基于 Hubble. net 的仓储系统设计与实现

杨伟超,马增军,耿 卫
(中国人民解放军信息工程大学 图书馆,河南 郑州 450001)

摘 要:科学技术的飞速发展,各类知识信息的爆发式增长,对高校的专业设置及其教学内容均带来了巨大的影响。图书馆作为学校文献资料的综合保障中心,面临着知识爆炸的严峻挑战。为此,借鉴当前最新的搜索引擎技术,设计并开发实现了基于 Hubble. net 的仓储资源服务系统。该系统根据数字资源存储的特点,通过对图书馆资源进行对接整合并构建面向读者的数据库系统来存储和索引数据,同时内置了 Web 服务功能,以实现对图书馆内部电子信息与文献资料资源的充分利用。运行情况表明,所设计构建的仓储服务系统能够满足用户快速获取知识的需求,具有较强的稳定性和实用性,能够更加便捷有效地为图书馆用户提供高质量的服务。

关键词:Hubble. net;搜索引擎;仓储管理;海量数据
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2017)10-0181-04
doi:10.3969/j. issn. 1673-629X. 2017. 10. 038

Design and Implementation of Massive Data Storage Service System Based on Hubble. net

YANG Wei-chao, MA Zeng-jun, GENG Wei
(Library, PLA Information Engineering University, Zhengzhou 450001, China)

Abstract: The rapid development of science and technology and emergence of various types of knowledge and information on the outbreak of growth have brought a huge impact on classification of discipline in universities and colleges. As a comprehensive safeguard center of the school literature, the library faces the severe challenge of knowledge explosion. Therefore, in reference of latest search engine technology, the storage resource service system based on Hubble. net integrated with library resources in accordance with the characteristic of digital resource storage is designed and implemented to construct a database system for readers. The built-in Web services capabilities have been employed the internal library of electronic information and document resources fully have been utilized. The operation status shows that it has met the needs of the users to get the knowledge rapidly with higher stability and practicability and has provided the library users with high quality service more conveniently and efficiently.

Key words: Hubble. net; search engine; storage management; mass data

0 引 言

信息社会数字资源极大丰富,呈知识爆炸的状态,如何使用庞大的知识资源库,为读者提供有效的服务,而不是让读者迷失在知识的海洋中,是长期数字文献资源服务工作中一直存在的难题。近年来,随着搜索引擎和云存储的飞速进步,出现了解决这一难题的曙光。

系统建设中专门成立课题组,充分考虑图书馆资源的特点,认真进行用户需求分析和详细设计,组织专家、教授对建设方案的内容和系统设计进行严格调研论证,明确每个成员的具体分工。实施阶段,成员各司其职,积极行动,把学到和掌握的新技术、新知识毫无保留地应用到系统设计中。经过多次争论,对于系统的结构和功能进行多次调整,使其趋于合理完善。在系统建设过程中,严格遵循技术规范,保证系统开发的规范性、统一性和标准化。

系统研制开发力图兼顾图书馆和读者两方面需求,尽可能保证有更好的稳定性,更快的检索速度,更强的兼容性;具有统一安装维护界面和一体化发布系统;提供多种数据库配置方案,方便灵活地增加检索数据库;提供友好简捷的管理系统,实现异构资源统一检索系统的一体化管理。

收稿日期:2016-04-11 **修回日期**:2016-08-04 **网络出版时间**:2017-07-11
基金项目:河南省科技攻关项目(132102210244)
作者简介:杨伟超(1980-),男,硕士,馆员,研究方向为信息技术、数据库。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20170711.1452.012.html

文中创新点在于:站在全局的高度,开展多方位、多层次的调查搜集,深入了解各种统一检索系统的特点及其变化趋势;集成了业界先进搜索引擎技术,建立专用的索引结构,使千万级的数据秒级响应;采用专门的采集和数据转换工具,快速转换异构数据,方便数据更新和管理;开发出章节检索和全文检索功能,向用户提供深度的知识点搜索功能;提供广泛参与的服务,为专业学科、教师提供建站云端,可以使更多的人或组织快速搭建专业的站点,提供专业的服务。

1 系统搜索引擎技术

1.1 Hubble. net 搜索引擎技术

系统的搜索引擎是基于 Hubble. net 开发而成的。Hubble. net[1] 是基于. net framework 的全文搜索引擎。目前关系数据库提供全文搜索的功能相对较弱,不能很好地满足实际应用需要,而一些组件只提供了全文搜索功能,缺乏和关系数据库的关联。Hubble. net 整合了全文搜索和关系数据库,可以方便地通过 SQL 语句对数据库进行全文搜索,Hubble. net 提供了一个和对应关系数据库的映射关系,通过 SQL 语句操作 Hubble. net 的数据库和数据表,Hubble. net 将自动和对应的关系数据库实体进行关联。Hubble. net 提供 Index cache,Query cache,Data cache 三种级别的缓存方案,设计了非常完善的并发控制机制,用户的增删改查可以同时进行,不会存在任何冲突。

1.2 Hubble. net 和 Lucene. net 的对比

Hubble. net 采用被动方式 Append Only 模式对数据库现有表进行索引。Lucene. net[2] 则是从数据库读取记录进行索引[3-4],数据存储在 Lucene. net 索引中。Hubble. net 以系统服务存在,不会像 Lucene 那样和应用程序共用内存。Hubble. net 设计了一套内存管理机制,设置最大内存使用数量,一旦 Hubble. net 使用内存超过这个数量,Hubble. net 就会自动启动内存整理程序,将一些不经常使用的缓存从内存中清理掉以腾出更多的内存空间给用户。多关键字情况下,Hubble. net 比 Lucene. net 具有明显的优势,Match 方法大概比 Lucene. net 快 5 ~ 10 倍,而 Contains 方法则比 Lucene. net 快上百倍。在单个关键字时,Lucene. net 和 Hubble. net 的搜索速度是接近的,但随着关键字的增多,两者的差距明显增大,Hubble. net 具有明显的优势。

2 系统总体设计

系统建设中主要基于“以用户为中心,综合考虑应用现状和发展需求,重点突出”的设计思想,根据高校图书馆的现状和需求分析,数字图书馆的建设将重点实现种类繁多、异构、分布式资源[5]数据库的整合服务;用 Hubble. net 作为其核心搜索模块,在成熟、经济、易用、可扩展、可持续发展等因素综合平衡下进行应用系统的选型和建设。

整个系统分为采集层、数据层、应用层和服务层四部分,如图 1 所示。

采集层:完成各种类型数据源的元数据的采集入库工作,并逐步实现部分全文数据的采集入库工作。

数据层:对所有采集来的元数据及数据实行统一

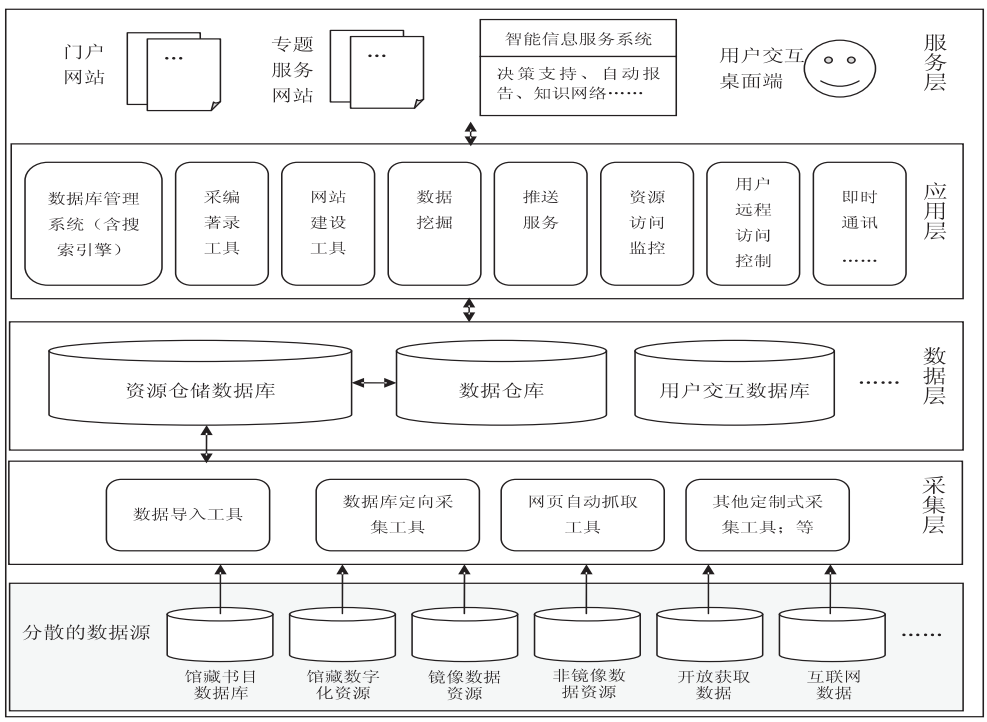

图 1 系统架构

的仓储化管理,统一放入资源仓储数据库,为实现资源的整合发布奠定数据基础。仓储化管理意味着所存储的不仅包含文字,还包括图片、文档、视频、flash 等各类数字对象。在资源仓储数据库的基础上,可建立面向主题的数据仓库,实现各种专题服务。通过对用户访问行为[6-7]的动态监控,以及通过交互软件积累下来的用户行为信息,形成用户交互数据库。此外,还包括实际建设过程中产生的其他各类数据库。

应用层:是从数据到服务的中间层,对数据进行加工处理,为读者提供多样化的数据服务模式和服务内容;此外,还包括针对用户需求开发的其他方面的服务。

服务层:读者获取资源和服务。可通过网页、桌面端以及信息系统等多种形式获取服务。

从以上四个层面有序开展建设,不但完成文献资源的整合、存储管理与网站发布等功能,还对整合资源进行多角度挖掘、加工处理,实现从数据到有价值信息(知识)的智能转化、智能提供,为用户提供高层次的信息服务;针对用户对特殊文献资源的需求,形成以全方位用户服务为核心的图书馆在线服务系统;将对特殊文献资源有着同类需求的用户[8],形成用户社区,为相互之间沟通交流、互助提供平台;针对图书馆对文献资源的管理与研究需求,可开发相应的研究系统或者工具等。最终形成一个以文献资源池为核心的、拥有活跃用户群体的一体化服务系统。

3　系统功能设计及运行试用

海纳仓储系统根据数字资源存储的特点,自主设计了面向读性能优化的数据库系统,来存储和索引数据,并内置搜索引擎[9-10]和 Web 服务。该系统基于 64 位系统,优化支持多线程,可充分发挥多 CPU 以及大内存的优势。

3.1　系统核心服务

(1)采集监视服务。

采集系统能够有效实现对列表式资讯类页面的定期自动监视和采集,采集结果统一进入仓储数据库。对入库数据可以即时发布,也可以利用数据采编工具进行编辑后再发布。采集系统采用先进的网页分析与提取技术,使用者只需进行简单的配置就可以实现对资讯类页面的有效抓取。

(2)数据采编、著录服务。

数据采编工具用来录入、修改数据库数据。该工具支持对 word、pdf、图片、影音文件等多类数据源的采编。既可以用于对数字资源进行方便的标引工作,支持截取封面图、数据源作为关联数字对象上传等功能;又可以用于网站各类动态信息、静态信息的录入、修改等。简单地说,既可以作为图书馆的编录系统,又可以作为网站的内容管理工具。

(3)快速搭建虚拟专题库服务。

可以方便地从实体数据库中通过关键词检索与分类检索、检索点检索、聚合检索等方式,迅速从仓储数据库中抽取所需数据,组成虚拟数据库。且生成的虚拟数据库可以便捷、迅速地发布到网页上。借此功能,使得建立各类专题数据库成为轻松容易的事情。

(4)分类聚合服务和特征聚类服务。

海纳仓储系统,采用渐进深入的搜索模式,提供对海量资源的强大搜索能力。用户使用简单方便,对检索词的命中分布一目了然,通过渐进深入的分类限定和特征限定可以快速缩小搜索范围。

(5)快速建站发布服务。

快速自助建站系统用于完成网页界面的搭建与服务模块的添加与管理维护。通过仓储管理与发布服务器将网页建设系统所生成的文件,解析成 HTML 文件,并且从数据仓储管理系统中提取网页建设系统中所配置服务模块指定的数据。快速建站系统在页面搭建方面具有强大优势,可以自由、灵活、快捷地构建个性化网页框架,在此基础上按需添加功能模块。快速建站系统在数据仓储管理方面,采用高速内存技术结合先进的倒排索引技术,支持海量数据的高并发、快速搜索服务;该搜索服务采用全新展示模式,只需输入一次检索词,便可一次性搜索出在各栏目下、不同检索点、不同数据仓储库以及各特征聚类点的检索结果集,只需在检索结果界面上点击,便可即时切换查阅各组合检索结果集,真正实现快捷、高效、一站式搜索服务。

(6)资源访问监控服务。

资源访问监控服务可以有效监控用户的网页、数字资源访问行为。对恶意下载等情况进行报警、追查封死 IP 等。可以按照多种方式对数字资源(网页)访问情况进行统计分析、自动生成统计报告。基于对用户访问行为[11-12]的记录和挖掘,可以进一步实现知识挖掘等服务。

(7)学术直通车服务。

基于数字资源仓储管理服务系统,为了能更好地为用户提供个性化服务[13-14],开发了专门的个人桌面服务客户端(学术直通车)。学术直通车集成了用户统一认证功能、访问代理服务功能、单点登录功能、跨库检索功能、数字资源仓储发布网页浏览功能。利用该客户端,可以有效实现数字资源的跨区域访问。

3.2　前台功能实现

资源检索提供输入框,只需输入检索词,在搜索框下列举整合数据库中的检索结果。比如“军事期刊(26692)”表示在军事期刊中所有检索点中的命中数

为 26 692 条。只需点击数据库,系统自动定位到该数据库的检索结果集。中心部分显示了所选数据库的检索结果。如图 2 所示,在该库所有检索点中共计命中 26 692 条,在所有整合数据库的所有检索点中共计命中 32 088 条。而搜索耗时小于 0.01 s。

3.3 后台管理

仓储服务器实现海量数字资源的仓储化管理。仓储管理工具实现对仓储服务器的管理应用。主要功能包括:

(1)物理数据库管理:新建、克隆、修改、删除后台数据库;对数据库进行预处理、索引、清除记录等操作;对数据库记录进行浏览查看、检索、增删改等操作;对数据库进行发布设置等。

(2)虚拟数据库管理:新建、删除虚拟数据库;设置、修改虚拟库提取条件,单个或批量进行数据提取、预处理、索引操作;人工对虚拟库所提取记录进行增删等操作。

(3)数据导入工具。针对不同来源的数据提供相应的导入工具。

仓储管理工具如图 3 所示。

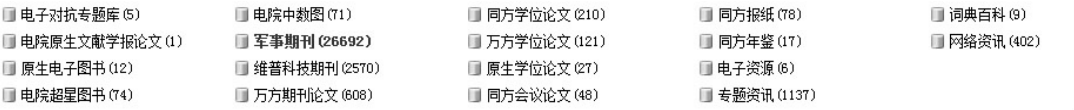

图 2 数据库检索结果

图 3 仓储管理工具

3.4 运行试用

目前基于 Hubble. net 的仓储系统已推广试用。系统采用的硬件环境:DellR710,CPU 为 6520 * 2,内存为 8 * 8 G,硬盘为 2TSAS * 6。软件环境:Win2008 R2。开发平台:基于微软的. Net 4.0 平台开发而成,开发语言为 C#。

经测试,系统搜索性能优异,数据量在 3 亿多条的情况下,搜索速度为 0.3 s,并发连接超过一千时,速度稍微下降。系统在推广使用中,学校读者整体反映良好,操作使用方便,在进行信息保障服务时,节省了广大读者查阅文献信息资源的时间,提高了查检信息资源的效率。

4 结束语

系统开发实现对文献资源元数据级的统一整合,形成统一的整合资源库;以此为基础,为读者提供统一入口的服务。基于统一的整合资源库,可以快速方便地从图书、论文、期刊、报纸、多媒体等多种形式的数据中,按照关键词、分类法及来源等抽取出相关数据,形成全方位的专题资源库,打破了以往需要到各处找数据再一条条录入的局面。这使得快速满足学科教学、读者学习成为现实。初期运行情况表明,该系统具有较强的稳定性和实用性,受到了广大用户的肯定。今

(下转第 188 页)

试者不需要提前训练,但是受试者情绪对实验有影响。

6 结束语

在传统的 SSVEP 脑机信号特征提取的过程中,使用傅里叶变换和功率谱等方法,不仅影响了系统速度、精度和传输率,而且还与受试人员的个体差异有关。为此,设计并实现了基于 SSVEP 的高传输速率脑机拨号系统。实验结果表明,该系统的确提高了系统的速度、精度和传输率,降低了受试人员个体差异性,为 SSVEP 信号的处理提供了新的思路和方法;此外,还能扩大 BCI 人群的使用范围,更利于 SSVEP-BCI 系统的推广。BCI 系统在脑科学、康复工程、生物医学工程有着广泛的应用前景,但是通信速率低依然是阻碍 BCI 系统应用的主要原因。总体而言,BCI 依然处于基础阶段,需要广大科技工作者更加努力。

参考文献:

[1] Wolpaw J, Birbaumer N, Heetderks W J, et al. Brain-computer interface technology: a review of the first international meeting [J]. IEEE Transactions on Rehabilitation Engineering, 2000, 8(2): 164-173.

[2] Chen Xiaogang, Wang Yijun, Nakanishi M, et al. High-speed spelling with a noninvasive brain-computer interface [J]. PNAS, 2015, 112(44): 6058-6067.

[3] Poryzala P, Materka A. Cluster analysis of CCA coefficients for robust detection of the asynchronous SSVEPs in brain-computer interfaces [J]. Biomedical Signal Processing and Control, 2014, 10(1): 201-208.

[4] Zhang Yu, Zhou Guoxu, Jin Jing, et al. SSVEP recognition using common feature analysis in brain-computer interface [J]. Journal of Neuroscience Methods, 2015, 244: 8-15.

[5] Chang M H, Lee J S, Heo J, et al. Eliciting dual-frequency SSVEP using a hybrid SSVEP-P300 BCI [J]. Journal of Neuroscience Methods, 2016, 258: 104-113.

[6] 赵 丽,孙 永,郭旭宏,等. 基于稳态视觉诱发电位的手机拨号系统研究[J]. 中国生物医学工程学报, 2013, 32(2): 253-256.

[7] 邓志东,李修全,郑宽浩,等. 一种基于 SSVEP 的仿人机器人异步脑机接口控制系统[J]. 机器人, 2011, 33(1): 129-135.

[8] 王行愚,金 晶,张 宇,等. 脑控:基于脑-机接口的人机融合控制[J]. 自动化学报, 2013, 39(3): 208-221.

[9] 徐光华,张 锋,谢 俊,等. 稳态视觉诱发电位的脑机接口范式及其信号处理方法研究[J]. 西安交通大学学报, 2015, 49(6): 1-7.

[10] 霍 涛,贾振堂. 基于 STM32 和 SIM900A 的无线通信模块设计与实现[J]. 电子设计工程, 2014, 22(17): 106-110.

[11] 翟 顺,王卫红,张 衍,等. 基于 SIM900A 的物联网短信报警系统[J]. 现代电子技术, 2012, 35(5): 86-89.

[12] 笪铖璐,陈志阳,黄丽亚. 基于 CCA 的 SSVEP 性能研究[J]. 计算机技术与发展, 2015, 25(5): 52-55.

[13] 薛 飞,杨友良,孟凡伟,等. 基于 Matlab GUI 串口通信的实时温度监控系统设计[J]. 计算机应用, 2014, 34(1): 292-296.

[14] 朱向荣,冯乔生,施少捷,等. 信捷 PLC 与计算机串口和以太网通信的 VC++编程技术[J]. 软件, 2015, 36(6): 75-82.

(上接第 184 页)

后,将对用户行为与仓储整合资源进行关联研究,以提高用户服务质量。

参考文献:

[1] 赵 英. 搜索引擎 Hubble. Net 的机制分析及基础应用[J]. 装备制造技术, 2011(12): 53-56.

[2] 郎小伟,王申康. 基于 Lucene 的全文检索系统研究与开发[J]. 计算机工程, 2006, 32(4): 94-96.

[3] 孙西全,马瑞芳,李燕灵. 基于 Lucene 的信息检索的研究与应用[J]. 情报理论与实践, 2006, 29(1): 125-128.

[4] A Pache Lucene6.2.1[EB/OL]. 2016-09-20. http://www.apache.org/dyn/closer.lua/lucene/java/6.2.

[5] 霍 林,黄俊文,潘英花,等. 大规模分布式资源搜索技术研究进展[J]. 计算机应用研究, 2010, 27(11): 4006-4009.

[6] 周满英,任树怀. 图书馆用户体验案例研究—以麻省理工学院图书馆实践为例[J]. 图书馆论坛, 2012, 32(6): 49-52.

[7] 包 凌,蒋 颖. 图书馆统一资源发现系统的比较研究[J]. 情报资料工作, 2012, 33(5): 68-73.

[8] 陈定权,卢玉红,杨 敏. 图书馆资源发现系统的现状与趋势[J]. 图书情报工作, 2012, 56(7): 44-48.

[9] 李学勇,欧阳柳波,李国徽,等. 搜索引擎中网络蜘蛛搜索策略比较研究[J]. 计算技术与自动化, 2003, 22(4): 63-67.

[10] 曹元大,贺海军,涂哲明. 中文 Web 文档全文检索系统的设计及实现[J]. 北京理工大学学报, 2002, 22(1): 68-71.

[11] Miller G A. WordNet: a lexical databas for english[J]. Communications of the ACM, 1995, 38(11): 39-41.

[12] Voorhees E M. Query expansion using lexical-semantic relations[C]//Proceedings of 17th annual ACM SIGIR conference on research and development in information retrieval. [s. l.]: ACM, 1994: 61-69.

[13] 郑 炜,梁战平,梁 建. 面向用户意图的智能搜索引擎框架研究[J]. 现代图书情报技术, 2014(3): 65-72.

[14] Massimo P, Takahiro K, Terry P R, et al. Semantic matching of web services capabilities[C]//First international semantic web conference. Sardinia, Italy: [s. n.], 2002: 333-347.