数字资源揭示

——海量数据环境下图书馆资源发现之路学术研讨会论文集

《数字资源揭示——海量数据环境下图书馆资源发现之路学术研讨会论文集》编委会　编

國家圖書館出版社
National Library of China Publishing House

数字资源揭示

——海量数据环境下图书馆资源发现之路学术研讨会论文集

《数字资源揭示——海量数据环境下图书馆资源发现之路学术研讨会论文集》编委会　编

國家圖書館出版社
National Library of China Publishing House

图书在版编目(CIP)数据

数字资源揭示:海量数据环境下图书馆资源发现之路学术研讨会论文集/《数字资源揭示:海量数据环境下图书馆资源发现之路学术研讨会论文集》编委会编. --北京:国家图书馆出版社,2015.6

ISBN 978-7-5013-5556-3

Ⅰ.①数… Ⅱ.①数… Ⅲ.①数字图书馆—文集 Ⅳ.①G250.76-53

中国版本图书馆 CIP 数据核字(2015)第 045960 号

书　　名　数字资源揭示——海量数据环境下图书馆资源发现之路学术研讨会论文集
著　　者　《数字资源揭示——海量数据环境下图书馆资源发现之路学术研讨会论文集》编委会　编
责任编辑　高　爽　王炳乾
封面设计　耕者设计工作室

出　　版　国家图书馆出版社(100034　北京市西城区文津街 7 号)
　　(原书目文献出版社　北京图书馆出版社)
发　　行　010-66114536　66126153　66151313　66175620
　　66121706(传真),66126156(门市部)
E-mail　btsfxb@nlc.gov.cn(邮购)
Website　www.nlcpress.com——→投稿中心
经　　销　新华书店
印　　装　北京科信印刷有限公司
版　　次　2015 年 6 月第 1 版　2015 年 6 月第 1 次印刷

开　　本　787×1092(毫米)　1/16
印　　张　12.5
字　　数　250 千字

书　　号　ISBN 978-7-5013-5556-3
定　　价　80.00 元

论文评审专家组成员

（按音序排序）

曹　宁　富　平　李春明　李志尧

吕淑萍　毛雅君　申晓娟　孙一钢

王乐春　魏大威　邢　军　张智雄

前　言

信息技术的发展将人类带入了全新的时代，也渗透到了图书馆的方方面面。为促进信息技术与图书馆业务的加速融合，推进数字图书馆技术与服务的不断进步，国家图书馆自2013年起开始举办图书馆现代技术学术研讨会，邀请国内外图书馆领域的专家、学者和业界同仁，共同探讨图书馆现代技术的发展前瞻、探索实践和应用创新。

此次研讨会的主题为“数字资源揭示——海量数据环境下图书馆资源发现之路”，围绕数字图书馆大数据、数字图书馆检索技术、数字图书馆服务整合等内容向业界进行了广泛征文和交流探讨，内容涉及资源发现、云计算、大数据、创新服务、资源共享及服务整合等诸多方面，取得了丰硕的成果。

现将此次研讨会的优秀论文整理成册，集结出版。希望通过这些论文，分享业界同行的研究成果和实践经验，促进现代信息技术对图书馆业务的支撑与引领作用，推动图书馆事业的蓬勃发展。

本书编委会

2015年1月

目　录

图书馆资源发现系统的资源组织和检索排序研究*

王乐春　杨东波　杨　帆(国家图书馆)

1　检索结果现状及分析

1.1　主要搜索引擎检索结果现状

2014年9月13日在百度、CALIS[1]、CADAL[2]、NSL[3]、NSTL[4]和文津搜索[5]中键入“信息安全”后所返回的搜索结果首页面情况,如图1至图6所示。

图1　百度搜索检索结果

在百度搜索结果首页中显示,共搜索出100 000 000(1亿)条搜索结果,其中头三条为广

*本文为国家科技支撑计划课题“文化资源服务平台解决方案及标准研究”(2012BAH01F01)成果。

告信息;紧跟其后的是百度百科、最新相关信息、百度贴吧、权威网站、维基百科、百度百科、百度文库等;搜索结果页的最后三条信息仍是广告信息。而与本次搜索相关的该搜索引擎推荐的栏目还包括:猜你喜欢、相关词汇、相关学科、专业知识早知道、相关搜索;该页面中还包括大量的广告性质的推广链接。

图2　CALIS检索结果

在CALIS搜索结果首页显示,共搜到12 407条相关资源,默认按相关度排序(还包含按首字母排序);第一条记录为学报,搜索结果涵盖图书、期刊等内容。包含“信息安全”词条解释。最左侧有详细资源分类及所含资源数量。让人印象深刻的是学位论文有10 258个。主题词除关联了中文词外,还关联了Information Security等英文单词。显著位置有和信息安全不相关的“直播公开课课程表”。

在CADAL搜索结果首页显示,共搜到143条相关资源,首页显示36条,2/3为图书,1/3为学位论文。搜索结果可以按照类型、标签和出版社分类。

使用NSL新版一搜即得,搜索结果首页显示,共搜到12 774条相关资源,默认按时间排序,最新为2014年资源(还包含按题名和相关度排序);首页全部为论文,全部来源维普中文科技期刊。进行了文本模式下和可视化模式下的详细分类。显著位置栏目为:“可以在如下数据库中检索”。首页最下方有一个小标签“期刊影响力排序”。

在NSTL搜索栏中,默认的检索资源类型为外文期刊和会议,为了一致选择全部资源进行检索,15.236秒后,搜索结果显示,中文期刊50 529条,学位论文10 123条等。最后首页显示中文期刊10条,未标明排序方式,但首页所列文章都为2014年发表论文,涵盖多个期刊。

图 3　CADAL 检索结果

图 4　NSL 检索结果

图5　NSTL 检索结果

图6　文津搜索检索结果

在文津搜索结果首页中显示，获得结果为 260 000 个，按照相关性排序（还包括题目 A—Z、作者 A—Z、出版单位 A—Z 和出版日期升/逆等排序方式），首页展示 10 条结果，全部命名为“信息安全”（9 本不同作者的专著，1 篇期刊论文）。查看指定类型分为：图书、文档、论

文、词条、多媒体和古文献。缩小检索范围分为:全文、年份、作者、语种。对来源数据库进行了分类。页面最后给出相关搜索。有查看其他版本和分册的特色展示。

1.2 检索结果分析

第一,搜索结果差距巨大。在键入同一个关键词后,不仅商业搜索引擎的检索结果和以学术为目标的数字图书馆检索结果差别巨大,各数字图书馆给出的检索结果也差别巨大。该差别不仅表现在返回结果的数量上,在内容上也差别巨大。

第二,搜索速度不再成为搜索面临的主要问题。除 NSTL 以外,在所有使用的各搜索引擎中,其搜索关键词提交后在秒级都反馈了大量的检索结果,反馈的搜索结果数量超出用户正常查看能力范围。经与 NSTL 机构确认,其检索速度 2015 年也将提升到秒级。目前搜索引擎的查全率、查准率和查询速度等纯技术指标已经不再是搜索引擎面临的主要问题。

第三,统一搜索已成为主流。在所有搜索引擎返回的检索结果中至少对两种以上的资源同时进行了检索,有些搜索引擎涵盖了十几个,甚至几十个资源类型和资源库。

第四,首页展示设计成为提高搜索引擎黏性的重要手段。各搜索引擎首页展示各有特色,突出了自身的资源组织能力,独特的资源评价体系,具有显性或隐形的排序特征。

总之,这些搜索引擎都在追求对搜索关键词的深刻理解;并在自己所擅长的领域里的尽可能广泛的范围内,搜寻相关的素材;对检索结果还要进行严格筛选,精心组织;最后呈现给用户的检索结果,试图在首页能够浓缩出最有价值的结果,通过首页展示反映出搜索引擎背后实体的真正实力。

2 搜索引擎

2.1 图书馆目录检索发展

目录检索是图书馆搜索引擎独特和最重要的应用场景,图书馆的检索目录可以追溯到 40 年以前,联机公共检索目录(OPAC,Online Public Access Catalog)[6] 系统自从推出以来,经历了几个发展阶段,20 世纪 70 年代 OPAC 采用传统图书馆卡片目录构建思路,提供与卡片相同的记录内容、记录格式及检索点。80 年代中期经过部分调整之后,OPAC 检索系统实现了关键词检索和布尔检索,用户操作界面实现了帮助、浏览、查询、用户导航和人机交互功能,有的 OPAC 系统甚至具备高级检索和词组检索。90 年代,OPAC 结合了增强式检索和匹配技术以及检索结果相关性排序等关键技术,真正实现了人机交互,并可以改善用户的检索策略和检索过程,最终帮助读者检索到较为理想的检索结果。2000 年在进入 Web2.0 时代以后,OPAC 检索也进入了崭新的历史时期。OPAC 系统完善了检索功能,包括多字段检索、多库检索、高级检索、命令语言检索等多种检索方式;丰富了检索结果的揭示,包括多样化的书目信息、分页、分面、摘要等;完善了读者借阅、预约收藏等功能;增加了更多与读者交互的功能,如评论、评级、标签和荐购等;增加了借阅排行和馆员推荐等特色功能。总之 OPAC 系统为读者信息检索和利用图书馆馆藏带来了极大的便利。

2.2 互联网搜索引擎

在图书馆搜索引擎蓬勃发展并愈加完善时，一个发展异常迅猛的搜索引擎应用方向在形成，甚至可以说是图书馆搜索引擎的噩梦。最早的互联网搜索引擎可以追溯到1990年，第一个互联网上的搜索引擎为Archie，Archie甚至不能被称为真正意义上的搜索引擎，它用于搜索FTP服务器上的文件，而这个时候基于HTTP协议的Web还没有出现。基于HTTP协议的Web出现后，先后出现了Wanderer和ALIWEB两个搜索引擎，前者是只收集网址而没有索引文件内容，后者开始索引文件元信息（标题与标签等）。1994年4月，第一个全文搜索引擎WebCrawler推出后广受欢迎，随后Lycos、Yahoo和Excite搜索引擎相继推出，成为早期流行的搜索引擎，但是这些搜索引擎以检索结果的数量衡量检索质量存在搜索结果相关性差等问题，用户无法有效找到满意答案，于是很快退出了历史舞台。如果说Yahoo是第一代搜索引擎的代表，那么Google就代表了第二代搜索引擎。第二代搜索引擎以关键字搜索为主要特征，在短时间内可以在海量的信息里准确找到用户需要的信息。相比第一代搜索引擎，第二代搜索引擎提高了检索速度与精度，使用了网站评级算法以及数据挖掘相关技术，并引入了人工智能和机器翻译等技术，并将搜索引擎处理数据的能力提高至EB级别。

2.3 资源发现系统

随着图书馆的数字资源不断增长和互联网搜索引擎的普及，图书馆的检索系统受到了前所未有的挑战。《图书馆与信息资源的理解：给OCLC成员的报告》显示：84%的用户使用搜索引擎进行检索，1%的人从图书馆网页上进行信息检索。《OCLC白皮书关于大学生信息搜索习惯》中指出：90%以上的大学生首选搜索引擎查询网络资源，他们更倾向于凭借自己的力量，使用搜索引擎如Google scholar等来迅速获取更加全面的学习知识。

为了应对互联网搜索对图书馆检索造成的冲击，借鉴互联网搜索引擎的先进技术，结合图书馆资源组织的优势，以及自身数据的格式特点，图书馆领域提出统一资源发现系统。2009年7月，Proquest旗下的Serials Solution公司推出第一款网络级资源发现系统Summon。同月，以色列Ex Libris公司介绍了Primo Central元数据仓储的建设理念，并于2010年1月发布统一资源发现系统Primo测试版，将原有的Primo架构到Primo Central和本地馆藏资源之上。此时，EBSCO公司也发布了EBSCO Discovery Service（简称EDS）系统。OCLC于2007年11月推出Worldcat Local系统，提供对馆藏印本和电子资源的一站式检索，随着OCLC与数据库商的不断合作，Worldcat Local集成了元搜索功能，并于2010年开始提供基于海量元数据的网络级资源发现服务。

这些新型的资源发现系统和OPAC比较起来特色明显，首先，他们实现了统一发现和统一检索。统一资源发现系统实现对图书馆纸本资源和电子资源的整合，能够同时检索图书馆各种类型的资源，甚至包括那些没有被图书馆订购但被中心索引覆盖的其他资源。通过使用统一界面上的单一检索框，提供类似Google的简单检索，用户不必在各个数据库系统之间跳转，不必花费很大的精力去学习和掌握各个数据库系统的使用方法。其次，他们有效提升了检索速度。由于统一资源发现系统是基于格式统一、结构清晰的元数据中

心索引进行的检索，因此检索速度可以达到秒级，甚至毫秒级。第三，他们可以有效集成多种服务。可以实现对图书馆书目系统（OPAC）、全文数据库、文摘和引文数据库，乃至原文传递、参考咨询、馆际互借等服务的集成。第四，他们有效提升了用户体验。检索结果提供特定资源推荐或者补充结果集的资源推荐；允许用户对检索结果创建标签、评分、发表评论等；提供可视化的标签云图；混搭 Wiki 词条、图书封面、网摘、目次和读者评论等。

虽然图书馆新推出的各个资源发现系统对资源的组织和揭示有了本质上的提升，单就搜索结果的用户满意度这一复合指标而言，他们与 Google 等一流的互联网搜索引擎仍有巨大差距。Google 的成功有许多因素，抛开查全率、查准率和查询速度等这些纯计算机技术范畴的指标，其制胜的优势是能够保证让绝大部分用搜索的人都能在搜索结果的第一“页”找到他想要的结果。究其原因，一是覆盖范围广，它的搜索范围覆盖全球的每一个有计算机的角落，二是 Google 对搜索结果的排序比其他搜索引擎都要好。

下面我们以国家图书馆文津搜索为例分析其资源组织和搜索排序，来探讨图书馆资源发现系统在资源组织和检索结果排序方面的发展趋势。

3 资源组织

3.1 资源现状

文津搜索是国家图书馆自主研发的资源发现和揭示系统，其需要组织的传统资源如图 7 所示，图 8 标示出了文津搜索需要整合的电子资源组成情况，其中电子图书 353.8 万种 399.4 万册；电子期刊约 5.6 万种；电子报纸约 1.5 万种；学位论文约 400.4 万篇；会议论文约 365.5 万篇；音频资料约 107.2 万首；视频资料约 10.2 万小时。表 1 列出了文津搜索需要考虑的外购资源库的情况。

图 7　文津搜索需要组织的传统资源

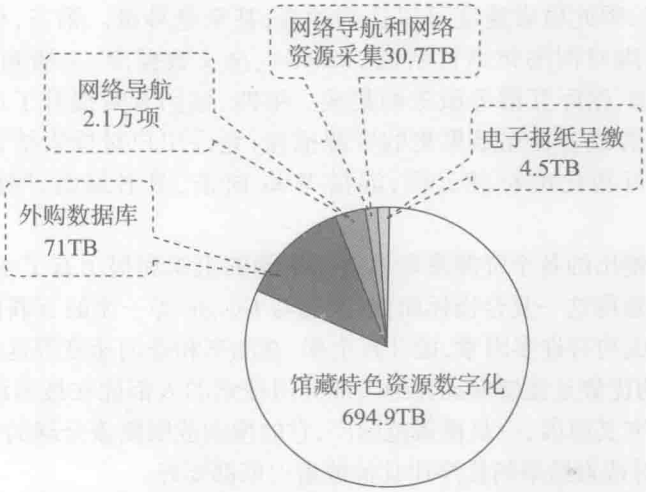

图8　文津搜索需要组织的数字资源组成比率

表1　外购数据库分类统计

文献类型	中文数量	外文数量	合计
电子图书	47.1万种	249.6万种	296.7万种
电子期刊	2.1万种	3万种	5.1万种
电子报纸	0.2万种	1.3万种	1.5万种
学位论文	341万篇	36.7万篇	377.7万篇
会议论文	364万篇	1.5万篇	365.5万篇
音频资料	55.6万首	0	55.6万首
视频资料	4.9万小时	0	4.9万小时

3.2　资源归类组织

文津搜索对所有进入搜索引擎的元数据进行了格式转换，形成统一元数据，不仅提高了检索速度，也避免了检索中不必要的检索噪音。根据文津搜索需要组织的资源是3.1节所列出的所有资源，参照ISBD和RDA最新资源类型标识，经过所有数据的分析和归类，最后在文津搜索中国家图书馆文献资源类型归纳如表2所列内容。

表2　文津搜索中文献资源分类列表

序号	文献资源类型	包含内容
1	图书	纸质图书包括专著、民语文献、海外中文图书、善本、新善本、普通古籍、再造善本、地方志、家谱、单独成册的联合国资料、敦煌资料等；数字图书包括纸质图书的数字化、外购电子书数据库、配合阅读器推出的电子书、网上下载的电子书、光盘或其他存储介质的图书

续表

序号	文献资源类型	包含内容
2	连续性资源	纸质连续性资源包括传统期刊、报纸等；数字连续性资源包括纸质连续性资源的数字化、外购电子期刊数据库、数字报纸与纸质连续性资源对应的光盘期刊或光盘报纸等
3	论文	纸质论文包括传统的博士论文、硕士论文、博士后报告、海外学位论文等；数字论文包括相应论文的光盘版、网上提交文档或数字化版本，同时也包括外购数据库中的期刊论文、会议论文和开放存取的论文等
4	专利标准	仅包括标准单行本和外购的专利、标准数据库
5	音频资源	包括不同载体形态的录音制品、自建录音资料、外购音频数据库、网络采集的音频等，如音乐数据库、语言类节目
6	视频资源	包括不同载体形态的录像制品、自建视频资料、外购视频数据库、网络采集的视频等，如学术性报告、专题节目、多媒体课程、舞台艺术
7	地图	纸质地图包括印刷、手绘、晒印及其他制作方法生产的古今地图、地图册及地图集划分到纸质图书类；数字地图包括纸质地图的数字化、光盘或其他载体形态的地图资料、网上地图等
8	缩微品	包括缩微胶卷、缩微胶片
9	手稿	包括原始稿、复制品及数字化
10	金石拓片	包括拓片、墓志、拓本、龟甲、獸骨等，实物和相应的数字化藏品都归入此类
11	静画资料	包括投影制品和图形资料，如投影片、幻灯片、美术复制品、闪现卡、图表、照片、招贴画、挂图、工程图等，不同载体形态
12	网络资源	包括政府信息采集、国内外图书馆信息采集、专题网络采集等
13	数据库资源	包括外购文摘/索引数据库、数值/事实数据库、工具型数据库以及其他不能归入前面 12 类的数据库
14	其他	不属于前 13 类的资源

3.3 资源组织思考

文津搜索通过元数据识别、解析、标准化、清洗等处理流程实现十几类元数据的整合，检索速度达到了亚秒级别，建立 5 亿条索引数据不超过 24 小时。文津搜索通过对元数据的操作实现了对结构化数据、半结构化数据以及非结构化数据的海量处理。纵观文津搜索资源发现系统的资源组织，可以说它整合了国家数字图书馆自建和其他方式（外购数据库和资源征集）获取的数字资源，已实现两亿多条元数据的整合，建立了分布式索引，为读者提供“一站式”检索服务。但相比一流的互联网搜索引擎而言，其在资源组织方面还有很长的路要走。

（1）资源集的构成上还不健全。文津搜索目前已经完成了两亿条元数据的整合，未来可以达到 5 亿条，在中文图书馆数字资源方面可以说是独一无二的，但是和面向全球的互联网搜索引擎来比，还是差别巨大。未来资源集的选取是文津搜索值得思考和研究的问题，也是每一个资源发现系统必须面临的挑战。

(2)资源更新和索引建立频率低。除了自建资源可以自控以外,文津搜索资源更新依托于数据库提供商的元数据提供情况,由于图书馆资源订购的特点,很多资源库的更新不是实时的。对广大读者来说,图书馆数字资源的检索如果没有最新的研究成果,很多研究人员势必倒向更新更快的资源提供商。

(3)检索深度亟需提升。图书馆在元数据的设计和制作上有先天的优势,因此在文津搜索中实现了元数据的整合和高速检索。对象资源的摘要和目次信息等内容,还没有进入文津搜索的索引建立范围,这些将是文津搜索要立即扩大的范围;长远来看,资源发现系统发展,势必首先要通过规范控制实现对对象数据的全文检索,再往后的发展将是通过自身的分词和自然语言理解实现在所有数字对象中的全文检索和深度数据挖掘。

4 检索结果排序

搜索引擎检索结果的排序是所有搜索系统的核心功能之一,是搜索引擎讳莫如深的部分,尤其是其缺省排序更是秘而不宣,是体现一个搜索引擎目标定位的试金石。

4.1 排序可依据的基础

元数据搜索引擎其数据基础是元数据仓储中的元数据,这些元数据有自身标准的格式,因此检索词只有落在这些元数据元素时,才能选出对应的对象资源,因此可以说各电子资源的元数据项决定了元数据搜索引擎搜索结果排序的范围和所能达到的效果。以文津搜索为例,其使用的典型对象资源的元数据著录所包含元素如表3所示。

表3 文津搜索中典型元数据著录元素

序号	文献资源类型	项数	详细内容
1	电子图书	19	题名、创建者、主题、描述、出版者、其他责任者、日期、类型、格式、标识符、来源、语种、关联、时空范围、权限、版本、价格、馆藏信息、书评
2	连续性资源	16	题名、创建者、主题、描述、出版者、其他责任者、日期、类型、格式、标识符、语种、出版频率、关联、来源、权限、馆藏信息
3	学位论文	16	题名、作者、主题、描述、导师、日期、类型、格式、标识符、来源、语种、关联、权限、时空范围、学位、馆藏信息
4	古籍	21	题名、主要责任者、其他责任者、日期、出版者、附注、相关资源、主题、时空范围、语种、类型、格式、标识符、来源、权限、版本类型、载体形态、收藏历史、文献保护、馆藏信息、其他复本信息
5	舆图	22	题名、主要责任者、其他责任者、日期、出版者、附注、相关资源、主题、时空范围、来源、语种、类型、格式、标识符、权限、版本类型、载体形态、馆藏信息、收藏历史、文献保护、其他复本信息、制图技法
6	视频资源	19	题名、创建者、主题、描述、出版者、其他责任者、日期、类型、格式、标识符、来源、语种、关联、时空范围、权限、版本、受众、馆藏信息、源载体

续表

序号	文献资源类型	项数	详细内容
7	音频资源	19	题名、创建者、主题、描述、出版者、其他责任者、日期、类型、格式、标识符、来源、语种、关联、时空范围、权限、版本、受众、馆藏信息、源载体
8	网络资源	19	题名、创建者、主题、描述、出版者、其他责任者、日期、类型、格式、标识符、来源、语种、关联、时空范围、权限、版本、馆藏信息、受众、采集地址

4.2 文津搜索检索结果排序考虑的原则

文津搜索向用户提供统一、实时、高效、精准、权威的元数据搜索服务，因此其反馈给用户的检索结果必须紧紧围绕其目标进行设计。第一，检索结果的排序必须反映统一检索和揭示特征，在首页的组织排序中要充分展现多种资源类型，包括图书、期刊、论文、古籍等，并且依据自身定义的展示策略进行归类调度，形成统一的排序结果。第二，检索结果的排序必须反映资源实时特征，搜索引擎揭示的资源要反映时效性，要反映出最新馆藏、时下热点，既要依据更新时间、上线时间，又要依据资源本身的出版时间，形成依据时间轴的排序结果。第三，检索结果的排序必须反映高效性，搜索结果有很好的分类和导航功能，可以引导用户迅速找到其目标结果，因此搜索结果排序中既要有单个资源的对象展示，也要有依据类别统计的排序展示，并且要有其独特的相关推荐，使读者如果在第一页找不到所需资料的情况下，迅速找到进一步寻找资源的入口。第四，检索结果的排序必须反映精准性，搜索引擎揭示的资源要和用户所寻找的资源密切相关，密切相关不能简单理解为完全匹配，完全匹配只是相关性的一个重要指标，相关度排序还要考虑是匹配的哪个字段，是作者字段、题名字段、关键词字段、还是摘要字段，排序展示要依据重要性给出优先级，相关性排序还要考虑该关键词出现的频率等因素，精准性还要考虑搜索词的规范表达，形成规范控制的相似词，以及对应的外文词，并且依据词意的内涵和外延，扩大或缩小范围后，提出进一步精准搜索的推荐内容。第五，检索结果的排序必须反映权威性，其给出的排序顺序不仅要反映该资源的影响因子、引用次数、是否为经典论文等资源对象本身权威要素，还要考虑对象资源的来源、位置、资源库是否公开等因素。

4.3 检索排序思考

在文津搜索结果排序中首选是相关性排序显示，其次可以根据用户需求，实现题目的拼音顺序排序，作者的拼音顺序排序，出版单位拼音顺序排序，实时性可以根据出版日期的升/逆序排序。导航可以引导读者根据年份、作者、语种等分类到相应的来源数据库进行针对性查找。文津搜索和其他图书馆资源发现系统一样，在面临海量检索结果的现实挑战面前，排序展示具有多个方面需要深入研究，需要用互联网思维完善资源发现系统，这样的资源发现系统才能在未来信息化环境中存活、壮大和发展。

(1)资源发现系统排序需要体现其背后资源的特点和自身对资源的组织理念。检索结果排序展示给用户，并非只是展示资源发现系统背后的资源量，更需要能够反映出背后依托