大数据环境中获取电子文献有效因子模型设计与实验

陈 俊 张伟云 赵 捷 (贵州师范大学图书馆 贵阳 550001)

摘要 为充分利用大数据资源提高图书馆主动服务质量,文章设计了大数据环境中获取电子文献有效因子模型。模型旨在通过安装节点插件,分布式采集数据,并针对读者获取电子文献情况,挖掘相关重要影响因素,结合适时电子调查问卷,建立获取电子文献有效因子数学模型。经实验验证基于该模型的主动服务达到了预期效果。

关键词 大数据 数据挖掘 有效因子 主动服务

Design and Experiment on the Effective Factor Model for the Electronic Document Capture in Big Data Environment

Chen Jun Zhang Weiyun Zhao Jie (Library of Guizhou Normal University, Guiyang, 550001)

Abstract In order to fully use big data resources to improve the quality of active service, this paper designed an access electronic document effective factors model in big data environment. The model aims to install the distributed data collection node plug-in, and it studies reader access to electronic documents, mining correlation important factors, and combined with timely electronic questionnaire, establishing access to electronic documents effectively factor mathematical model. Experiment shows the model active service achieves the desired effect.

Keywords big data, data mining, efficiency factor, active service

基于数据的海量增长,大数据时代已经到来。2012 年 3 月美国筹资 2 亿美元推出“大数据的研究和发展计划”。该计划旨在美国国家科学基金、美国国防部等六家部门合作下,推动并改善针对大数据关联的收集、组织、分析、决策技术。IBM、甲骨文、EMC、Google 等大公司也纷纷进入到涉及大数据的软件与硬件的技术整合中,并进行大数据信息处理技术的研究。

图书馆是计算机新技术应用的前沿阵地,结合图书馆的泛在服务性,图书馆无疑正处于一个大数据环境。图书馆大数据来源于:(1)网站行为数据,对读者获取电子资源行为的数据进行收集。(2)图书馆内的传感器网络,针对所处环境进行感知并不断生成数据。(3)图书馆资源中嵌入 RFID 所生成的数据。(4)通过移动设备获取读者移动位置与个人行为数据[1]。

电子文献服务作为图书馆的重要服务,其正处于大数据环境。在提供电子文献服务的过程中,单条或少量数据并无价值可言,但海量的数据则蕴涵着趋势和规则,将已有非结构化、半结构化与结构化数据进行整合分析,能为图书馆读者服务提供可供借鉴的有益信息。

后文详细叙述了设计的研究模型,并在其后依据其模型框架进行了以贵州师范大学为实验点的模型实验。模型实验中进行了推送服务,推送后有效文献比例提高了约 6.2%。

1 获取电子文献研究现状

电子文献作为科研工作者即时获取信息资源的重要渠道一直备受关注,其较纸本资源具备便利、快速、易检索等突出优势[2]。美国犹他州州立大学图书馆针对学校教师进行了电子资源意识评估。评估发

本文系贵州师范大学资助博士科研项目“基于 MDT-FHMIPv6 协议 IPv4/IPv6 虚拟机微移动迁移系统研究”,贵州省高校人文社科研究基地招标项目“喀斯特景区环境管理信息系统开发研究”(编号:12JD059)和贵州省科学技术基金项目“‘贵州省(州、地)级地方志全文数据库’研究与建设”(编号:黔科合 J 字 LKS[2010]50 号)的研究成果。

现，相对纸本资源，教师群体更偏重于电子资源的使用[3]。北京大学图书馆在 110 周年馆庆中公布的近五年读者借阅情况统计数据，亦得出了阅读电子读物读者量逐年增长、阅读纸质读物读者量逐年减少的结论。可见读者针对电子资源的使用意识在逐年加强，特别是科研工作者获取电子文献的意识明显提升[4]。相关研究表明，自然科学方面的读者使用电子资源的比例高于人文社科方面的读者，本科生使用纸本资源的比例远高于教师和研究生[5]。

虽然电子文献资源日益得到重视，国内读者使用电子文献资源的状况却不容乐观。张谦，张大庆 2005 年针对广东地区中山大学、深圳大学、华南理工大学、华南农业大学、华南师范大学、广东工业大学、暨南大学、广东教育学院等 10 所院校图书馆进行了调查访问，读者普遍反映存在电子文献资源数据库重复建设率较高、统计功能不能满足实际需要、外文电子文献资源获取困难等问题。最终调查结果显示，真正有意识地获取电子文献资源的用户只有 18%，且用户中真正得到有效阅读的读者数量更是远小于这个比例[6]。魏争光和余迎娣通过武汉大学图书馆链接了 18 个 CALIS 成员馆进行了有关数字资源组织和揭示的调查。电子文献资源利用方面存在的问题为：(1) 不同图书馆数字资源的组织和揭示发展不平衡；(2) 普遍缺乏对数字资源的深层次组织和揭示[7]。

对待电子文献资源我们应当注意到，电子文献资源不仅仅是纸质资源的数字化，还要有效地进行知识的检索和组织管理，在纷繁复杂的信息流中发现新的知识点和知识间的联系，以达到知识服务。电子文献资源越来越多地受到学术界关注，但存在的诸多不利因素应引起我们足够的重视。不利因素有来自于读者本身的，也有来自于电子文献资源建设方面的，二者相互关联、互相影响。

电子文献资源是读者获取知识信息的重要渠道，电子文献被下载的次数亦被各电子资源数据库提供商作为某文献被关注的重要参数。然而，只有下载文献被有效阅读，该文献的该次下载才是有效的。针对文献被有效阅读的具体研究，将为评估某电子文献的学术价值提供参考数据，并为实现高效电子文献资源服务提供参考依据，且有利于提高图书馆的电子文献主动服务质量，大幅度减少电子文献资源使用成本。

2 模型设计

图书馆正处于大数据产生的环境中，因其自身服务特性，图书馆成为大数据产生的母体。大数据并不等同于海量数据，大数据要求针对非结构化、半结构化与结构化的海量数据进行及时的数据分析并进行规则挖掘[8]。基于此，本文设计模型基于分布式计算插件方式，利用心理测量及相应标准进行调查读者群体选定与数据分析，进行分布式数据采集，利用树型结构进行数据挖掘，以期得到跳离于传统研究所能考虑的范围之外，针对获取电子文献行为数据进行系统、全面分析，以建立获取电子文献有效因子的数学模型。该数学模型为图书馆的高效、优质服务提供决策参考。本文用三个月内某电子文献被下载的总次数做分母，其中下载后被有效阅读的次数做分子，计算得到的数值称为获取该电子文献的有效因子。

首先对项目中涉及的关键因素类别作如下定义：

(1) 备选关键因素：针对已有数据归纳推导定义的关键因素。

(2) 指导性关键因素：针对采集数据在备选关键因素基础上扩展数据挖掘得到的关键因素。

(3) 确定关键因素：针对指导性关键因素进行定性、定量处理后的关键因素。

下面是模型针对获取电子文献大数据的处理步骤。

步骤一：针对已得读者行为数据，进行影响获取电子文献有效因子备选关键因素研究。

研究依据前期已采集到的大量读者行为数据，采用心理测量及相应标准选择读者数据，并配合调查问卷形式采集相关信息，分析数据以确定获取电子文献有效因子备选关键因素[9]。

步骤二：针对设定备选关键因素，进行扩展泛化数据挖掘研究。

在步骤一基础上，针对设定的备选关键因素进行读者行为数据采集（读者的选择采用心理测量及相应标准选择），并以备选关键因素为核心进行行为元扩充。利用分布式插件技术采集信息点读者行为数据，配合实时电子调查问卷进行汇聚泛化数据挖掘。

步骤三：新指导性关键因素加入，数据回归挖掘研究。

在步骤二的基础上，加入新确定的指导性关键因素，适当调整调查读者对象，进行二次数据采集挖掘，从而确定指导性关键因素的准确性，并可望发现新的指导性关键因素。

步骤四：更改指导性关键因素后定性、定量研究。

在步骤一、二、三的基础上，首先确定被试范围的读者群体，适当更改已确定的指导性关键因素，进行数据采集挖掘，并针对步骤一、二、三的数据进行对比研究，以期针对确定的指导性关键因素进行定性、定量研究，得出确定关键因素。模型采用分布式计算插件实现，保证了大数据环境内数据处理的横向扩展，有利于海量数据的技术分析处理。

采用的技术路线图见图 1。

3 模型实验

本文设计的研究模型是普适性的，因实验条件所限，本文模型实验仅针对贵州师范大学实验点，并进行了小范围模型实验。

针对购买电子资源供应商提供的后台统计功能单一，不能满足研究需要的客观情况，我们开发了基于 C/S 架构的采集系统。该系统通过客户端采集读者行为数据，传送至服务器 S 端，并进行数据挖掘。项目组在贵州师范大学图书馆面向研究生、教师免费开放的数字资源教室安装了客户端，在贵州师范大学图书馆数字中心服务器上安装了系统服务器端，服务器选

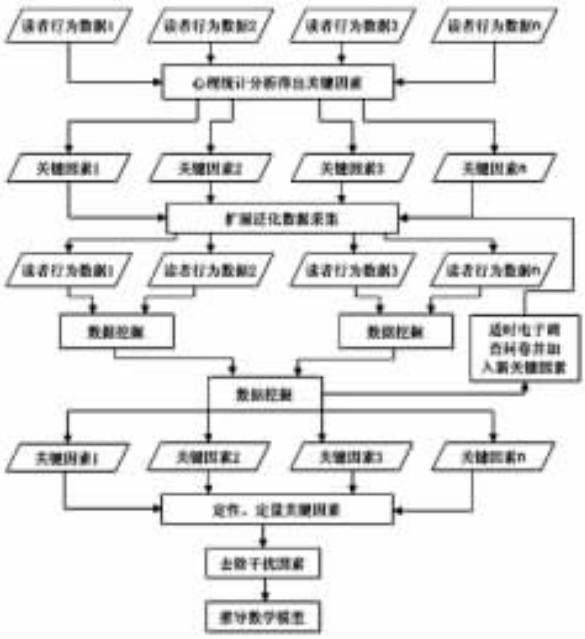

图 1　项目技术路线图

定为小型机 IBM Power720，存储设备选定为 EMCCX 340 并分配了 5TB 容量。项目组针对安装客户端的 30 台客户机，进行了数据采集。现针对 2013 年 4~ 6 月数据进行部分介绍。在此，数据采集采用数据泛采与抽样采集方式进行。表 1 反映了基本泛采情况（服务使用基本情况表，出现读者检索但未下载情况，因此平均下载时长与平均检索时长出现数据不匹配）。表 2 反映了项目组按月不定期抽查读者情况。抽查方式为，项目组在读者使用完毕离开数字资源教室时，通过查看服务器（服务器提供了 Web 访问功能），选择曾进行过搜索下载行为的读者群体。项目组在征得读者同意后，登记基本信息，并向其电子邮箱发送调查问卷，要求该读者在 1 个月内针对本次下载效果进行问卷答返。其中调查号为本次调查编号，项目组人员在选定调查对象时进行分配，并在服务器中针对本时间段该读者使用机器信息采集情况做出调查编号标识。需特别注意的是，保存文献数并不等同于下载文献数，下载文献数甚至下载文献篇名，我们皆可通过安装在客户机上的 C 端采集。而读者在离开数字资源教室时，使用 U 盘拷贝或发送至电子邮箱进行保存的文献并不是所有下载文献，该处希望读者提供其实际保存文献数。还可在后针对其做出一定说明，此处我们特别要求读者尽可能在说明栏记录下有效阅读该文献的具体时间。最后，综合泛采数据与抽样数据进行数据挖掘，获取相关指导性关键因素。实验中我们发现平均检索时长与平均下载时长，对提高有效因子成正比关系，进一步发现平均下载时长的影响度大于平均检索时长的影响度[14]，且两个指导关键因素存在互影响关系，如保持相同的平均下载时长，则具有较长平均检索时长的使用者具备较好的有效因子概率。但需要注意的是，在数据挖掘中，我们发现以上关联规则需满足一定平均下载时长，如平均下载时长超过约 1000s，则上述规则成立度明显下降。项目组依据采集数据内容挖掘出检索词关联频度、获取关联文献数量、下载文献时间点、使用间隔时间为指导关键因素（因前期工作未对挖掘出的关键因素进行定性、定量，因此未能排除干扰因素，所以在此称其为指导关键因素）[13]。

表 1　服务使用基本情况表

表 2　抽查读者情况表

项目组在前期工作中进行了主动推送服务。主动推送服务分为三级推送：一级推送服务的目的为提高读者获取电子文献资源的检索有效性；二级推送服务的目的为提高读者已获取电子文献资源的利用有效性；三级推送为后期关联性文献推送，以期提高针对读者有效阅读的持续性[13]。针对一级推送，项目组搜集了有效下载实例中读者使用的检索词，并做了相关推送。针对二级推送，项目组根据检索与下载实时的时长数据，推断出读者已下载文献中的高概率有效文献，亦做了相关推送，见表 3。以下对表中各项内容做出相关定义。

（1）检索平均时长：上次下载时间点到本次完成下载时间点的时间段做分子，上次下载时间点到本次完成下载时间点所完成的检索次数做分母得出。

（2）下载时长：本次下载时间段时长。

（3）标准下载比标差：根据调查读者认可的有效下载文献的检索平均时长除以该次下载时长所得到的数值取其平均值，为检索下载比标值；用本次文献下载的检索平均时长除以该次下载时长所得到的数值减去检索下载比标值（取绝对值），则为本项内容数值[14]。

（4）检索词数标差/ 80：根据调查读者认可的获取有效下载文献所输入的检索词个数（多个检索词检索视为一个检索词检索记数），通常为检索次数，取其平均值，为检索词数标值；用本次文献下载所使用的检索词个数减去检索词数标值（取绝对值）除以 80（该 80 为影响缩放比例），则为本项内容数值。

（5）推送率：1 减去检索下载比标差与检索词数标差/ 80，拥有较大推送率值的文献较可能是高概率出现有效文献。项目组经过调查数据计算得到检索下载比标值为 0. 12，检索词数标值为 5。从表 3 数据我们可以得出文献 2 有较高可能成为有效文献（工作中推送率达到 0. 9 以上的文献进行主动推送）。

表 3　推送读者获取电子文献行为次表

针对三级推送，项目组对关联文献作了数据挖掘，例如，文献 1 被下载并被有效阅读后，读者后继以达到预定概率获取了电子文献 2，则我们在后继其他

读者获取文献 1 后，即可主动向其推送文献 2，因前期工作挖掘数据未能满足量度需求，关联文献挖掘未达到预期要求。该情况说明，项目挖掘需具备一定的数据母本，这正是项目拟采用分布式泛采的原因。三级推送特别需注意的是推送时间，即文献 2 的推送时间越靠近文献 1 被有效阅读的时间点，该次推送才越可能成为一次有效推送。为解决推送时间问题，我们针对 100 篇有效文献做了统计分析，见表 4。将表 4 转化为折线图，为图 2，横坐标为时间轴，以日为单位；纵坐标为被有效阅读文献数量，以篇为单位。其中 96 篇文献有效阅读时间出现在下载后 10 日以内，其他 4 篇文献被有效阅读时间为 15 日、34 日、42 日、58 日。利用图 2 的统计数据，计算文献被有效阅读的数学期望[13]。

计算公式为：$\sum_{n=1}^{58} n \cdot \frac{n_n}{100}$

注：n 取值为 1~ 58 日 100 篇文献被有效阅读的时间段，n 篇为第 n 天被有效阅读的文献数量。

经计算得到值为 4.91 日（据问卷调查得出，文献推送时间在有效阅读时间后推送，可取得较好效果）。因此第一次推送定义为下载后的第 5 日。而通过统计数据我们亦发现第 5 日为下载后大部分文献被有效阅读的高峰时段，至此文献被有效阅读已达 84%。因此，项目组将例中文献 2 的再次推送时间定为下载后的第 6 天。该项数据研究中经数据挖掘我们还发现，文献针对同一读者有可能出现多次有效阅读，后续研究我们将对其多次有效阅读做权重评估，如未达到影响标值，则作为干扰因素予以去除，否则需考虑其对推送行为的影响。

表 4　有效文献阅读时间表

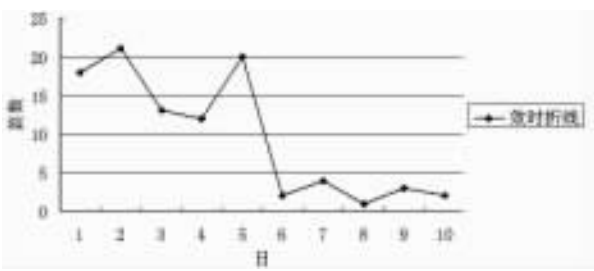

图 2　有效文献阅读时间图

经过相关推送处理（如图 3），并再次进行推送服务后相关读者的读者调查，有效文献比例提高了约 6.2%。

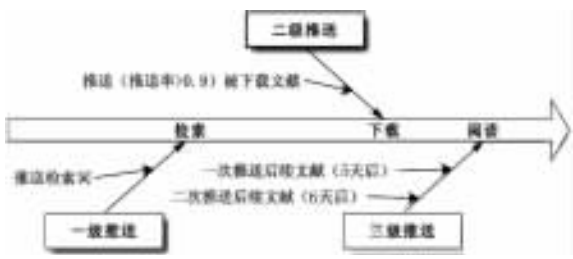

图 3　推送步骤图

4　结语

本文设计了针对大数据环境下获取电子文献有效因子的研究模型，并在小范围内基于分布式采集模式进行数据收集，完成数据挖掘。实验取得了一定的预期效果。下一步工作旨在现有工作基础上增加采集的读者行为数据，加深拓宽大数据采集处理程度与范围。可望在分析大量泛化数据后得到更为确切的科学论断，进行指导性关键因素定性、定量，从而定义确定关键因素，并推导得出具备较好普适性的获取电子文献有效因子数学模型。

参考文献

[1] 樊伟红，李晨晖，张兴旺，等. 图书馆需要怎样的"大数据"[J]. 图书馆杂志，2012(11)：63-68.
[2] Mihlrad Leigh. The Journal Usage Statistics Portal(JUSP): Helping libraries measure use and impact[J]. Journal of Electronic Resources Librarianship, 2012,24(3):229-230.
[3] Weingart Sandra J, Anderson Janet A. When questions are answers: Using a survey to achieve faculty awareness of the library's electronic resources[J]. College & Research Libraries, 2000, 33(3):127-135.
[4] Bashorun M Tunji, Isah Abdulmumin M Y Adisa. User perception of electronic resources in the university of Ilorin, Nigeria[J]. Journal of Emerging Trends in Computing and Information Sciences, 2011,2(11):554-562.
[5] 杨　毅，邵　敏，李京花，等. 电子资源建设与利用的读者调查——由读者调查结果分析读者利用电子资源的方式与倾向[J]. 大学图书馆学报，2006(6):39-48.
[6] 张　谦，张大庆. 高校图书馆数字文献资源读者利用状况分析及其对策[J]. 图书馆论坛，2006,26(3):204-207.
[7] 魏争光，余迎娣. 我国高校图书馆数字资源组织和揭示现状与分析[J]. 图书馆学研究，2004(11):50-52.
[8] Vaughan Jason. Investigations into library Web-scale discovery services[J]. Information Technology & Libraries, 2012,31(1):32-82.
[9] Michele Behra, Rebecca Hill. Mining e-reserves data for collection assessment: An analysis of how instructors use library collections to support distance learners[J]. Journal of Library & Information Services in Distance Learning, 2012,6(3): 159-179.
[10] Kao S-C, Chang H-C, Lin C-H. Decision support for the academic library acquisition budget allocation via circulation database mining[J]. Information Processing and Management, 2013, 39: 133-147.
[11] Francis Decroos, Kris Dierckens, Vincent Pollet, et al. Spectral methods for detecting periodicity in library circulation data: A case study[J]. Information Processing & Management, 1997, 33(3):393-403.
[12] Toshiro Minami, Eunja Kim. Data analysis methods for library marketing[C]. FGIT'09 Proceedings of the 1st International Conference on Future Generation Information Technology, 2009: 26-33.
[13] Durante Kim, Wang Zheng. Creating an actionable assessment framework for discovery services in academic libraries[J]. College & Undergraduate Libraries, 2012,19(2):215-228.
[14] San-Yih Hwang, Ee-Peng Lim. A data mining approach to new library book recommendations[J]. Lecture Notes in Computer Science, 2002, 2555:229-240.
[15] Zhang Zhongfei, Guo Zhen, Pan Jiayu. Amultiple-instance learning based approach to multimodal data mining[J]. International Journal of Digital Library Systems, 2010, 1(2):24-42.

[作者简介] 陈　俊，男，1979 年生，贵州师范大学图书馆副教授。
张伟云，女，1955 年生，贵州师范大学图书馆研究馆员。
赵　捷，男，1969 年生，贵州师范大学图书馆副研究馆员。

收稿日期：2013-08-21