●索传军

论电子资源在线使用统计数据的收集与分析*

摘　要　可以从以下方面对电子资源(数字资源)进行数据统计:某段时间内用户访问电子资源的任务数,某段时间内的检索次数,某段时间内下载记录、文献和数据数量,某段时间内拒绝访问的平均任务数,等等。通过对它们的分析,便于对电子资源的服务绩效进行评价。参考文献 9。

关键词　图书馆　电子资源　在线使用　统计数据
分类号　G253

ABSTRACT　The author discusses parameters for the statistics of usage of electronic resources, including session numbers, search numbers and download numbers in a period of time, numbers of documents and data, average session numbers rejected in a period of time. Then, the author proposes a method for the evaluation of the performance of electronic resource services. 9 refs.

KEY WORDS　Library.　Electronic resource.　Online usage.　Statistical data.
CLASS NUMBER　G253

1　电子资源在线使用统计的作用与意义

电子资源或者说数字资源,已经成为图书馆馆藏资源建设的重要内容。其建设经费已经占到某些图书馆文献信息资源建设总经费的 1/3 左右。而且随着网络的普及,读者特别是年轻的读者(如高校的学生和青年教师等),更多地习惯于借助网络来获取自己所需的文献信息资料。但图书馆为读者订购的电子资源是否符合读者的需要,其利用率如何,都是需要研究的问题。

图书馆通过调查也许能够获得读者对电子资源的反映情况,但要知道电子资源客观质量和利用情况,或者说对用户的使用价值,就需要进行科学评价和评估。科学的评价能够帮助采访人员正确地选择电子资源,而对电子资源的服务绩效进行评估,能够使图书馆了解其利用效率,调整服务策略。因而无论是对电子资源质量的评价,还是服务绩效的评估,都具有十分重要的意义。但这些,都有赖于对电子资源使用情况的统计分析。

2　国内外电子资源在线使用统计的现状

2.1　国内外研究现状分析

对中国期刊全文数据库、维普中文科技数据库、全国报刊索引、人大报刊复印资料和中国高校学位论文等数据库检索得知:关于电子资源使用统计的文献仅有何雄的《图书馆电子资源使用统计》[1]和《图书馆电子资源使用统计的简易实现》两篇论文,其内容只是通过分析 IIS 日志文件,对页面和 IP 访问量进行统计。

对 Emerald Fulltext, ProQuese Digital Dissertation, ACM Digital Library, SCI Expanded, IEEE/IEE Elecero-nic Library 等 10 几个外文数据库检索和 Google 搜索可知,从上世纪 90 年代中期以来,国外不仅有一些文献探讨,而且关于电子图书馆服务评价、绩效评估,在线数据库商使用统计等相关研究项目有 10 余项,其中美国、欧洲研究较活跃。国外对电子资源的管理问题重视比我国早,研究更深入。一个十分明显的特点就是,多数项目是应用性研究。如美国图书馆和信息科学委员会的"在线数据库使用统计数据及报告机制"[2]和美国博物馆和图书馆服务研究所的"公共图书馆网络统计和绩效测度的国家数据收集模型"[3],试图为公共图书馆的统计和绩效测度设计一个可靠的、及时的数据收集、分析和报告系统。英国出版和图书馆理事会的"在线数据商使用统计数据研究"[4]等都非常重视信息资源管理过程中数据的可获取性

* 本文系国家自然科学基金项目(项目批准号 70573099)"电子资源在线使用统计与绩效评估"的阶段性成果和河南省高校新世纪优秀人才支持计划资助项目阶段性成果。

等应用基础问题研究。

近几年,一些国际学术会议也开始关注数字图书馆或电子资源的服务绩效问题。如 Northumbria 国际绩效评估研讨会的四届会议,均涉及电子资源、数字图书馆的绩效评估问题[5]。美国国会图书馆和信息科学委员会也组织召开了4次"网络绩效评估"研讨会[6]。

2.2 实践现状分析

通过 Google 搜索发现,国内个别图书馆、数据库商和软件商已经意识到使用数据统计的重要性,已开始收集、统计有关电子资源的使用情况,如 CALIS 的试用数据库统计报告、沈阳师范大学图书馆的"网络数据库统计表"[7],万方公司的"万方数据资源系统北大镜像站使用统计"[8]等。但这些统计不规范,不系统,内容过于简单。

2.3 电子资源在线使用的统计标准建设

为了适应图书馆科学管理的需要,国际标准化组织和美国国家标准局等对图书馆统计标准做了修改和补充,增加了有关电子资源的内容。如国际标准化组织对只覆盖传统服务的《ISO2789 信息和文献:图书馆统计数据》进行了修订,颁布了国际图书馆统计数据修订版,增加了一个附件:《ISO2798 Annex A:电子图书馆服务使用评价》。2002 年美国 NISO 发布了新版图书馆统计标准(*The Library Statistics Standard*, *ANSI/NISO* Z39.7)。

概括地说,关于电子资源质量评价、服务绩效评估和其在线使用统计等方面的研究还没有引起国内学者的重视,但国外的有关组织和学者已经开始了相关研究,并取得了一定成果。

3 电子资源在线使用的统计数据

对电子资源使用过程中所产生的数据,用户使用电子资源的感知情况等,需要运用专业技术(如数据挖掘等)和统计分析系统工具,从服务器日志文件和检索结果中分析、判断、整理出管理所需的数字,以便对电子资源的服务绩效科学地进行评估。

以往个别图书馆或数据库商通过用 IP 和访问某数据库页面的日志文件,对单位时间内(例如1个月等)读者的访问次数进行统计。而要真正了解电子资源的利用绩效,就必须知道读者对某种电子资源的每次访问或检索时间、下载记录或文献数量、系统拒绝访问的数量以及电子资源单位时间内正常的服务时间等等。这些数据是反映服务绩效的重要指标,虽然不能直接获取,但是通过对电子资源的一些使用数据的统计分析可以获得,因而,对电子资源的在线使用统计显得尤为重要。可以从以下7个方面进行数据统计。

(1)某段时间内用户访问电子资源的任务数。该指标反映电子资源的利用率。可通过服务器获得。

(2)某段时间内的检索次数。它反映系统服务的效能,可通过电子资源检索系统获得。

(3)某段时间内的访问时间和检索时间。该指标某种程度上能反映某资源对用户的重要程度,或用户对该资源的偏爱程度。可以从数据库服务器获得。

(4)某段时间内下载记录、文献和数据数量。下载数据说明用户已经找到了与自己需求相关的内容。单位时间内下载数据越多,说明用户找到自己所需的信息越多,该电子资源的绩效也就越好。从"平均每次下载文献数量"看,用户每次下载的越多,说明检索到自己所需的信息越多,与用户需求的相关度就越高。

(5)某段时间内拒绝访问的平均任务数。拒绝访问可能是因为系统超出并发用户数限制,也可能是系统故障。拒绝访问的任务数越多,用户满意度越低,感知服务绩效越差。

(6)系统平均无故障时间。该指标主要测度电子资源系统的稳定性,是测度电子资源可获取性的重要指标之一,也是用户感知服务绩效测度的重要指标。它需要数据库管理员统计。

(7)某段时间内正常服务时间。它是指单位时间内,除去电子资源系统、服务器等设备、网络和电源等故障,以及数据更新造成电子资源不能使用的时间后剩余的时间。这项指标反映了电子资源服务的及时性与可获取性及用户的满意度。可以通过系统管理员统计获得。

前4项指标可以客观地反映电子资源的服务效率和效能;而后3项能反映用户的满意度,测度用户对服务的感知绩效。这7个指标,基本上反映了电子资源服务的数量和质量。

4 电子资源在线使用统计数据对比分析

4.1 任务数与检索次数

一个任务(session)是指对一个数据库的一次成功的请求(request),是用户使用数据库从连接成功到任务完成后退出或超时为止的过程[9]。一次检索即

代表一次唯一的情报需求,向服务器提交一次检索请求记录为一次检索。一次任务可以包含多次检索。一次检索必然是某个任务的一部分或全部(当某个任务就进行一次检索时)。任务和检索是密切联系的,没有不包含检索的任务,也没有与任务不相关的检索。一个任务的时间一定大于或等于一次检索的时间。

4.2 下载(或浏览)文献数量与记录的数量

一条被下载的记录是检索一个数据库后全部显示出的一条编目记录和数据库条目。一篇被下载的文献就是一篇文献的全文和其中的一部分,在电子馆藏中就是传递给用户的文献,包括从电子期刊和数据库中下载的全文文献。

用户检索的根本目的是获取文献信息资料。但用户通过数据库检索,首先得到的不是文献,而是若干条检索记录的集合(记录数大于等于0),也就是数据库条目。然后用户再根据检索记录去获取文献的全文信息。

如果用J表示一次检索获得的记录数,用W表示一次检索用户下载文献数,那么二者之间具有如下关系(其中J,W都是大于等于0的正整数):

(1)J≥W,即一次检索,用户可能下载的记录数量一定大于等于下载文献的数量,而且,1≥W/J≥0。

(2)当W=0时,W/J=0,即用户下载文献数量为0,表示本次检索记录没有与用户需求相关的文献信息。

(3)当J=W时,W/J=1,即用户下载的文献数量与记录数量相同,表示本次检索记录与用户需求完全相关。

每次检索,W/J越大,表明系统的检准率越高,该电子资源的服务绩效也就越好。否则,相反。

4.3 单位时间内拒绝访问的任务数与任务数

这二者是紧密相关的,二者之和是某段时间内对电子资源访问的总任务数。

4.4 系统平均无故障时间与单位时间内正常服务时间

系统平均无故障时间与单位时间内正常服务时间是相关联的,二者成正比。系统平均无故障时间越长,单位时间内正常服务时间就越长。但正常服务时间不仅仅与系统平均无故障时间相关,而且还与系统其他因素,如更新时间、停电时间等相关。它们虽然不是影响电子资源服务绩效的主要指标,但对用户满意度影响很大。

5 电子资源在线统计数据与服务绩效指标的计算

对单位时间内用户访问电子资源的平均任务数和每次检索平均下载记录、文献和数据数量等指标,如何获取或计算,是评估中需要解决的问题。下面将详细分析每个指标的目的、定义、获取方法以及影响因素等。

另外,为了论述的方便,首先作几点说明或假设:

(1)每个用户指图书馆的合法读者群中的任何人,例如大学图书馆的学生和教师等。

(2)所有的统计数据都是针对某一段时间内某一种电子资源的服务。

(3)假设要统计的指标都是通过某种方法可以获取的。

(4)假设每种资源所处的环境是相同的,读者的信息素质和获取电子资源服务的条件等是相同的。

5.1 每个用户使用电子资源的平均任务数

(1)目的:获得用户使用电子资源的数量。

(2)定义:特定时间内(例如24小时)用户访问电子资源的任务总数除以用户总数。

(3)方法:如果用A表示特定时间内用户成功访问电子资源的任务数,B表示用户总数,那么,每个用户使用电子资源的平均任务数为A/B,其中A≥1、B≥1的正整数。

(4)解释及影响指标的因素:这个指标是一个没有上限的数值,数值越高,表明用户对电子资源的访问量越大,服务绩效越好。

5.2 对于每项电子资源每次任务平均下载(或浏览)文献和记录的数量

(1)目的:通过下载(浏览)文献或记录数量,可以表明用户感兴趣的记录有多少,与用户需求相关的文献有多少。这是反映电子资源服务绩效最重要的指标之一。

(2)定义:特定时间内每次任务下载(浏览)文献的数量除以其间该项服务的任务数。

(3)方法:特定时间内每次任务下载(浏览)文献数通常从数据库商提供的使用数据中获得。对有些服务这项数据可能无法获取,它应排除或通过用户调查得到。

对于一些电子资源,文献或记录可能只是下载而不在屏幕上浏览,例如篇幅长的期刊论文的记录可以很快扫过。这种情况下,如果这些数据能获取,下载

文献或记录可以与浏览文献加起来评价。

如果C表示特定时间内用户使用每次任务下载(或浏览)文献的数量,A表示特定时间内用户使用某种电子资源的任务数。那么,每次任务平均下载(或浏览)的文献和记录的数量可表示为:C/A,其中,A≥1、C≥0的正整数。

(4)解释及影响指标的因素:这项指标的数值是一个没有上限的正数,数值高表明用户检索到感兴趣的文献多,数值低可能表示满足用户需求的信息少。

(5)相关指标:每次任务的平均成本,每次任务下载(或浏览)每篇文献或记录的成本,用户满意度。

(6)每次检索平均下载的文献或记录数的计算法与此相同。

5.3 被拒绝的任务数占全部任务数的百分比

(1)目的:通过发现与电子资源服务不成功的链接,确定是否有充分的网络资源来满足用户请求。不成功的链接是由于访问超出并发用户的限制,并且表示用户对基础设施的提供是否满意。

(2)定义:被拒绝的电子服务的任务数占总任务数的百分比。

(3)方法:这些信息可从数据库商那里获得,特别是那些在一定用户许可限制基础上的服务。如果这些信息不能从某一特定电子资源得到,则可排除。

如果A表示特定时间内用户成功访问电子资源的任务数,F表示特定时间段内某一电子资源被访问的总任务数,则特定时间段内电子资源的被拒绝的任务数=F-A。那么,被拒绝的任务占总任务数的百分比:[(F-A)100/F]%,其中,F≥1的正整数。

(4)解释及影响指标的因素:这个指标值是一个0~1之间的数字,数值大表明访问不能成功的比率越大,这里不包括由于输错口令或密码引起的拒访,结果数值大表明对于该项服务需要更多的用户许可。

对于每一种电子资源,这项指标应该单独考虑,如果汇总计算全部电子资源服务的拒访任务数比例,结果意义不大。

(5)相关指标:用户满意度,每项电子资源每次任务的成本。

5.4 某段时间内的正常服务时间

如果某电子资源的更新时间不影响用户的访问(例如在夜里0~4时,在线更新),如果停电时间可以忽略不计,那么,我们就可以计算出某段时间内电子资源系统的正常服务时间和系统平均无故障时间。

(1)目的:通过该指标评价电子资源的可获取性。

(2)定义:某段时间内,电子资源系统正常服务时间等于总服务时间减去系统故障时间。

(3)方法:如果用T表示电子资源某段时间内的总服务时间,t表示某段时间内的电子资源系统的故障时间,Z表示正常服务时间,那么Z=T-t。

参考文献

1 何雄. 图书馆电子资源使用统计. 科技情报开发与经济,2004(9)
2 Denise Davis. Electronic Access And Use Related Measures: Summary of Findings. September 8, 2000. http://www.nclis.gov/statsurv/2000ven.pdf[2005-01-01查询]
3 Charle R. Mc Clure, John C. Bertot. Developing National Data Collection Models for Public Library Network Statistics and Performance Measures (funded by IMLS). http://www.ii.fsu.edu/getProjectDetail.cfm?pageID=9&ProjectID=8[2004-12-01查询]
4 http://p105.lib.nctu.edu.tw/2001conference/pdf/1-1.pdf[2005-01-20查询]
5 刘文梅. 国外数字图书馆绩效评估研究综述. 津图学刊,2003(6)
6 NCLIS. See website. http://www.nclis.gov/statsurv/statsurv.cfm[2003-05-12查询]
7 网络数据库统计表. 沈阳师范大学图书馆. http://210.30.208.249/net/database2/list.asp[2004-02-10查询]
8 万方资源系统(北大使用统计). http://162.105.138.185/database/wfdata.htm[2005-05-10查询]
9 ISO Information and Documentation - International Library Statistics. Switzerland, ISO27892003(E)

索传军 郑州大学信息管理系主任,郑州大学文献信息中心副主任,博士,教授,硕士生导师。通信地址:河南省郑州市。邮编450072。 (来稿时间:2005-10-12)