大学图书馆 OPAC 系统用户信息搜寻路径的可视化分析*

□姜婷婷　陈舜昌　高慧琴

摘要　从武汉大学图书馆 OPAC 系统获取为期 18 天的访问日志数据并对其进行清洗和处理，基于点击流数据分析框架的路径层开展数据分析。经划分、编码、筛选得到 51410 条待分析路径并按长度分为 3 组，利用 Levenshtein Distance 算法提取各组中心路径，以 2D 折线图的形式对其进行可视化。可视化分析揭示，用户访问 OPAC 系统留下的主要是包含 3—10 个页面的短路径，他们习惯于从图书馆主页进入 OPAC，很少使用复杂的搜索方式，倾向于将 OPAC 作为查阅资源所在馆藏位置的工具。图书馆应考虑从界面和功能两个方面改善 OPAC 设计，以帮助用户更高效地利用馆藏资源。

关键词　OPAC　信息搜寻　路径　可视化

分类号　G258.6

DOI　10.16603/j.issn1002—1027.2017.01.009

1　引言

尽管网络已经成为人们在日常生活和工作中获取信息的主要途径，但对于高校师生和研究人员而言，他们仍然在很大程度上依赖着图书馆馆藏资源以获取系统、权威的学术信息和知识。联机公共查询目录(Online Public Access Catalogue，OPAC)是图书馆用户获取馆藏图书纸本或电子书的入口。随着网络技术的发展，OPAC 的发展已经进入了一个全新的阶段，即“下一代图书馆目录”(Next Generation OPAC)：凭借优化的排序算法、分面导航工具、可定制的界面等为用户的信息搜寻过程提供更丰富、更有效的支持；除了通过搜索、浏览等方式获取感兴趣的书目信息外，用户还可以访问在线个人图书馆服务，完成预约、续借等操作。

目前，国内外许多知名高校的图书馆都通过向获得广泛认可的发现服务提供商(如 The Summon Service、Ex Libris Primo、EBSCO Discovery Service 等)购买功能强大的发现工具(Discovery Tools)完成了下一代 OPAC 的升级。一直以来人们都十分注重开展用户测试以了解 OPAC 系统的运转情况，从而为系统的优化升级管理提供科学的策略参考。而针对下一代 OPAC 的用户行为研究在网络学术资源日益丰富的今天更是尤为必要，因为以谷歌学术(https://scholar.google.com)和微软学术(https://academic.microsoft.com)为代表的学术搜索引擎正借助其独特的技术优势逐步取代 OPAC 成为主要的学术资源获取与检索工具。

在以往的 OPAC 用户行为研究中，人们较为关注用户与系统搜索功能的交互，包括检索式的构造、搜索结果的查看以及搜索改进等方面，所采取的研究方法主要有调查、实验、访谈、搜索日志等。文章从武汉大学图书馆 OPAC 系统服务器中提取了连续 18 天的事务日志，对其中所包含的 757 万条点击流记录进行了处理与分析，通过路径可视化的方式完整展现了用户使用 OPAC 系统获取馆藏资源的行为特征，而不仅仅局限于其搜索行为。

* 国家自然科学基金青年项目“用户探寻式搜索策略分析及系统构建研究”(71203163)和教育部人文社会科学研究青年基金项目“社会性标签系统中用户信息搜寻行为研究”(12YJC70011)的研究成果之一。

通讯作者：陈舜昌，ORCID：0000-0003-0562-2944，whuchenzjc@163.com。

2 相关研究

2.1 OPAC系统用户信息行为研究

早期的OPAC系统用户信息行为研究主要关注两种类型的搜索,即已知书目搜索(Known-item Search)与主题搜索(Subject Search)[1]。这二者各有侧重:已知书目搜索指已知特定书目题录信息(题名、作者或其他字段等),用户利用题录信息进行精确搜索;主题搜索是查找与某个主题相关的全部书目,与前者相比更具开放性[2-3]。实践中,这两种类型的搜索有时无法确切区分开来,比如说"信息检索"既可作为一个书名,也可作为一个宽泛的主题。主题搜索成功率往往较低,因为它要求用户具有较高的搜索技能,可以构建精准表达信息需求的查询式,并在搜索失败的情况下对查询式进行重构[3-4]。后来主题搜索逐渐被关键词搜索所代替,后者的搜索效果更好[5-6]。

下一代OPAC对发现工具的引入引起了人们对OPAC研究的进一步关注。发现工具可提供类似谷歌的搜索体验,因而获得了大学图书馆用户的青睐。今年的研究发现显示,用户偏好能够接受任何关键词的单一搜索框以及拼写检查与查询式建议等搜索工具[7-9];在查看搜索结果时,他们一般只会查看搜索结果的第一页[10],对相关性排序和搜索结果的质量较为满意[11]。更重要的是,由于具有分面导航和Web 2.0功能,发现工具为下一代OPAC用户的信息探索与发现创造了卓越体验。

分面导航的基础是一组分类层级,每个层级对应着集合的一个方面[12]。就大学图书馆而言,其资源集合的分面一般包括作者、主题、出版年、地区、语种等。用户可以按照任意顺序查看任意个分面,选择其下的分类并浏览其中包含的条目。这种方式大大减轻用户认知负担[13]。在实际应用中,用户确实非常依赖分面来区分不同类型的字眼,他们认为分面导航是一种直观的工具,可以帮助他们洞察整个搜索结果空间并对结果进行精炼[9-10,14-15]。一系列的实证研究表明,采用分面导航的发现工具可以提供更好的搜索体验[16-17]。不过也有研究注意到用户在利用分面的时候可能遇到困难,因而分面及其所包含分类的设计需要比较谨慎[18-19]。

Web 2.0功能主要包括标签、用户评论、评分和RSS订阅等,这些功能在下一代OPAC中也起到了重要的作用[20]。"图书馆2.0"(Library 2.0)概念的提出表现了独立用户参与OPAC重构的价值[21]。人们认为用户不仅愿意贡献自己的知识,而且也希望能够利用他人贡献的内容[22-23]。然而在OPAC系统中提供Web 2.0功能的做法还存在一些争议,许多用户对其有用性存在怀疑,在信息搜寻过程中也不愿用到这些功能,主要是因为用户已经习惯了简单搜索界面,而对更新的Web 2.0技术缺乏必要的了解[9,24-25]。

2.2 点击流数据分析及路径可视化分析

网络服务器中储存的事务日志分为两种:搜索日志(Search Logs)与点击流数据(Clickstream Data)。前者是由用户与网络搜索系统(如通用搜索引擎和站内搜索应用等)之间的交互而产生的,一般包含用户ID、访问日期和时间、用户查询式、搜索结果页面和结果点击等字段[26]。后者记录的则是用户的点击情况。从用户进入直到离开网站的这段时间内,所有对页面、按钮的点击都被记录下来。因此点击流表现了用户在网站中的导航路径,反映了他们在访问网站过程中所提交的一系列页面请求及其顺序[27]。常见的数据字段包括用户ID,日期和时间,请求方式,请求资源,指引页面(Referring Page)等[28]。

搜索日志分析已经被广泛应用于各种网络搜索系统中的用户信息搜索行为研究,其中也包括OPAC系统[29-32]。目前获得普遍认可的搜索日志分析框架由简森(Jansen)于2006年提出,包含关键词、查询式和搜索会话三个层次,研究人员一般都会基于其中一个或多个层次开展分析。但是针对OPAC系统的点击流数据分析却并不多见[33-35],而且并未形成统一的研究方案。尽管点击流数据在信息行为领域未能得到很好的利用,但在电子商务领域却得到了充分的重视,常用于了解网站的使用情况和用户的导航模式,以及营销策略的有效性和客户购买转化率等[36]。

电子商务领域的研究人员对点击流数据的利用存在着分析方法过多、难以选择的问题。有研究者认为有必要形成"结构化的方法论",因而提出了一套全新的分析框架,采用"脚印"(Footprint)、"踪迹"(Track)和"路线"(Trail)这三个概念来描述用户访问网站的行为[37]。脚印表示由用户与网页之间的交互产生的一条点击流记录;踪迹是脚印的集合,按照时间先后顺序提供了用户所有的浏览操作历史;最后,对相似踪迹的聚类产生了路线,反映了相似的

行为、属性、信仰和价值观。该框架适用于在线购物网站的研究,而对于研究信息内容丰富的在线环境(如 OPAC)也具有重要的借鉴意义,因为前者中客户需要找到商品以满足购买需求,而后者中用户需要找到信息资源以满足信息需求。

考虑到人类信息行为的多样性以及点击流数据的格式特征[38],对以上框架进行了改进,创建了一套更适合信息行为的点击流数据分析框架。该框架包含三个层次,即"脚印"(Footprint)、"移动"(Movement)和"路径"(Pathway)。简单而言,当用户在访问网站时,其每一次页面请求都会在页面上留下一个脚印,将两个连续发生的脚印连接起来便形成一次移动,最后将所有的移动按照时间顺序链接起来便形成了用户访问网站的一条路径。这个新的分析框架已经应用于社会性标签系统用户的信息行为研究,其有效性得到了验证[13]。

相关文献中已有研究专注于路径层次的分析,旨在通过直观的可视化图形揭示用户行为模式。其中,研究人员所采用的可视化方法主要可以分为两大类:一类是对较长一段时间内累积的页面访问统计情况进行可视化,针对用户整体观察其行为特征及变化趋势;另一种则是首先对单个用户的访问路径分别进行可视化,然后依据机器可读表达对它们聚类,挑选处于聚类中心的路径作为代表开展分析。

对于反映整体统计情况的可视化方法,人们一般会采用不同形式的树状图来描绘网页之间的层级关系[39-42],由连接线的方向和粗细表示访问量的流向和大小[43]。除此之外,色谱图(Stratogram)[44]、自组织地图(Self-organizing Map,SOM)[45]以及社会网络图[46]等都曾用于展示用户访问情况的累积结果。这一类可视化方法具有明显的缺点,因为统计异常值和偏差会导致研究人员忽略用户行为上的细节差异。另一方面,以单个访问路径为展示对象的可视化方法则主要采用拓扑图[47-49],其中结点表示页面,可添加序号表示页面访问顺序,连线表示跳转关系。特别值得一提的是首创的足迹图(Footstep Graph)[50],这是一种基于方格网中折线图的可视化图形,纵轴标记为页面类型,横轴标记为时间,通过折线反映用户在页面之间转移的方向与耗时。该可视化方法可转换为机器自动识别模式[51],但无法揭示页面在站点结构中的上下级关系。

3 数据与方法

选取武汉大学图书馆 OPAC 系统作为研究对象。该 OPAC 于 2009 年由 Ex Libris Primo 开发,是典型的下一代 OPAC,主要服务于武汉大学的师生。用户可以在系统中进行简单搜索、高级搜索、Aleph 命令搜索以及分类浏览。OPAC 主页上默认为简单搜索,而图书馆主页上也可以在"馆藏目录"选项下进行 OPAC 简单搜索,此外系统还为每一种搜索方式额外提供了独立的入口界面。

在 OPAC 搜索结果页面上,用户可以利用各种工具查看结果条目并且对搜索进行精炼。除了基本的排序和格式选择功能外,他们还可以通过在搜索结果中再次搜索或是采用右侧的分面导航栏来限定搜索范围。其中默认的分面包括主题词、年份、语种、馆藏、分类、作者、关键词以及格式。点击搜索结果条目会将用户带往资源详情页面,从而查看到简介、目录、馆藏位置等信息。资源的完整记录可以加入收藏、保存或是邮件发送。

3.1 数据收集

数据来源于武汉大学图书馆服务器中为期 18 天的 OPAC 访问日志,即 2014 年 10 月 13 日 0:00:00 至 2014 年 10 月 30 日 23:59:59,为学期中的常规时段。原始日志文件共包含 7574170 条记录,以 W3C 扩展日志格式存储。考虑到研究的需要,从中提取了 6 个基本字段,分别为:

- 用户 IP:用户的 IP 地址,用于区分不同用户;
- 访问日期:页面请求发送的日期;
- 访问时间:页面请求发送的具体时间;
- 请求类型:客户端对服务器的请求类型,主要为 GET(获取数据)和 POST(提交数据);
- 资源地址:用户请求访问资源的 URL;
- 协议状态:服务器返回的 HTTP 状态码,如 404 和 200 等。

3.2 数据准备

(1)数据清洗。

数据清洗是网络日志分析中的重要环节,对原始日志文件的清洗主要是为了去除其中的崩溃记录和冗余记录。前者是由服务器记录数据时发生错误造成的,可以通过对每个字段依次进行排序来快速识别,格式异常的数据会集中在字段列的顶部、底部或是聚集在一起[52]。后者是与研究无关的数据,无法反映用户的信息行为,过滤掉这些数据可以极大

压缩文件的大小,从而提高分析效率。数据清洗的具体步骤如表1所示。经过数据清理后的日志文件共包含800320条有效记录。

表1 数据清洗步骤

清洗步骤	无效数据类型	数据清洗措施
1	错误记录、不完整记录、重复记录等	对各字段依次排序,删除格式异常的记录
2	外部链接	删除URL以"http://"开头的记录,例如"http://www.baidu.com/"
3	图片或网页样式加载记录	删除URL以"png"、"jpg"、"gif"、"ico"、"css"、"js"等结尾的记录
4	失败的请求	删除协议状态码不属于200类(成功的请求)的记录,例如404(找不到)、500(内部服务器错误)
5	提交数据的请求	删除请求类型为"POST"的记录

(2)访问路径划分。

由于本研究主要关注路径层上的用户行为特征,路径划分是数据准备阶段的必要环节。路径包含了一次访问中用户与系统之间的所有交互活动,从用户进入到离开OPAC的整个过程中所有的访问请求按时间顺序排列起来就构成了一条独立、完整的路径。在划分路径时,采用了搜索日志分析中的搜索会话划分的方法,即不同用户IP的记录属于不同路径;对于同一用户IP的记录,若两条记录之间的时间间隔超过30分钟这个阈值,则也属于不同路径。借助Python程序对用户访问路径进行抽取划分共得到104080条路径。但值得注意的是,这其中也包括了非人为路径,即由计算机代理(如网络爬虫)产生的路径,其特征是包含了大量的记录。因此将阈值设为100,即记录数超过100的路径都视为非人为路径,将其去除后共得到103542条人为路径。

(3)访问路径表示。

OPAC用户的访问路径在本质上可以表示为页面的访问和页面之间的跳转,因此需要考虑的关键元素主要包括页面的类型和跳转之前在页面上耗费的时间。耗时情况可以根据两条相邻记录的时间差来计算,而页面类型则需要人工干预标注。基于武汉大学图书馆OPAC的系统结构与页面功能,最终分析了5大类型的页面,分别为:

首页(H):OPAC系统首页;

搜索界面(S):该路径选择的搜索方式,如简单检索、高级检索等;

搜索结果列表(L):搜索结果返回的展示页面;

资源详情(D):用户点击结果条目后跳转到的详情页面;

个人图书馆(I):用户登录后个人信息页面。

每种类型的页面都允许用户采取相应的行动实现特定的功能,而页面类型可以通过URL中的关键字符串来进行识别。附录列出了5大页面类型及其子类型,对每个子类型都分配了一个专有编码,然后利用Python程序自动解析日志中每条记录的URL并将编码添加到新字段中,这样一来路径就可以采用编码串来指代,如['H1', 'S1', 'L2']。

4 分析与结果

为了揭示用户访问路径的特征,本研究采取的方法是从以上人为路径中找到具有代表性的路径,以可视化的形式将其发生模式展现出来。在典型路径的提取过程中,首先删除了长度小于3个页面的超短路径(所能反映的用户行为非常有限),从而得到51410条路径。由于路径长度差别较大,须对这些路径按其长度进行分组,通过计算最小平均编辑距离来确定组内中心路径,将其作为典型路径。计算接近中心性(Closeness Centrality)是在一组对象中寻找中心(最具代表性的对象)的常用方法[53],而接近中心性和平均距离成反比关系;因此,平均距离最小的路径是同类路径中处于中心地位的一条,可以很好地反映同类路径共同特征。

最小平均编辑距离的计算采用的是Levenshtein Distance算法,这是一种判断两条任意长度的字符串之间的相似性的方法,反映的是两个字符串之间相互转换所须增删改操作的最小次数,且不要求两字符串等长,可以有效地用于两个短字符串,或一长一短两个字符串之间[54]。由于访问路径长度不一,大部分都很短(77.29%的路径含有不超过10个页面,且所有路径包含页面数量平均值仅为2.48),因此使用Levenshtein Distance比较合适。以下是用Levenshtein Distance计算路径间两两距离、寻找中心路径的简单示例。

示例:给定若干路径,根据以下方法判断其中典型路径。

步骤一:将路径表达为数组。

Path 1=['H1', 'S1', 'L2']
Path 2=['H1', 'I1', 'I4', 'I1']
Path 3=['H2', 'L1', 'L2', 'D2']

步骤二:计算数组两两之间平均距离。

Path 1	Path 2	Path 3
LD(1,2)=3	LD(2,1)=3	LD(3,1)=3
LD(1,3)=3	LD(2,3)=4	LD(3,2)=4
Avg=3	Avg=3.5	Avg=3.5

步骤三:选择典型路径。

因为 Path 1 具有最小平均距离,所以典型路径为 Path 1。

通过反复比较对路径长度不同划分条件下类中心编辑距离的方差发现,以页面数 10 和 20 作为划分节点时,各类路径呈现出最具代表性的特征。因此,将 51410 条待分析路径分为 3 组,它们包含页面数分别为 3～10(Cluster 1),10～20(Cluster 2)以及 20 以上(Cluster 3),这三组所包含的路径数量分别为 39736、6439、5235。在每组中分别取最小平均编辑距离的那条路径作为中心路径,最终得到 3 条用户访问 OPAC 的典型路径(Typical Paths, TP):

- TP 1: ['L1', 'L1', 'I12'];
- TP 2: ['L1', 'L1', 'L10', 'L2', 'L2', 'L9', 'L9', 'D1', 'L2', 'L2', 'L2', 'L2'];
- TP 3: ['L1', 'L1', 'L9', 'L2', 'L2', 'L10', 'L10', 'L2', 'L9', 'I11', 'I1', 'I4', 'I5', 'L1', 'L1', 'L1', 'L1', 'L10', 'L2', 'L2', 'L1', 'L1'].

对这 3 条典型路径分别进行了的可视化表示,采用的是 2D 折线图的形式,横轴表示时间,用户纵轴用以区分页面类型,每条路径都是由节点和箭头组成的。需要指出的是,所有的典型路径都是从简单搜索结果页面(L1)开始的,对应的前一步是在武汉大学图书馆主页上进行简单搜索,但是该主页不属于 OPAC 系统,因此未被记录在 OPAC 访问日志中。

Cluster 1 中最具代表性的用户访问路径 TP 1 如图 1 所示。在该路径中,用户访问了 3 个页面,时间总共持续了 4 分 14 秒。结合附录中的页面类型编码,对此路径进行如下描述:用户在武汉大学图书馆主页上的"馆藏目录"选项下进行简单搜索进入 OPAC,首先到达"多库检索结果"页面(L1),此时用户需要在"西文文献库"和"中文文献库"中做出选择,选定数据库后进入相应的搜索结果页面(L1);在第一个搜索结果页面上,用户花费了将近 4 分钟的时间查看结果条目,最后直接退出了系统(I12)。需要特别说明的是,OPAC 系统会在搜索结果页面上根据鼠标悬停弹出资源的馆藏位置,这样用户不一定需要进入资源详情页面。因此,这很有可能是一条线性的查寻路径,用户找到所需资源的馆藏位置后,信息需求得到满足而离开系统。

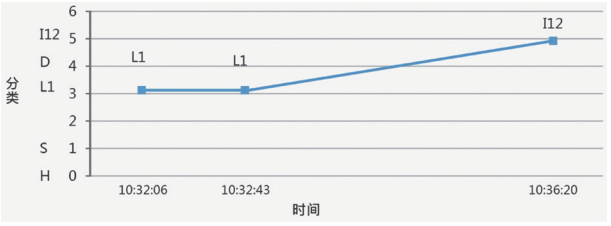

图 1 Cluster1 典型访问路径

图 2 展示了 Cluster 2 中的典型访问路径 TP2。此路径属于中等长度路径,包括了 12 个页面,时间总共持续了 6 分 23 秒。该用户同样是在图书馆主页上搜索进入 OPAC,经由"多库检索结果"页面(L1)到达搜索结果结果页面(L1),在花费了 1 分钟左右查看结果条目后决定对搜索进行精炼(L10)。然而这并没有满足用户的需求,他又连续进行了两次简单搜索(L2),并在第二次搜索结果页面上进行了两次翻页操作(L9),最终进入资源详情页面(D1)。遗憾的是,无法确定该资源是否满足了用户需求,因为紧接着他又连续多次进行简单搜索(L2),这一系列动作可能是希望找到更多相关结果,也可能是因为结果不理想而调整查询式。但可以肯定的是,TP2 与 TP1 相比表现出明显的探索性,用户以多种形式与搜索结果发生交互。

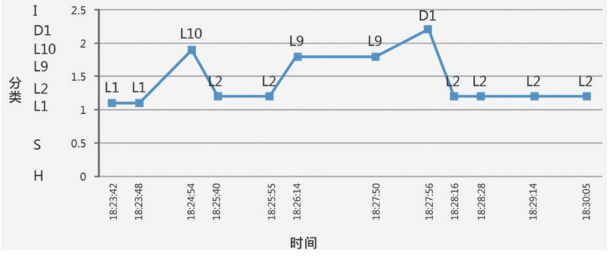

图 2 Cluster2 典型访问路径

在图 3 中可以看到 Cluster 3 的典型访问路径 TP3,持续时间比 TP2 稍长,为 7 分 14 秒,但是用户一共访问了 22 个页面。在该路径中,用户的访问大致可以分为三个阶段。首先,在前面的 3 分多钟时

间里,他从图书馆主页搜索进入OPAC(L1－L1),经历了一个非线性的搜索过程,多次执行翻页(L9)、精炼(L10)、查询式重构(L2)等操作。接着,用户中断了搜索行为,转而登录个人图书馆(I11),默认进入个人信息页面(I1),选择查看了个人借阅信息(I4)和个人借阅历史(I5),这一行为很有可能是因为刚才的搜索并不顺利,他突然想起以前曾经借阅的资源可能会提供一些有用的线索。最后,用户又回到图书馆主页进行了多次搜索(L1－L1),其中也穿插了结果精炼(L10)和查询式重构(L2)。与TP2一样,TP3也是一条探索路径,用户采用了多元化的搜索策略,其中额外利用了个人图书馆中存储的信息。

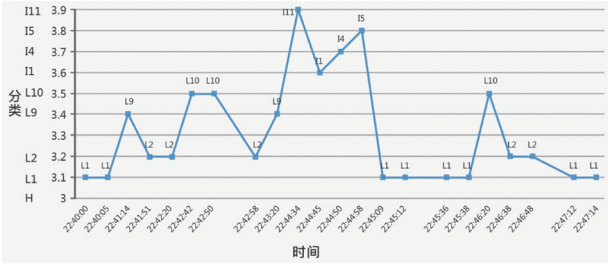

图3　Cluster3典型访问路径

5　讨论

5.1　界面设计

从用户的实际访问情况来看,武汉大学图书馆OPAC系统在界面设计上存在一些冗余之处,有时可能影响用户搜索资源的效率。武汉大学图书馆将馆藏资源与电子期刊、数据库等资源区分开来,专门提供OPAC搜索,而且对各种传统搜索方式进行了区分。然而这些搜索方式的独立入口界面(S)使用率极低,用户甚至很少从OPAC主页上开始搜索,而是从图书馆主页上的"馆藏目录"搜索进入OPAC系统,这一过程中还需要完成语种选择才能查看到搜索结果。这样做无疑增加了搜索活动的复杂度,用户可以明显感受到搜索流程的中断,也可能对图书馆网站界面和OPAC系统界面的视觉设计差异产生疑惑。

在这一点上,国外许多大学图书馆(如哈佛大学图书馆、斯坦福大学图书馆等)则采取了不同的做法,他们提供的是单一搜索框的整合式搜索,用户可以直接提交任何查询式,由系统识别其意义,再从多个来源聚合搜索结果,返回到统一的界面上,同时将结果类型作为一个分面允许用户选择所需类型。这样做更加符合新一代互联网用户的使用习惯,作为出生在数字时代的年轻人,他们早已对通用搜索引擎(如Google、Amazon)的单一搜索框习以为常,能够熟练运用自动完成、查询建议等工具。因此,武汉大学图书馆可以考虑整合其搜索界面,减少语种选择、查询式类型选择等冗繁的搜索步骤,将精力放在简单搜索的交互和算法设计上。

5.2　功能设计

以上讨论根据路径长度的不同对所有路径进行了分组,其中Cluster 1的规模最大,包含了77.29%的路径,也就是说绝大多数路径都是页面数为3～10的短路径。从TP1的可视化视图来看,武汉大学图书馆OPAC系统主要起到了了解资源馆藏位置的作用。也就是说,用户已经通过其他手段找到所需资源的题录信息,如书名、书号等,而在OPAC中搜索这些已知条目的目的就是在现实图书馆内对其进行定位,从而获取物理资源。从某种意义上讲,OPAC只是被简单地用作查阅工具,这并不符合下一代OPAC的特征。

虽然TP2和TP3的可视化视图表现出一定程度的探索性,但是频繁的交互并不一定是有复杂或模糊的搜索任务造成的。从时间轴来看,用户在每个页面上花费的时间都不算太长,大多不超过1分钟,这段时间或许足够用户大致查看页面内容,不过如果需要对这些内容进行理解、思考探索策略,用户将需要更多的时间。此外,用户访问的基本上都是搜索结果页面,极少进入资源详情页面,前者仅提供了书名、作者、年份、出版社等有效信息,这对于用户深入了解搜索主题帮助很有限。因此,即使用户多次精炼结果、翻页、重构查询式,这些行为很可能是他们在应对系统功能缺陷的表现,例如结果匹配和排序不理想、查询式构建工具缺失等。

另外特别值得一提的是,武汉大学图书馆OPAC系统为用户提供了个人图书馆,方便他们利用传统的图书馆服务,包括借还书、续借、预约、提醒等。目前个人图书馆只是起到了个人信息管理工具的作用,但是如果能够引入Web 2.0元素,用户体验将得到大幅提升:基于用户借阅或收藏的资源对搜索结果进行个性化地排序;基于相似用户的借阅或收藏情况推荐可能有价值的资源;对用户个人行为进行聚合形成一些总体的行为趋势。

6 结论

以武汉大学图书馆 OPAC 为例对用户访问下一代 OPAC 系统的典型路径进行了可视化分析。具体来讲,对来自武汉大学图书馆的为期 18 天的 OPAC 访问日志进行了采集、清洗和处理,将分析重点放在路径层次,依次实现了路径的划分、编码表示、分组、中心提取以及可视化表示。分析结果显示,用户访问 OPAC 系统留下的主要是包含 3～10 个页面的短路径,他们习惯于从图书馆主页进入 OPAC,很少使用复杂的搜索方式,倾向将 OPAC 作为资源馆藏位置的查阅工具。以上研究发现及其讨论对大学图书馆 OPAC 改善其界面和功能设计具有重要的启示作用。在后续研究中,将进一步探讨组内中心路径的提取,试图采用更理想的路径间距离计算方法,同时考虑页面跳转和停留时间这两个因素,也会将网站页面层级关系纳入权重计算考量。

参考文献

1 Large A, Beheshti J. OPACs: a research review[J]. Library & Information Science Research, 1997, 19(2):111－133.
2 Wildemuth B M, O'Neill A L. Research notes the "known" in known-item searches: empirical support for user-centered design [J]. College & Research Libraries, 1994, 56(3):3778－3781.
3 Hunter R N. Successes and failures of patrons searching the online catalog at a large academic library: a transaction log analysis [J]. American Library Association, 1990, 30(3):395－402.
4 Connell T H. Subject searching in online catalogs: metaknowledge used by experienced searchers[J]. Journal of the American Society for Information Science, 1995, 46(7):506 - 518.
5 Larson R R. The decline of subject searching: long-term trends and patterns of index use in an online catalog[J]. Journal of the American Society for Information Science, 1991, 42(3): 197－215.
6 Tillotson J. Is keyword searching the answer? [J]. College & Research Libraries, 1994, 56(3):199－206.
7 Gross J, Sheridan L. Web scale discovery: the user experience [J]. New Library World, 2013, 112(112):236－247.
8 Lown C, Sierra T, Boyer J. How users search the library from a single search box[J]. College & Research Libraries, 2013, 74(3):227－242.
9 Tam W, Cox A M, Bussey A. Student user preferences for features of next - generation OPACs: a case study of university of Sheffield international students[J]. Program Electronic Library & Information Systems, 2009, 43(4):349－374.
10 Sarah C. Williams, Anita K. Foster. promise fulfilled? an EBSCO discovery service usability study[J]. Journal of Web Librarianship, 2011, 5(3):179－198.
11 David J. Comeaux. Usability testing of a web-scale discovery system at an academic library[J]. College & Undergraduate Libraries, 2012, 19(2):189－206.
12 Hearst M A. Search user interfaces[M]. Cambridge University Press, 2009. [2016－09－25]. http://people.ischool.berkeley.edu/～hearst/talks/sui_google09.pdf.
13 Jiang T. A clickstream data analysis of users' information seeking modes in social tagging systems[J]. Ischools, 2014. [2015－09－25]. https://www.ideals.illinois.edu/bitstream/handle/2142/47288/091_ready.pdf? sequence＝2.
14 Melissa Becher, Kari Schmidt. Taking discovery systems for a test drive[J]. Journal of Web Librarianship, 2011, 5(3):199－219.
15 Denton W, Coysh S J. Usability testing of vufind at an academic library[J]. Library Hi Tech, 2011, 29(2):301－319.
16 Calvert P. Using mobile technology to deliver library services: a handbook[J]. Australian Library Journal, 2014, 62(4):333－334.
17 Fagan J C. Usability studies of faceted browsing: a literature review[J]. Information Technology & Libraries, 2013, 29(2):58－66.
18 Emanuel J. Usability of the vufind next-generation online catalog[J]. Information Technology & Libraries, 2012, 30(1):44－52.
19 Susan C. Wynne, Martha J. Hanscom. The effect of next-generation catalogs on catalogers and cataloging functions in academic libraries [J]. Cataloging & Classification Quarterly, 2011, 49(3):179－207.
20 Osborne H M, Cox A. An investigation into the perceptions of academic librarians and students towards next-generation OPACs and their features[J]. Program Electronic Library & Information Systems, 2015, 49(1):23－45.
21 Ozel N, Cakmak T. Users' expectations on restructuring opacs through social network applications[C]see: Green Computing and Communications. IEEE, 2010:798－803.
22 Sadeh T. User experience in the library: a case study[J]. New Library World, 2008, 109(1/2):7－24.
23 Yang S Q, Wagner K. Evaluating and comparing discovery tools: How close are we towards next generation catalog? [J]. Library Hi Tech, 2010, 28(4):690－709.
24 Barilan J, Haustein S, Peters I, et al. Beyond citations: scholars' visibility on the social web[J]. 2012. [2016－09－25]. http://sticonference.org/Proceedings/vol1/Bar-Ilan_Beyond_98.pdf.
25 Osborne H M, Cox A. An investigation into the perceptions of academic librarians and students towards next-generation OPACs and their features[J]. Program Electronic Library & Information Systems, 2015, 49(1):23－45.
26 Jansen B J. Understanding user-web interactions via web analytics[J]. Synthesis Lectures on Information Concepts Retrieval & Services, 2009, 1(1):102. [2016－09－25]. http://downloadpdfs.net/pdf183/understanding_user_web_interactions_via_web_analytics_bernard_j_jansen.pdf.
27 Montgomery A L, Li S, Srinivasan K, et al. Modeling online browsing and path analysis using clickstream data[J]. Marketing Science, 2004, 23(4):579－595.
28 Tatnall A. Web technologies: concepts, methodologies, tools, and applications (4 Volumes) [J]. 2010. [2016－09－25]. http://www.igi-global.com/Files/BookBrochures/9781605669823.pdf.
29 Blecic D D, Bangalore N S, Dorsch J L, et al. Using transaction log analysis to improve opac retrieval results[J]. College & Research Libraries, 1998, 59(1):39－50.
30 Lau E P, Goh H L. In search of query patterns: a case study of a university OPAC[J]. Information Processing & Management, 2006, 42(5):1316－1329.
31 Wolfram D. Search characteristics in different types of web-based ir environments: are they the same? [J]. Information Processing & Management, 2008, 44(3):1279－1292.
32 Niu X, Zhang T, Chen H. Study of user search activities with two discovery tools at an academic library [J]. International

Journal of Human-Computer Interaction, 2014, 30(5): 422-433.
33 Villén-Rueda L, Senso J A, Moya-Anegón F D. The use of OPAC in a large academic library: a transactional log analysis study of subject searching[J]. Journal of Academic Librarianship, 2007, 33(3):327-337.
34 Lown B, Cory. A transaction log analysis of NCSU's faceted navigation OPAC [J]. 2008.[2016-09-26]. https://ils.unc.edu/MSpapers/3387.pdf.
35 Asunka S, Hui S C, Hughes B, et al. Understanding academic information seeking habits through analysis of web server log files: the case of the teachers college library website[J]. Journal of Academic Librarianship, 2009, 35(1):33-45.
36 Bucklin R E, Sismeiro C. Click here for internet insight: advances in clickstream data analysis in marketing[J]. Journal of Interactive Marketing, 2008, 23(1):35-48.
37 Sen A, Dacin P A, Pattichis C. Current trends in web data analysis[J]. Communications of the Acm, 2006, 49(11):85-91.
38 Jiang T. Characterizing and evaluating users' information seeking behavior in social tagging systems[J]. Dissertations & Theses-Gradworks, 2010.[2016-09-25]. http://d-scholarship.pitt.edu/10412/1/Jiang_Tingting_etd2010.pdf.
39 Hong J I, Landay J A. Webquilt: a framework for capturing and visualizing the web experience[C] International Conference on World Wide Web. ACM, 2001:717-724.
40 Brainerd J, Becker B. Case study: e-commerce clickstream visualization[J]. Internet Computing IEEE, 2001:153-156.
41 Hofgesang P I, Kowalczyk W, Hofgesang P I. Analysing clickstream data: from anomaly detection to visitor profiling[J]. Ecml/pkdd Discovery Challenge, 2005.[2016-09-25]. http://www.cs.vu.nl/ci/DataMine/DIANA/papers/hofgesang05pkdd.pdf.
42 Chi E H. Improving web usability through visualization[J]. Internet Computing IEEE, 2002, 6(2):64-71.
43 Zhao J, Liu Z, Dontcheva M, et al. MatrixWave: visual comparison of event sequence data[C] SIGCHI Conference on Human Factors in Computing Systems. 2015:259-268.
44 Berendt B. Detail and context in web usage mining: coarsening and visualizing sequences.[J]. Lecture Notes in Computer Science, 2002, 2356:1-14.
45 Shen Z, Wei J, Ma K L, et al. Visual cluster exploration of web clickstream data[C] Visual Analytics Science and Technology. 2012:3-12.
46 Ortega J L, Aguillo I F. Network visualisation as a way to the web usage analysis[J]. Aslib Proceedings New Information Perspectives, 2013, 65(1):40-53.
47 Dömel P. Webmap: a graphical hypertext navigation tool[J]. Computer Networks & Isdn Systems, 1995, 28(95):85-97.
48 Cugini J, Scholtz J. Visvip: 3d visualization of paths through web sites[C] International Workshop on Database & Expert Systems Applications, 1999:259-259.
49 David Canter, Rod Rivers, Graham Storrs. Characterizing user navigation through complex data structures[J]. Behaviour & Information Technology, 1985, 4(4):93-102.
50 Ting I, Kimble C, Kudenko D. Visualizing and classifying the pattern of user's browsing behaviour for website design recommendation[C]. International Workshop on Knowledge Discovery in Data Stream, 2004:101-102.
51 Ting I H, Clark L, Kimble C, et al. APD-a tool for identifying behavioural patterns automatically from clickstream data[C]. Knowledge-Based Intelligent Information And Engineering Systems: KES 2007-WIRN 2007, Pt II, Proceedings, 2007:66-73.
52 Jansen B J, Spink A. How are we searching the world wide web? a comparison of nine search engine transaction logs[J]. Information Processing & Management, 2006, 42(1):248-263.
53 Tsvetovat M, Kouznetsov A. Social network analysis for startups finding connections on the social web[J]. 2011.[2016-09-25]. http://www.nhmnc.info/wp-content/uploads/fbp-dfs2014/Social-Network-Analysis-for-Startups-Finding-connections-on-the-social-web-by-Alexander-Kouznetsov-Social-Network-Analysis-For-Startups-.pdf.
54 Beusekom J V, Poulsen P G. Textual similarity[J]. Bachelor thesis [Academic thesis], 2012.[2016-09-25]. http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/6344/pdf/imm6344.pdf.

作者单位:武汉大学信息管理学院,武汉,430072
收稿日期:2016 年 8 月 4 日

Visualizing Users' Information Seeking Pathways on OPAC of Academic Library

Jiang Tingting Chen Shunchang Gao Huiqin

Abstract: An 18-days transaction log file was firstly obtained from the OPAC of Wuhan University Library, and then cleaned, prepared, and analyzed at the pathway level based on the clickstream data analysis framework. After data parsing, coding, and filtering, there remained 51,410 pathways which were divided into three clusters. The most typical pathway in each cluster identified with the Levenshtein Distance algorithm, was visualized in the form of 2D polyline graph. According to the visualizations, users' visit to the OPAC mainly created short pathways consisting of 3 to 10 pages. The library should consider improve the design of the OPAC in terms of both interface and functionality, in order to help users make better use of library collections.

Keywords: OPAC; Information Seeking; Pathways; Visualization

附录:武汉大学图书馆OPAC网页编码

页面类型	页面编码	页面内容	页面URL特征
首页	H	搜索入口	Start with:"/" OR "/F" OR "/F?"
检索入口	S1	简单检索	End with: "func=find-b-0"
	S2	多字段检索:最多可同时选择六个不同字段的检索框	End with: "func=find-a-0"
	S3	高级检索:最多可同时选择三个字段(可重复)的检索框	End with: "func=find-d-0"
	S4	通用命令语言检索	End with: "func=find-c-0"
	S5	全面检索/多库检索	End with: "func=find-m"
	S6	分类浏览书目	End with: "func=cat-list"
	S7	借阅排行	End with: "func=file&file_name=hotinfo"
检索结果列表	L1	书目语种分类选择;全面检索结果页	End with: "func=find-m" OR "func=find-m-results"
	L2	快速搜索栏检索结果页	End with: "func=find-b"
	L3	多字段检索结果页	End with: "func=find-a"
	L4	高级检索结果页	End with: "func=find-d"
	L5	通用命令语言结果页	End with: "func=find-c"
	L6	分类浏览结果页	End with: "func=cat-list"
	L7	上次检索结果	End with: "func=short"
	L8	检索历史(最近7次)	End with: "func=history"
	L9	检索结果翻页	End with: "func=short-jump&jump"
	L10	在结果中检索	End with: "func=short-refine"
	L11	结果筛选	End with: "func=short-filter"
	L12	结果排序	End with: "func=short-sort"
	L13	改变结果格式	End with: "func=short-format"
	L14	全选	End with: "func=short-select-all"
	L15	取消选择	End with: "func=short-deselect-all"
	L16	定题服务请求	End with: "func=short-sdi"
结果详情页	D1	书目详情	End with: "func=full-set-set"
	D2	馆藏信息(书目在分馆馆藏情况)	End with: "func=item-global"
	D3	添加到个人的收藏夹	End with: "func=myshelf-add-sel-1"
	D4	邮寄(以电子邮件邮寄书目信息)	End with: "func=short-mail"
个人信息	I1	用户信息	End with: "func=bor-info"
	I2	更新个人信息	End with: "func=bor-update"
	I3	修改个人账户登录口令	End with: "func=file&file_name=bo" OR "r-update-password"
	I4	个人借阅信息	End with: "func=bor-loan"
	I5	个人借阅历史	End with: "func=bor-history-loan"
	I6	预约结果	End with: "func=bor-hold"
	I7	个人账户现金记录	End with: "func=bor-cash"
	I8	定题服务	End with: "func=bor-sdi"
	I9	消息	End with: "func=bor-note-display"
	I10	个人收藏夹	End with: "func=myshelf"
	I11	登陆个人账户	End with: "func=bor-info"
	I12	退出个人账户	End with: "func=logout"
	I13	续借历史	End with: "func=bor-renew"