库检索系统往往起到帮助用户发现资源、选择资源的作用,它在以下情景中尤其适用:

- 用户对图书馆信息资源不熟悉,不知道到底使用哪一信息资源合适。
- 作为图书馆资源的知识发现工具,通过跨库检索系统可以概览数字图书馆在线信息资源的大致情况。
- 经由跨库检索系统访问具体的信息资源系统。

参考文献:

1 Jackson M E, Preece B G, Peters T A. Consortia and the portal challenge. Journal of Academic Librarianship, 2002,28(3):160-162

2 Thomas S E. Abundance, attention, and access: of portals and catalogs. [2004-05-20]. http://www.arl.org/newsltr/212/portal.html

3 Mischo W H. Library portals, simultaneous search, and full-text linking technologies: University of Illinois at Urbana-Champaign. Science & Technology Libraries, 2001,20(2/3):133-147

4 Tennant R. Digital libraries-cross-database search: one-stop shopping. [2004-05-20]. http://www.libraryjournal.com/article/CA170458

5 [2004-05-20]. Searchlight. http://searchlight.cdlib.org/cgi-bin/searchlight

6 [2004-05-20]. Gateway. http://gateway.nlm.nih.gov/gw/Cmd

7 [2004-05-20]. Flashpoint. http://lib-www.lanl.gov/

8 Mahoney D, Giacomo M D. Flashpoint @ LANL.gov: a simple smart search interface. [2004-05-20]. http://www.istl.org/istl/01-summer/article2.html

9 CSDL跨库集成检索系统. [2004-05-20]. http://159.226.100.20:8080/metasearch/jsp/index.jsp

(下转第63页)

的,因此,获取这些指标需要相关人员有一定的学科背景、数据库知识和使用经验,并需经过多次试用、反复使用检索系统才能得到。

例如:河南HALIS管理中心在组织中文财经类数据库团购时,对国研网、中经网、中国资讯行、中宏网4个数据库试用后得到了如表2所示的定性评价结果。这些指标中有些具有客观性,可直接获取,如布尔逻辑检索。有些指标需要相关人员凭个人知识和经验进行综合考虑及主观判断后才能得到,如数据库特色等。

3.3 半定量指标的获取方法

半定量指标不能简单地通过定性或定量方法直接获取,需要经过调查问卷、统计、分析才能得到一个量值。为此,首先需要设计调查问卷,问卷应包括数字馆藏服务中可以通过此方法获取指标的具体内容,包括一般数字馆藏服务的满意度、特定数字馆藏服务的满意度、数字馆藏用户培训的满意度以及对用户帮助服务、图书馆计算机工作站和其他相关设备提供情况的满意度等[15]。调查问卷可采用三种方式:在图书馆现场填表、网上发布调查表、专家征询表。对每一项内容的评分不能简单地设为优、较好、一般、差,而应该进行详细分级记分(见表3)。问卷收集后进行统计、分析和数据处理,进而得出可以比较的指标的相对量值。

表3 用户计分标准

差	一般	优
1、2、3	4、5、6	7、8、9

每项数字馆藏服务的用户平均满意度可用A/B公式表示。A表示用户给出的每项数字馆藏服务期望值的总和,B表示参与该项数字馆藏服务调查的总人数。

用户期望虽然是一个非常重要的因素,但用户的选择有较强的主观性。如果用户对某一项服务的期望很低,就很容易表现出较高的满意度。如果用户对某项服务的期望很高,那么就容易表现出不满意。因此,在处理和分析数据时应该把这些因素考虑进去。

随着图书馆数字化进程的加快和数字图书馆的发展,馆藏中数字资源的比例将越来越大,为了使数字馆藏与传统馆藏的配置更加合理,充分发挥其资源效益,最大限度地满足用户需要,对数字馆藏进行评价研究是十分必要的,我们认为有关数字馆藏服务绩效、使用及数字馆藏保存功能的评价指标宜采用定量方法获取;有关数字馆藏内容、结构等的评价指标宜通过定性方法获取;而关于数字馆藏用户满意度和存取能力的评价指标宜运用半定量方法获取的观点,仍需要通过进一步的应用评价来检验。

参考文献:

1 张 咏. 网络信息资源评价的方法及指标. [2003-07-11]. http://www.chinalibs.net
2 盛小平. 数字图书馆馆藏评价. 图书情报工作,2003(5):40-43
3 肖希明. 网络环境下的馆藏评价标准. 中国图书馆学报,2002(5):21-24
4 ISO. Information and Documentation-International Library Statistics. In:ISO27892003(E). Switzerland,ISO27892003(E):35-43
5 Poll R. Performance measures for library networked services and resources. The Electronic Library,2001,19(5):307-314
6 Blixrud J C. Measures for electronic use:The ARL E-Metric Project. [2004-03-01]. http://www.arl.org/stats/newmeas/emetrics/index.html
7 EQUINOX library performance measurement and quality management system performance indicators for electronic library services. [2004-09-15]. http://equinox.dcu.ie/reports/pilist.html
8 刘雁书等. 网络信息资源评价指标体系及可获取性研究. 情报杂志,2002(6):10-12
9 胡昌平. 信息管理科学导论. 北京:高等教育出版社,2001:85
10 [2004-08-18]. http://ultra2.lib.tsinghua.edu.cn/cpx/plsql/cust_usage.usage_by_month
11 期刊引文报告数据库. [2004-09-04]. http://isi4.isiknowledge.com/portal.cgi? DestApp=JCR&Func=Frame
12 元智大学图书馆馆藏发展政策. [2004-09-04]. http://www.yzu.edu.tw/library/orienation/law/cdp.htm
13 沈继武,肖希明. 文献资源建设. 武汉:武汉大学出版社,1991
14 陈光祚. 计算机检索导论. 北京:书目文献出版社,1993
15 [2004-07-15]. http://www.arl.org/libqual/events

〔作者简介〕张宏玲,女,1970年生,馆员,硕士研究生,发表论文多篇。
索传军,男,1964年生,研究馆员,博士,硕士生导师,系主任,发表论文40余篇,出版著作4部。

(上接第91页)

10 CNKI源数据库跨库检索系统. [2004-05-20]. http://www.cnki.net/index.htm
11 西安交通大学图书馆跨库检索系统. [2004-05-20]. http://202.117.24.50:8080/interpub/index1.html
12 郭少友,Web环境下分布式信息检索模式. 情报科学,2003(4):632-635
13 Park S Y. Usability, user preferences, effectiveness, and user behaviors when searching individual and integrated full-text databases: implications for digital libraries. Journal of the American Society for Information Science,2000,51(5):456-468

〔作者简介〕张 秋,女,1978年生,博士研究生,发表论文7篇。