

doi:10.3772/j.issn.2095-915x.2015.02.011

大数据环境下图书馆信息服务模式探析

王丽敏

(中国矿业大学(北京)图书馆 北京 100083)

摘要: 本文首先阐述了大数据的概念、特征和技术,介绍了现阶段大数据的应用现状,探讨了大数据时代图书馆信息服务的模式和方向,最后指出了图书馆在应用大数据时需要注意的问题。

关键词: 大数据,图书馆,学科信息门户服务,个性化信息推送服务

Research on Library Information Service Patterns in Big Data Environment

WANG Limin

(Library of China University of Mining & Technology, Beijing 100083, China)

Abstract: The article states the concept, characteristics and technology of big data; introduces the current application of big data; discusses the pattern and direction of library information service in the era of big data; and finally points out the problems which should be paid attention to in the application of big data.

Keywords: Big data, library, subject information gateway service, personalized information recommendation service

早在1990年,数据仓库之父比尔·恩门(Bill Inmon)就开始关注大数据,但最早提出“大数据”概念并被广泛传播则始于2008年9月《自然》推出大数据专刊,对大数据的起源、特征、组织形式、长久保存和利用进行了介绍,认为需要进行海量数据处理的天文学和基因学创造出了“大数据”概念。2011年《科学》推出数据处理专刊,认为对科学学科而言,大数据既是挑战也是机遇。谷歌利用用户的海量检索关键词近乎实时地估测出世界上不同国家和地区的流感传播情况,并与官方数据相关度高达97%^[1]。科学数据处理的需要,数据的指数增长,学术界的研究及大数据应用,说明大数据时代已经到来。

1 大数据概述

1.1 大数据的涵义

目前关于大数据的概念还没有标准的定义,Informatica中国区首席顾问但彬认为:大数据包含了海量数据的含义,而且在内容上超越了海量数据,可以简单地描述为“海量数据+复杂类型的数据”。NetApp认为大数据包括A、B、C三个要素:大分析、高带宽和大内容。大数据分析帮助用户获得更高的资源价值,高带宽使数据处理更快速,大内容是指信息不丢失并实现高扩展性。Gartner公司提出:大数据是高速度、大容量和多样化的信息集合,需要新的处理方式以提高决策能力、洞察力,实现流程优化。综合众家之言,大数据描述的是随着数据量和数据类型激增而出现的一种大规模、多样化的数据集,及其对数据集高速采集、分析、处理以提取知识价值的技术架构与过程^[2]。

1.2 大数据特征

大数据如下四个特征。(1)大量化(volume):

数据量巨大,一般指在10TB规模以上的数据。遍布世界各个角落的电脑、传感器、移动设备、在线交易和社交网络每时每刻都在产生大量数据。大数据遍及各行各业,企业、机构甚至个人的活动都在产生和积累大量数据。(2)多样化(variety):数据类型繁多,包括结构化数据和非结构化数据。大数据囊括了网络日志、视频、声频、图片、地理位置信息、交易信息、科学数据集等。目前非结构化数据已占数据总量的80%以上^[3]。(3)快速化(velocity):对大数据快速处理,才能了解迅速变化的环境并做出反应,才能快速制定出合理准确的应对策略。(4)密度低(veracity):价值是大数据的终极意义所在,大数据的价值密度很低,比如用户观看2小时的视频资源,其中有价值的往往只有几十秒。

1.3 大数据技术

依据大数据生命周期阶段可以将大数据技术分为大数据抽取与集成、大数据存储、大数据挖掘与分析、大数据解释与展示四个方面。大数据挖掘与分析是整个大数据处理流程的核心,这一部分能够发现数据的相关关系,产生大数据价值。大数据分析(Big Data Analytics)就是将先进分析技术应用于大数据。TDWI最佳实践报告认为,根据忠诚度和技术发展可能性,高级分析技术(预测分析、数据挖掘、统计分析、复杂SQL、人工智能、自然语言处理)和可视化发现技术在现在和未来3年都将是大数据分析领域的关键技术;Hadoop分布式文件系统(HDFS)、MapReduce、复杂事件处理(CEP)、云计算技术虽然出现晚,但发展前景很好^[4]。大数据解释与展示是面向用户的大数据技术,帮助用户理解大数据的分析结果。由于大数据分析结果是海量的,结果之间关联复杂,常采用可视化技术和人机交互技术。

2. 大数据研究现状

目前,大数据研究与应用已经成为计算机与信息技术领域中的热点问题。大数据理论与应用提出后受到社会的广泛响应,世界各国都高度重视大数据的研究、探索与应用,西方不少国家推出了国家战略层面的研究规划。2012年3月,美国政府拨款2亿美元启动了“大数据研究和发展计划”,致力于提高从大型复杂数据集中提取知识和洞见的能力。2012年4月,美国、英国、德国、澳大利亚和芬兰研究者联合开展“世界大数据周”活动,极力推动政府制定大数据发展战略措施。2012年7月,日本推出重点关注“大数据应用”的“新ICT战略研究计划”。我国对大数据研究同样重视,科技部发布的“十二五”国家科技计划信息技术领域2013年度备选项目征集指南中,关于大数据的研究(存储设备)排在第一位^[5]。

随着IT业巨头如IBM、Microsoft、Google、Oracle等跨国公司相继投入资金,推动大数据处理技术的快速发展,大数据的分析和研究在许多领域取得了巨大成功。与此同时,大数据也已引起图书馆、信息与情报界的广泛兴趣,与大数据关联紧密的网络计量学和文献计量学因此得到了快速发展。图书馆学研究者也广泛参与到诸多的大数据研究项目中。美国亚利桑那州立大学、伊利诺伊州立大学、密歇根州立大学都已经开设了大数据相关的课程和研究方向。例如,亚利桑那州立大学围绕数据迁移、数字格式和元数据主题开设了数字馆藏课,内容包括与数字馆藏管理有关的资源采购、政策发展和组织结构、内容存储等。在应用实践领域,2012年4月,哈佛大学图书馆准备把由73家分馆数据组成的图书馆大数据公之于众,具体涵盖1200多万种资料^[6]。清华大学利用元数据仓储进行数据挖掘,建立可检索多种数据源的检索平台,并利用关键词分析作者与合作者的关系,建立知识关联网络^[7];2013年10

月25日,北京大学图书馆承办了“大数据时代数字图书馆的变革与创新”学术研讨会。

3 大数据时代图书馆的信息服务模式

大数据正在成为当前图书情报领域研究的热点之一,作为当今社会知识信息服务中心的应以积极的姿态迎接大数据时代的到来,积极调整信息服务的模式和方向应从以下四个方面着手。

3.1 建立学科信息门户服务模式

大数据时代的文献资源建设要求图书馆的信息服务必然向着学科化服务方向发展,建立学科门户服务模式十分必要。即针对某个学科领域,根据特定的资源选择标准对其进行规范性描述与组织,收集具有一定学术价值的网络资源,进行组织整理后再面向用户提供检索及导航等增值服务。学科信息门户的形式包括以下几种:(1)以网络学科信息导航为主的主题信息门户,其主要作用是提供网络信息资源选择、描述及检索服务,并且要保证资源的权威性、可靠性、可持续性等特点。(2)专业信息门户,该形式就是按照图书馆信息服务的具体要求,以图书馆服务系统为基础,面向用户提供一种组合各类资源的信息服务。

(3)跨学科信息门户,顾名思义即可以提供多个学科信息门户间的互相检索、选择及描述。(4)分布信息门户,以信息门户体系为基础,整合若干分布门户,便于读者通过统一的信息门户体系进行信息资源的检索及调用。(5)开放信息门户,这种学科门户服务形式实现了学科信息门户资源与服务的集合,并且可以根据用户的个性化需求提供信息门户定制服务,按照不同的逻辑业务流程对多个信息服务环节进行整体规划。

3.2 建立智能化的信息服务模式

图书馆智能化信息服务不仅仅指的是解决对海量数据进行智能收集及处理问题,还指对图书

馆服务对象进行智能化的信息服务,包括信息服务应用、工具、平台的智能化推荐,让用户智能轻松获取所需结构化数据以及非结构化数据,如:文本、音频、视频、网络社交信息、现实生活资料等数据。最为重要的是,智能化的信息服务还包括知识信息的流通过程中对于隐性知识信息转变为显性知识信息的智能转化,促进知识信息由个体向多个主体进行传播和交流,利于知识的重新发现与组织。

3.3 准确高效的个性化信息推送服务模式

信息推送服务是指图书馆根据读者的需求,主动将读者所需要的信息内容推送到读者端。目前,图书馆的推送服务往往没有考虑到读者需求,而是根据图书馆所提供的服务进行推送,或者只在某些领域考虑了读者需求,在相关方面做了一些推送,并不能满足读者的全部要求,同时也没有预测性,当大数据环境下,对于读者的各类行为,例如查阅资料或者浏览网页等,可以细化到某一篇文章,甚至某一个词,将读者各类微行为都汇集到数据库中,然后根据不同层次人群的兴趣爱好进行详细分析,分层次进行准确地推送服务,另外,根据读者的行为热点分析,大数据能够告诉图书馆读者的兴趣变化曲线,以预测读者需求,从而为图书馆的资源采集、资源更新提供依据,最终达到超前读者需求、引导读者需求行为的效果,大大提高图书馆的社会效益。大数据时代下,数据资源迅速膨胀,人们能够获取的真正有用的信息却越来越少,而这个个性化信息服务推送模式恰恰可以最大程度上满足用户的信息需求,且过程更加主动,为图书馆实现细粒度个性化服务提供了可能,真正体现出“以用户为中心”的服务理念。

3.4 深度的参考咨询服务模式

传统的参考咨询服务主要是利用图书馆现有

的参考工具书、检索工具书、外购数据库和自建数据库,由有一定专业知识和熟悉检索工具的专职参考馆员来开展此工作,在大数据时代,咨询馆员可以将符合大数据特征的咨询课题融入大数据系统,系统将根据咨询主题,利用数据分析技术,可视化技术,瞬时以图形展示查询结果,从而缩短了咨询馆员获取资源的时间,扩大了咨询馆员对资源的获取面,从而最大限度的满足读者(特别是企业方面的读者)需求,提高了参考咨询的效率,并使图书馆深度参考咨询服务成为可能。

4 应用大数据应注意的问题

4.1 大数据格式的统一问题

高校图书馆大数据中不同格式的数据越来越多,价值密度低,只有通过统一格式,进行整合才能发挥出最大价值。这就需要通过来自不同结构资源的海量数据进行抽取、映射、收割、导入等手段进行预收集,归并映射到一个标准表达式,进行预聚合及融合,形成格式统一、内容丰富、结构清晰的数据,灵活构建各种分类和界面,按照知识本体进行组织和揭示,进而保障强大高效的检索能力和良好的结果相关度排序。

4.2 大数据的高效率、安全存储问题

大数据的高效率存储是高校图书馆面临的一个关键问题。随着数据库技术以及云计算技术的迅猛发展,大规模数据存储要借助非关系型的数据库分析技术——NoSQL、MapReduce 和 Hadoop。这些技术简单易用,能够并行处理大规模数据,非常适合非结构数据的处理,也成为大数据分析领域的主流技术。大数据存储也可采用基于云计算的分布式存储技术,利用分布式的数据云存储技术和与之相关的虚拟技术能够使高校图书馆的数据更加统一有序,使通过网络进行的访问与存储更加方便快捷。数据存储的安全性可

以从两个方面采取措施:一是个人数据保存多久,被用于什么用途将被设置权限;二是高校图书馆要树立良好的职业形象,完善保障机制。读者对自己的检索历史等数据有知情权,可以决定自己的数据是否被图书馆利用,以及被用于什么用途,以及久之内删除这些数据。解除读者对个人隐私的顾虑,让高校图书馆能够合法合理地采集到更多的读者数据。

4.3 非结构化数据的处理问题

大数据中的数据信息复杂多样,传统的数据分析和挖掘对关系型数据,非结构化的、半结构化的数据力不从心。高校图书馆中待处理的非结构化数据与读者兴趣密切相关,通过分析为读者提供个性化服务非常重要。当前使用的方法是基于内容的推荐方法和协同过滤的推荐方法 [8]。协

同过滤法是目前最为成功的算法,但也存在着读者兴趣变化、数据稀疏性、读者评分的真实性及差异性问题,根据读者信息数据构建个性化读者行为模型,结合基于读者行为的协同过滤算法,挖掘模型中存在的规则,从而产生个性化服务值得研究。

“大数据”这一概念已经成为时代的关键词之一,从大数据中去捕捉、分析并向读者推送有潜在价值的数据,将成为大数据时代图书馆的一大主要业务,并且这些业务开展的水平也将决定着大数据时代的图书馆发展水平及方向。因此图书馆要把握住这次机遇,找准切入点,坚定图书馆为社会和高校科研服务的宗旨,扩大服务面,提高服务质量,开创新的服务方式,以更开放的服务模式进一步提升图书馆的社会价值,使其真正成为社会的知识中心、学习中心和文化中心。

参考文献:

- [1] GINSBE R G J. Detecting influenza epidemics using search engine query data[J]. Science,2009 (457): 1-5.
- [2] 刘高勇,汪会玲,吴金红. 大数据时代的竞争情报发展动向探析 [J]. 图书情报知识,2013(2):105-111.
- [3] 姜山,王刚. 大数据对图书馆的启示 [J]. 图书馆工作与研究,2013(4):52-54.
- [4] RUSSOM P. Big data analytics [R]. TDWI Best Practices Report, 2011 (Fourth Quarter).
- [5] “十二五” 国家科技计划信息技术领域 2013 年度备选项目征集指南 [EB/OL][2014-03-06].http://www.most.gov.cn/tztg/201203/t20120329_93437.htm.
- [6] WATTERS A.Strata Week:Harvard Library releases big data for its books[EB/OL].[2013-05-11].<http://strata.oreilly.com/2012/04/harvard-book-data-clouder-hadoop-splunk-ipo.html>
- [7] 邓景康. 大数据环境下清华大学图书馆的实践 [N]. 中国新闻出版报, 2013, 8(29):5.
- [8] 朱静薇,李红艳. 大数据时代图书馆的挑战及其应对策略 [J]. 现代情报,2013(5):9-13.