Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

图情资源发现系统的研究与实现*

王灏 张正锋 冯巍 / 北京万方数据股份有限公司 北京 100038

摘要：资源发现系统是一种面向图书情报领域的垂直搜索引擎。文章从图情机构馆藏和服务的问题入手，介绍了资源发现系统的概念和技术路线，并以“中国学术搜索网”为例，介绍了该平台在资源发现服务中的总体设计、关键技术、核心功能和服务方式，旨在帮助图书馆等单位在产品技术选型和建设过程中提供可能的参考依据。

关键词：资源发现系统，发现服务，搜索引擎，云计算，软件即服务
DOI：10.3772/j.issn.1673—2286.2013.06.009

1 图情机构面临的问题

图情服务机构从纸本馆藏转向电子馆藏，数字资源建设已经成为非常重要的工作。然而随着图书馆用户阅读习惯的改变以及互联网信息搜索产品如Google学术、百度文库等迅速发展，图书馆提供的电子资源查询系统，已经开始为很多新生代用户所忽略。业界甚至已经开始有“图书馆消亡”的预测[1]。

高质量的、来源可靠的优质资源，专业的学科服务人员，专业的信息服务经验和数据组织管理能力，是图情服务机构的优势。从根本上来说，图书馆等机构面临的现实问题是如何能够充分利用各类资源，提升服务水平，让图书馆的投资回报率最大化，让高质量的学术资源真正成为用户的科研助手和工具。

为了实现更有效的馆藏资源的揭示和利用，馆藏资源的整合是必经之路，搜索则是资源整合后的第一应用。在5-10年前，图书馆普遍采用的是虚拟资源整合技术[2]为主导的联邦检索（跨库检索）系统。

联邦检索需要分别向各个数据库系统提交检索词，从不同的异构数据库获取检索结果后，进行整理排序并展示给用户。由于系统的检索速度和检索效果受制于每一个目标资源的索引方式、检索算法、网络环境等因素，系统从各数据库获取检索结果的速度不一，因此不能很好地实现全部结果的相关性排序、查重和归并。所以它只能在一定程度上解决数字资源一站式检索的问题，在用户的深层次需求上存在难以克服的缺陷，且只能整合本馆资源，对馆外资源的揭示是不可能的。因此用户实际使用效果和反馈并不好。

2 资源发现系统

资源发现系统则是站在元数据层面。

随着云计算、云服务的普及以及图书馆馆藏观念的转变，越来越多的图书馆开始接受公共数据的云端存储，开始逐步放弃原有虚拟整合技术，选择元数据整合方案来实现一站式检索服务。

资源发现系统是对海量的来自异构资源的元数据通过抽取、映射、收割、导入等手段进行预收集，并通过标准的索引方式进行加工，形成统一的元数据索引，以搜索引擎的应用形式向终端用户提供基于本地分布或者远程中心平台的统一检索和服务的系统[3]。资源发现系统基于庞大的元数据，将图书馆关注的各种数据类型资源实现统一存储、统一索引、统一检索、统一展示，从而实现对资源的发现和获取。

资源发现系统可以说是搜索引擎技术在图书馆领域的高级应用。相对于百度、Google等通用搜索引擎的网页搜索，图情资源发现系统是锁定了信息搜索范围的垂直搜索引擎，它针对图书馆馆员和读者的资源管理需求和文献查询需求提供特定的信息输出和相关服务，因为具有图书情报检索领域的专业色

* 本文系国家高科技发展计划（863计划）“云计算一期”重大专项课题“以科技文献为主的搜索引擎研制”子课题（编号：2011AA01A206）成果之一。

彩，相比较通用搜索引擎的海量信息无序化，资源发现系统会更加专注、具体和深入。

由于大规模分布式检索技术以及云计算平台的发展和应用，发现服务系统能够借助更强大的平台运算优势，绑定图书馆需要的元数据仓储，整合各种图书馆资源，包括内部的、外部的、纸质的、电子的、自有的、许可的以及可自由获取的数据源，使用统一标引的数据格式，通过链接解析器链接到全文，并提供统一的界面实现分面筛选和高级检索功能。

资源发现系统从根本上改变了联邦检索系统运行速度慢、返回结果有限的缺点，又进一步实现了资源的深度揭示和融合，在检索范围、检索速度和检索结果质量等方面都有很大程度的改善。

3 资源发现系统的设计

万方数据公司因为承担了搜索引擎相关的国家863重大课题，正在开展该课题的应用系统构建，推出了“中国学术搜索网”平台（以下称为“学术搜索网”），下面笔者将从数据来源、关键技术、核心功能和服务方式等对学术搜索网的资源发现服务作详细的介绍。

3.1 数据来源

学术搜索网的核心是它所覆盖的数据。正如前文所述，发现系统依托的海量数据仓储并非从互联网上抓取的网页信息，它对于信息的规范性、完整性以及正确性都有相应的要求，否则无法满足图书馆学术研究的基本需要。

学术搜索网的元数据仓储的数据来源主要是三个方面：

一是直接获得数据资源供应商授权的元数据，这种来源的数据能够充分保证数据的完整性、时效性和稳定性。同时，学术搜索网的元数据包含大量内部数据加工人员和外部合作机构长期建设的二次文献数据。这些数据经过专业人员的标引，数据质量高。

二是采用元数据收割的方式。学术搜索网基于OAI-PMH协议及其他网络访问协议的元数据收割技术，收集网络开放资源以及未授权的第三方元数据。

三是如何实现图书馆内部资源的整合，这是资源发现系统要覆盖的重要内容。这种情况需要在系统部署实施时对本地资源的数据进行收割和上载，并定期更新。例如将图书馆的馆藏目录MARC数据加以映射，使其转换成标准化的元数据后进行上载。实在不能进行映射收割的可通过人工导入等方式生成相应索引，上传到厂商的中心索引库或图书馆本地的索引库。

3.2 技术框架

学术搜索网的发现服务的核心模块分为索引服务、检索服务、定位服务。系统的总体架构和流程如图1。

1）索引服务

与通用的搜索引擎类似，学术搜索网的索引是为了实现词语和文档之间的对照关系的数据结构。

为了适应不同的检索场景，学术搜索网实现了多种索引方式：

（1）对异构数据类型抽象建模，利用“知识获取五要素”[4]的数据组织理论，从“主题、学科、人物、机构、基金”建立通用的索引项，实现异构数据的统一检索问题；

（2）支持某一个数据字段的多索引方式，如对主题字段实现“按字”索引或“按词”索引，实现了完全精确或模糊检索支持。

（3）支持对多个字段的联合索引，即将若干同类字段合并索引，如标题、摘要和关键词合并为主题索引。

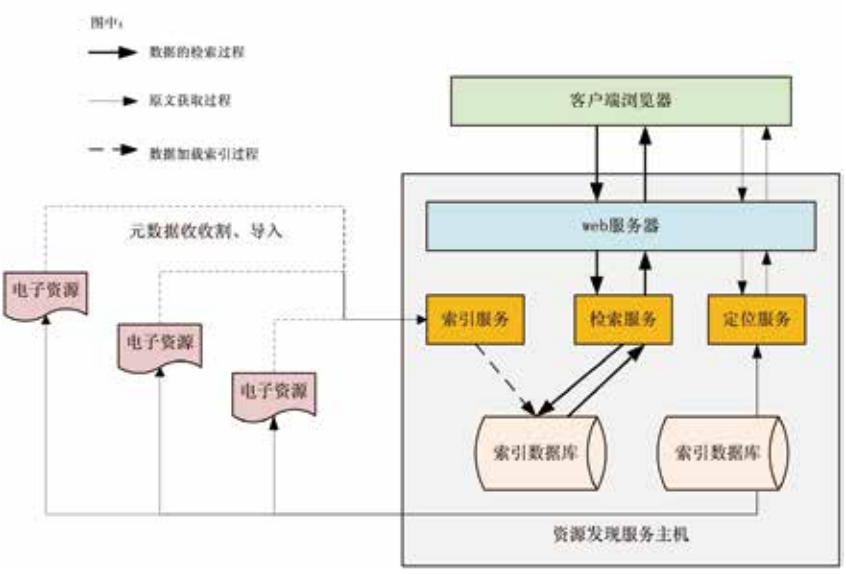

图1 资源发现系统架构

当系统面临巨大的文档集合时，索引文件也会迅速膨胀，靠单机往往难以承担重任。此时的索引方案是将整个文档集合划分为若干子集，建立分布式集群，即每台机器维护整个索引的一部分，由多台机器共同完成索引的建立和对检索的响应。

2）检索服务

图2是分布式索引方案对用户的检索响应过程，检索分发服务器接收到用户的检索请求后，会以广播模式发送给所有的索引服务器，每个索引服务器会负责本服务器内的文档子集的索引维护和检索响应，当索引服务器接收到检索请求后，会计算相关文档，将得分最高的K个文档返回，检索分发服务器将综合各个索引服务器的检索结果，合并完成后将利用某种特定的打分规则对文档进行排序，得分最高的M个文档作为最终的结果返回给用户。

打分规则是资源发现系统中最重要的部分，即检索评分算法。

学术搜索网采用的是基于文档内容与用户检索需求的匹配程度、文档的学术价值以及用户的信息背景的综合评分算法。

文档内容与用户检索需求的匹配程度包括用户检索内容的出现频率、位置权重等。

文档的学术价值包括发文期刊的影响因子、被引次数、论文作者影响力等。

用户的信息背景包括检索和浏览历史等。

3）定位服务

定位服务解决的是用户通过检索结果返回某个资源数据库系统找到原文的问题。

由于数据已经被资源发现厂商进行了清洗和排重，而图书馆购买

	文档1	文档2	文档3	文档4	文档5
词语1	√			√	
词语2		√	√		
词语3	√	√		√	
词语4		√			√
词语5	√		√		√
词语6			√	√	

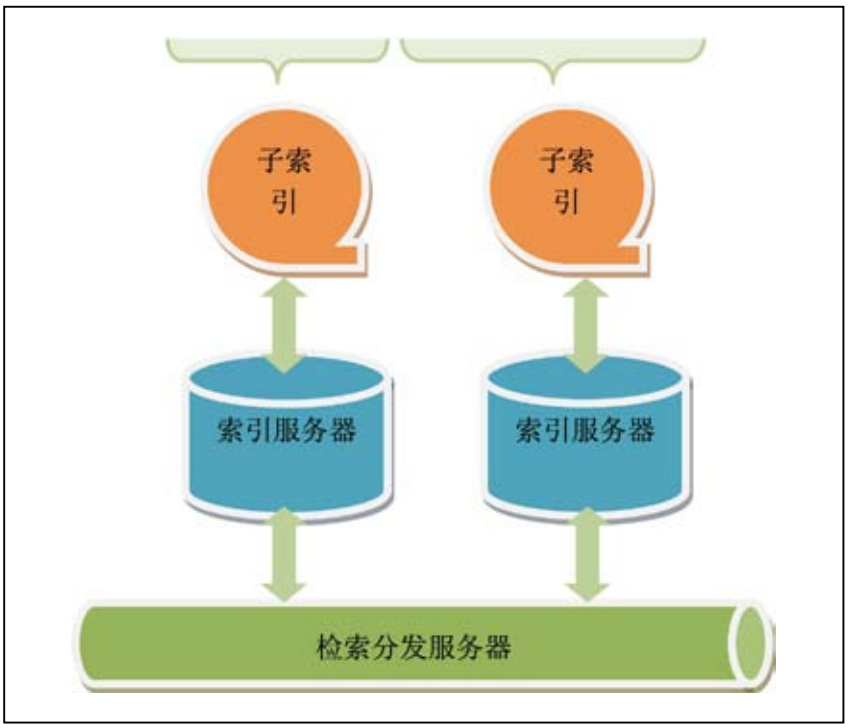

图2 索引集群和检索响应流程

数据库厂商资源的同时，可能存在多种情况，如购买的方式不同，可能为网络或镜像；购买的资源范围不同，可能是全库或某几个学科分类；购买的镜像的版本不同。

为了解决这些问题，保证原文链接的准确性，学术搜索网提供了相应的原文链接定位服务iSFX。

学术搜索网保证每个文档拥有唯一文献标识，iSFX定义URL链接规则知识库，通过唯一标识获取数据来源和多个可能的规则，并自动完成规则组配。一种数据来源加上其对应的规范化的OPENURL规则和识别机制可以称为“定位方案”，在学术搜索网中“定位方案”可能是固定的，也可能是根据用户的个性化需求定制的。

3.3 核心功能

学术搜索网提供的资源发现功

能主要包括：

（1）统一检索功能

界面使用一框式的检索，用户通过在框内去表达检索意图或者构造表达式，获得检索结果。同时学术搜索网引入了后控词表检索，词表类别包括主题词表、人物词表、机构词表等，如用户输入“水污染”时，搜索系统会同时获取词表中的相关词如“水体污染”进行扩展检索，提高系统的查全率。

（2）检索结果分面筛选

检索结果集可进行不同数据类型、不同维度的聚类显示，可进行多种排序，如相关度、时间、题名、馆藏位置等。在命中结果较多时，能够根据用户对聚类方式、排序方式实现分面检索。

（3）资源导航实现

学术搜索网实现的是搜索引擎倡导的傻瓜式的检索方式，提倡给用户最低的学习成本。但是面对诸多厂商过亿的数据总量和动辄百万级的检索结果来看，再高级的分面筛选其实还是无法替代图书馆传统的A-Z的数据库导航和出版物导航。在这样的情况下，学术搜索网的出版物的导航和订阅功能才是资源发现另一大核心价值。

（4）资源原文链接与获取

学术搜索网整合了若干家资源厂商的元数据，其重要功能就是揭示一次原文的获取途径，也就是原文获取链接功能，包括链接指向数据库厂商官方站点或者内部镜像服务站点，从而帮助用户下载全文。由于图书馆的特殊服务地位，原文链接服务往往不是单一的，在与图书馆用户合作的过程中，还会实现原文链接请求与原文传递、参考咨询、馆际互借等服务的集成。

（5）数据挖掘分析

学术搜索网存储的是“厚数据”，即数据覆盖全面、标引完整、质量高、规范性好。利用这些规范数据，针对用户科研需求，学术搜索网从主题、学科、学者、机构等用户最关心的若干要素生成了若干知识库，借助知识库开发了若干分析服务，如提供学科研究趋势和主题变迁、机构/学者的科研状况跟踪等分析功能，这些功能将从资源发现系统的信息检索服务上升到科技情报服务。

（6）科研辅助服务

高素质的科研人员以及科研管理部门往往会面临更多实际的场景，如科研立项选题、科技论文写作、期刊评价、学者/机构科研能力评估、学科发展潜力评估、主题研究趋势分析等，这些场景需求是嵌入在整个科研过程中的。学术搜索网根据实际应用需求，提供了多种应用工具和在线服务，如学位论文开题服务、创新助手、机构创新能力透视等。

3.4 服务方式

学术搜索网主要采用云服务模式，即软件即服务（Software as a Service，简称SaaS）模式，同时支持云端与本地数据支持的混合模式。其中SaaS模式是完全的托管方式，图书馆不需要任何硬件投入，也无需安装任何软件，只需要在供应商在云端将各个系统之间的接口配置好就能使用，所有的厂商系统均支持这种方式。这种模式对图书馆管理者来说压力最小，代价最低，但是很难实现资源的查全率，同时对云端数据很难作太多个性化的操控。

混合模式是指图书馆将在本地服务器安装系统，实现对本地馆藏的OPAC数据、本馆机构仓储数据、特藏数据的管理和索引，并提供本地检索服务，而庞大的中心索引库及系统仍存放在云端，以SaaS的模式提供公共数据检索服务。这种模式能够较好地解决元数据资源覆盖全面的问题，而且在本地数据的即时更新和数据保护方面具有优势，但是由于受限于整个发现系统的体系架构，在资源的数据收割、清洗、规范化的过程中，不可避免需要投入人力和物力，也是图书馆在具体实践中需要权衡的重要因素。

4 待改进之处

图情市场提供搜索发现服务的厂商也有很多，市场上的资源发现产品的首次出现当推2008年OCLC的WorldCat Local，在此之后，分别在Proquest旗下的Serials Solution公司和以色列的艾利贝斯公司分别于2009年7月和2010年1月推出了Summon和Primo发现服务[6]。紧接着，EBSCO联合推出了EDS发现服务，后又与Innovative Interfaces公司开展了深度合作，并做了诸多技术改进[7]。

综合比较多家厂商，主流的资源发现系统还存在很多的改进空间。

4.1 资源覆盖增强

目前所有的资源发现系统获得授权的元数据都是集中在图书、期刊等文献型数据上，对图片、音乐、影片和数值型数据覆盖很少，而一些专业型分析工具类型数据库则更少。而中外文厂商在语种覆盖上也存在各自相应的短板。

实际上，资源发现系统囊括所有馆藏并不现实，而元数据收录的工作让厂商独立完成也是一项艰巨的任务。在扩大资源合作渠道和收割力度的情况下，借助互联网很多UGC（用户创造内容）的理念，厂商提供在线的数据完善任务和数据加载工具，鼓励终端用户参与数据共建，实现资源互惠服务或有偿服务，或许也是一个实现资源覆盖增强的途径。

4.2 互动服务支持

资源发现系统相对于以往的电子资源数据库的查询功能，补充了很多互联网应用中的互动功能，有效地提升了用户体验。但现在提供的应用主要是用户创建个性化标签、发表评论等，允许高级用户注册后并对数据进行维护，系统与其他互联网产品的关联如百科、Wiki、豆瓣书评等。图书馆提供的发现服务，是与师生的科研活动密切相关的。未来的资源发现系统如果能够嵌入到科研过程之中，会有更长久的生命力。

4.3 由资源发现转向知识发现

目前所有厂商还是停留在文献数据的揭示上，也就是数据库厂商发现级别的服务。资源发现系统若能够从信息检索延展到数据分析，帮助用户从海量文献中获取规律性或特征性情报，则可能对学校长期践行的学科服务提供更多益处。

5 结语

资源发现系统的引进已经开始对图书馆的馆藏方式、资源采购方式和信息服务方式带来了积极而深远的影响。首先，由于发现系统统一了资源检索入口，简化了检索方式，降低了使用门槛，大大提高了资源的使用率。

正是如此，当80%以上的用户集中涌入到发现系统进行“先集中，后发散”的搜索访问时，由于“长尾效应”，很多原本被忽略的电子资源数据库的价值被挖掘出来。很多图书馆因为资源发现系统的引入，带来数据库访问量的大幅提升，与此同时，发现系统统一的日志访问记录提升了图书馆员对各类馆藏数据的监控管理能力和使用程度的了解，有效地提升了图书馆的工作效率，加之SaaS服务模式减少了相应的系统投入和人员配备，使得图书馆可以在经费使用效率上更好地平衡。

同时，资源发现系统不仅整合了数据和检索入口，也整合了多种Web 2.0的服务方式，如维基百科、用户评论、网摘、标签等，使得图书馆以用户为中心的理念更为强化和实际，缩短了用户与图书馆的距离。

当然，云服务模式下的资源发现会进一步改变图书馆的工作模式，传统的馆藏建设工作都是图书馆人员的分内之事，而有了云端的大数据共享[8]，当数据存储和数据管理有了专门的公司来承担，图书馆的管理人员的工作重点将会从以往的加工数据转移到分析数据上，真正利用资源的原生数据和用户的行为数据开展更为有效的知识服务。

参考文献

[1] 李正祥.数字化时代的图书馆消亡论刍议[J],情报资料工作,2002(1).
[2] 端木瑜.基于数字图书馆的异构资源检索[D].北京:北京信息科技大学,2006.
[3] 聂华,朱玲.网络级发现服务——通向深度整合与便捷获取的路径[J].大学图书馆学报,2011(6):5-10.
[4] 吴广印,杨奕虹,杨贺.从知识获取看知识组织——基于“知识获取五要素”的知识组织研究与实现[C]//国家科技图书文献中心:数字图书馆高层论坛2010年年会论文集,2010.
[5] 张俊林.这就是搜索引擎:核心技术详解[M].电子工业出版社,2013:68.
[6] ExLibris. Primo: Empowering Libraries to Address User Needs [OL]. [2013-04-20]. http://www.exlibrisgroup.com/category/PrimoOverview.
[7] EBSCO出版社和Innovative Interface合作改进EBSCO发现服务TM和Encore发现平台[J].现代图书情报技术,2011(11).
[8] FRANKS B.驾驭大数据[M].人民邮电出版社,2013:16.

作者简介

王灏（1977-），女，北京万方软件产品经理。“863”专项课题“以科技文献为主的搜索引擎研制”的课题组成员，参与了中国学术搜索网总体设计工作。研究方向：大规模中文信息处理与文本挖掘。E-mail: wanghao@wanfangdata.com.cn

Research and Implementation of Libraries' Resource Discovery System

Wang Hao, Zhang Zhengfeng, Feng Wei / Beijing Wanfang Data Co., Ltd., Beijing, 100038

Abstract: Resource discovery system is a kind of vertical search engine focused on library and information science. For the general problem in libraries and information centers, this article gave a brief introduction of the concept and main technologies. With the example of "Chinese Academic Search," it has provided some practical details about the design, main function and service model of resource systems in order to provide suggestions for others' purchase or construction of discovery service if possible.

Keywords: Resource discovery system, Discovery service, Search engine, Cloud computing, Software as a Service

（收稿日期：2013-05-14）