

●索传军 焦玉英

基于 Web 的信息服务效果的优化研究

摘 要 在因特网上寻找信息费时费力。基于 Web 的信息推送技术可以帮助图书馆开展主动信息服务。基于 Web 站点的信息服务效果的优化是一个系统工程,通过信息的收集、整理等过程,使信息系统处于优化状态,改善服务效果。参考文献 15。

关键词 Web 站点 信息服务 组织管理 优化研究

分类号 G252

ABSTRACT As everyone knows, it is time- and effort-consuming to search information on the Internet. However, the Web-based push technology can help library to provide active information services. It is a systems engineering to optimize the results of Web-based information services. 15 refs.

KEY WORDS Website. Information services. Organization management. Optimization research.

CLASS NUMBER G252

因特网上信息资源丰富但利用难的问题,一直困扰着人们。这不仅是因特网自身的因素,它涉及到与之相关的各个方面。这些问题可以简单地归纳为三方面:网络自身技术问题,网络管理问题和网上信息的开发利用问题。从理论上讲,每一方面问题的解决,都会有效地促进网上信息资源的利用。但前两方面问题的解决,涉及到的因素很多,而且还有很大的技术难度,解决起来很困难。本文主要从信息管理科学的角度,论述如何提高网上信息服务效果的方法。

1 影响因特网上信息资源有效利用的因素分析

1.1 影响网上信息利用的客观因素

通常,人们是利用某个搜索引擎(如 Yahoo, OpenText)来查询因特网上信息的。查询的结果,除与所使用的搜索引擎直接相关外,还与该搜索引擎网站的机器人性能和它的信息资源索引文档的组织结构有关,并与网上主页制作标引的规范程度有很大的关系。

网站是信息存储的单位,主页是网上信息发布的基本单元。目前的 Web 页面大都为 HTML 文档,文档中的 meta 标记用于提供该网页的标引词,以供广大搜索引擎

中自动标引程序识别并抽取,组织搜索引擎中的倒排档。HTML 中 meta 标记中间的词既有 HTML 文档自动生成的,也有 HTML 文档制作者赋予的,前者大都是有关文档属性的描述,后者赋词的随意性很大。一个网站的好坏、利用率的高低,与主页上 meta 标引的规范程度和对网站信息内容的揭示程度有很大关系,从信息利用角度看,这是评价一个主页质量的最重要的指标。因为,一个主页缺少必要的反映其资源内容的元数据,网络机器人就难以获得准确或充分反映它的信息,这样机器人就会对它错误地标引和分类,一旦这样的信息被添加到搜索引擎网站的数据库中,它就会成为“垃圾”,查询时产生信息噪音,影响查询效果。

1.2 图书情报部门的服务因素

面对因特网这样新的服务环境,图书情报部门应如何变革,以适应社会发展的要求,这一问题已经有许多有识之士作了精辟论述。结论是,利用图书情报部门的人才、资源优势,当好桥梁,作好中介,继续向读者或用户开展信息服务。但问题是,该怎么做,现在又做了什么?现在,我国许多图书情报部门没有自己的 Web 站点。已建站点的一些单位,Web 服务器上既无信息资源,也没有信息服务项

- 3 北京图书馆. 中国机读目录格式使用手册. 北京:华艺出版社,1995
- 4 中国科学院文献情报中心. 文献机读目录数据处理手册. 北京:中央广播电视大学出版社,1999
- 5 熊光莹等. 计算机编目技术手册. 北京:北京图书馆出版社,1999
- 6 北京图书馆图书采选编目部. 中文普通图书编目手册.

- 1998
- 7 全国图书馆联合编目中心,国家图书馆图书采选编目部. 中文图书机读目录格式使用手册. 北京:华艺出版社,2000
- 陈源蒸 研究馆员. 通信地址:北京市东城区东厂北巷四号楼群 553 室. 邮编 100006.

(来稿时间:2001-01-12)

目。仅有少量的站点,不仅建设和组织大量的信息资源,而且还开展一些服务项目。例如:清华大学图书馆(<http://www.lib.tsinghua.edu.cn>),开展大量的中外文数据库检索、馆藏书目查询、资源共享等服务;国家图书馆(<http://www.nlc.gov.cn>),不仅开展大量的中外文数据库检索,还开展电子图书查询浏览等服务。

我国一些图书情报站点的信息资源主要有三大类:一是馆藏书目信息;二是中外文数据库;三是开展导航服务的各类“网址集”。服务项目主要是信息检索、资源导航服务和辅助性的用户信息服务(如图书馆新闻、读者意见箱等)。服务方式还基本上停留在被动的用户查询阶段,缺乏主动的信息服务。值得注意的是,现在有许多 ICP,主动进行信息资源的数字化生产,主动在因特网上开展丰富的信息服务活动。例如:中国期刊网(<http://www.cnki.net>)建立了“网上研究院”,不仅在因特网上提供中文期刊、报纸、学位论文等数据库的全文信息检索,而且能将作者、出版社等联系在一起。北京书生科技有限公司(<http://www.shusheng.net>),在因特网上建立了一个集数据库应用、信息资源电子商务和资源数字化加工三位一体的数字图书馆。它不仅提供全息化的电子图书查询浏览服务,而且还可以根据用户需要,提供个性化的定制服务。

目前我国图书情报部门的站点并未成为人们网上信息服务的中介,并未成为人们查询、利用信息必需的“门户”站点。这一点,从我国图书情报部门的 Web 站点建设情况就可以看出(Web 站点是利用因特网开展信息服务最基本的物质条件之一)。

1.3 影响网上信息利用的主观因素

人的主观因素也是决定网上信息利用能力的重要方面。归纳起来,主要是使用者的个人知识背景(如计算机、外语水平等)、信息意识和对网上信息的获取能力等。

2 信息资源的优化、评价和选择

信息资源是开展信息服务的物质基础。信息资源的收集是从事信息服务的首要环节。信息服务效果的好坏,与信息资源的数量和质量有很大关系。国内图书馆的 Web 站点与国外图书馆的 Web 站点在服务内容方面明显的差距主要表现在数字化信息资源太少,馆藏数字化的程度太低^[1]。而数字化信息资源数量是制约图书馆 Web 站点服务的一个主要因素。在购买和自建能力较低的情况下,对因特网上信息资源的优化和选择,也是提高信息服务效果的一个有效途径。

2.1 对网上信息资源的描述

通常在信息服务手段和方式不变的条件下,信息自身的质量是影响信息服务效果的最重要的因素之一。在网络环境下,利用搜索引擎在因特网上查询信息是人们获取信息最常用的方法。但由于 HTML 或 XML 文档缺少必要的

描述信息——元数据(metadata)。这样由搜索引擎网站机器人遍历所获得的信息(有关主页内容的描述)往往包含许多未知内容。在建立索引文档——指引库时,就会产生分类不当或错误,用户查询时就成为信息噪音。为了优化查询结果,图书情报学专家提出,借鉴传统图书馆对书目信息进行规范标引(MARC)的办法对网上资源进行规范标引。为此出现了 DC, PICS, Web Collections, CDF 等元数据格式。其中最有影响的是都柏林元数据(Dublin Core)。DC 是旨在促进信息资源发现的一个元数据集,能较好地解决网络资源的发现、控制和管理问题,并对现在的数字图书馆研究也有很大的意义^[2]。但它自身的规范仍是需要研究的课题。

Dublin Core 资料(data)之有用,必须依赖其资料单元是否有标准,如 Title 之权威控制及相关规则如何^[3]。一旦所创作的 HTML 或 XML 文档中都包含有 DC 描述,网络机器人将能够生成高质量的标引数据,将会提高搜索引擎的检准率。

2.2 对网上资源的评价和选择

在科学文献的搜集工作中,图书情报人员可以用直接评定法、间接评定法评定文献信息源的价值。还可以用布拉德福定律、文献相互引证规律来确定和评选各学科领域的核心期刊,检测信息搜集的完整程度。而网络信息资源的评价,还是一个急待研究的课题。目前已有许多图书馆依据一定的标准,评价、选择一些网上资源,建设了学科专业导引系统,有效地减少了广大科研人员上网查询、选择信息的时间,受到人们的欢迎。如美国密执安大学图书馆和信息与图书馆学校联合编制的《因特网学科资源指南》,按学科分类指引用户利用因特网上的各种资源。

评价资源是筛选出优秀网站的过程。关于网上信息资源的评价标准也是人们普遍关注的焦点,也出现了许多评价方法和标准。在统一科学的评价标准之前,我们可以依据一些原则进行定性评价,选择适合自己需要的信息资源也是非常可行的。如文献 4 所述的 5 个原则:(1)网站被访问次数和频率;(2)建站机构的权威性和知名度;(3)网站传统媒体的重要程度;(4)用户的评价;(5)专家推荐。

3 因特网上信息资源有效组织方法

为了提高因特网上信息的可利用性,人们从信息的有效组织和管理角度做了许多有益的尝试。因特网上信息资源的组织方式,从技术角度可以分为因特网检索工具、虚拟图书馆、书目控制及联机分析挖掘系统(OLAM,又称 OLAP mining)。从信息服务角度考虑,可以分为网络导航、专业学科指引系统、指引库和虚拟图书馆建设等。它们有各自的优缺点,可以简单概括如表 1。

表 1 因特网信息资源组织方式比较

	网络 导航	学科专业指 (导)引系统	指引库	虚拟图书馆
分类	有	有	有	有
标引	无	无	有	有
检索软件	无	无	有	有
组织难易	易	易	较难	难
查全率	差	一般	较高	较高
查准率	一般	高	较高	较高
可用性	一般	好	较好	好

3.1 专业指引系统建设

学科专业指引系统建设就是图书情报工作人员利用各种搜索引擎从网上查询、选择获取一些本专业学科所需的网址,然后经过专业工作人员简单加工(如翻译标题、编写摘要等)后,再按照某种体系结构(不一定是科图分类法、中图法等),将这些物理上分散的大量原始信息资源进行合理整理和组织,从逻辑上联系起来,并且通过各种导航手段,为读者方便地定位、迅速获取所需信息资源提供引导^[5]。如清华大学图书馆的学科网络资源导引(<http://166.111.120.8>)。

学科专业指引系统具有建设简单,见效快,易利用等特点,深受我国一些图书馆的喜欢。现在我国的一些大学图书馆正在建设学科专业指引系统。但本质上学科专业指引系统也只是一种简单的导航,只是学科专业针对性更强。其最根本的问题是没有规范标引和不能检索。这样一旦信息量过大,会给用户带来查询困难。从长远看,其发展必将受到很大的限制。但对于一些学科专业面很窄的小图书馆,也不失为一种有效的选择。

3.2 指引库建设

指引库主要有 3 部分组成:一是反映信息资源的 URL (Uniform Resource Locator, 统一资源定位器)即信息资源的网址;二是原始信息,包括访问频度较高的原始信息资源的镜像等;三是一套方便信息组织和读者查询的支持技术(检索软件)。为了搜集专业信息资源的 URL,所需的关键技术之一是因特网自动漫游技术。从国外的经验看,人们搜集因特网信息资源的手段是利用一种“网络机器人”的软件(automated robot, Wanderer, Spider, Harvest 等)在因特网上漫游,自动搜集网上 WWW, Gopher 和 ftp 站点上的各种相关信息(连接,文本等),并且把这些存于本站点上,然后进行加工、组织和提供服务^[6]。

专业信息资源指引库的建设,我国一些单位也做了开发研究。如南京大学软件新技术国家重点实验室研制开发的“基于 WWW 的信息搜集系统 IDGS (Information Discovering and Gathering System)”。IDGS 是为了在 WWW 上自动进行中英文技术资料的搜集而设计开发的。IDGS 能够根据用户所提供的训练文本和系统设定,提取目标资料的特征,并根据目标特征自动在 WWW 上搜集资料,然后将所搜集到的资料经分类整理导入文档索引库,供用户查询。IDGS 系统能够保持资料库的不断更新,能够向用户提供及时的网上技术资料服务^[7]。

3.3 虚拟图书馆建设

虚拟图书馆是指将分布在因特网上相关网页的 URL 收集起来,并对其进行标引,形成倒排档。倒排档中每条记录的文献标识(一种元数据元素)均指向相应网页的 URL。检索时,用户输入检索词,虚拟图书馆的检索软件将输入的检索词与相应倒排档中的文献标识进行匹配,并将匹配结果的 URL 集合显示在荧光屏上,用户通过超链可以调出相应的 URL 的网页^[8]。

虚拟图书馆建设通常是很有针对性的,通常是某信息服务机构针对自己的服务群体而建设的。虚拟图书馆不是简单地进行网页的搜集分类,而是经过专业人员,根据某些原则和标准对因特网上信息资源反复筛选、加工后才提供读者使用的。从对信息的组织的角度看,它较好地解决了基于特定主题的族性检索问题^[9]。虚拟图书馆在因特网上信息资源的组织方面的确有许多优点。但因为专业人员手工加工难以应付因特网上信息的快速增长,因而还必须加强虚拟图书馆的智能化研究。

用户对查询的结果是否满意,就存在一个评价标准问题。因特网是一个高度开放的海量信息服务系统。若我们严格用检全率和检准率来评价它的性能,显然有些不太恰当。也许我们用用户能够接受的相对检全率和检准率来评价更科学些。这样,尽管目前学科专业指引系统和虚拟图书馆都很难做得较为系统(主要是人工查询、筛选和数据加工,数据量较少),查全率相对较低,但较高的查准率和较为详细的资源内容说明,能够较好地满足读者的基本需求,因而这些方法较为可行。不过在对因特网资源的组织方面,国外一般采取主题和字顺两种方式,主题分类比较正规详细,而国内图书馆的 Web 站点在资源组织方面集成度较低,并且主题分类模糊不正规,缺乏统一的字顺标准^[10]。

4 推送技术对网络信息服务方式的变革

4.1 一种新的网络信息服务技术——推送技术 (Push technology)

为解决因特网上寻找信息费时费力的缺陷,现在因特网上正流行着一种新的服务方式——基于 Web 的信息推送技术。美国、英国等发达国家的图书情报部门,正在研究利用推送技术开展主动信息服务的新方式,近两年就发表相关论文 160 多篇,例如文献[11~14]。

“推送”技术,又称 Web 广播技术、频道技术,实质上是一种软件。从技术上看,“推”模式网络信息服务是具有一定智能性的,可以将其描述为,基于 Internet/ Intranet 网络环境的一个高度专业化、智能化的网络专题信息服务系统。该软件不仅能够了解、发现用户的兴趣,还能够主动从网上搜寻信息,经过筛选、分类、排序,按照每个用户的特定要求,主动推送给用户。美国华盛顿大学图书馆、加州洛杉矶分校图书馆等利用因特网,为用户提供“一对一”(One-to-One)的服务,以满足用户的个性化信息需求^[15]。这就是基于“Push”技术的一种新的信息服务模式。

我国的图书情报机构,目前仍然借用传统的信息服务模式,单纯提供信息检索(或查询)、网络导航等被动服务项目。可以说,利用推送技术建立主动信息服务 Web 站点,不仅可以节省读者的宝贵时间,而且可以克服读者对计算

机网络、外语等知识的不足,极大提高网络信息资源的利用效果。我们应当加快对这些新的信息技术的研究与开发。

4.2 专业信息服务频道的建设

频道是一个定期更新、定期通知的 Web 站点。由于采用了推送技术,使得用户不必每次来访问固定的站点就可以自动获得由网站自动发送的最新资源,它还提供了拨号用户离线浏览的功能。只要用户预先订阅所需的频道,并设置好更新时间,就可以获得由浏览器自动下载、更新的资料。对于一个图书馆的 Web 站点,只要建设一个专业信息服务频道,就能够面向自己的用户开展具有很强针对性的主动信息推送服务。一个图书馆的专业信息服务频道是这样工作的:

(1) 用户登录到图书馆 Web 站点,提出获取主动推送信息服务申请(仅第一次获取服务时需要填写)。

(2) 图书馆 web server 发送一个申请表单给用户,具体项目包括:用户名、密码、所需信息的主题、关键词、推送信息的地址、推送周期、要查询的数据库等等。

(3) 用户填写好申请表后,提交给图书馆 web server。图书馆 web server 将用户的特征信息、查询要求等传送给“推送服务代理”。

(4) “推送服务代理”根据用户的请求信息,在用户特征信息库(专门记录用户需求特征信息的数据库)和用户信息库(专门记录用户希望获取信息的数据库)中分别增加一条记录。

(5) “推送服务代理”会根据用户的要求,定期将用户的查询需求传递给“查询代理”。

(6) “查询代理”根据“推送服务代理”传送的用户要求,定期检索相应的数据库,并将查询结果返回“推送服务代理”。

(7) “推送服务代理”按照用户的要求,定期将最新信息推送到用户指定的地址。

从系统论的角度看,基于 Web 站点的信息服务效果的优化是一个系统工程。通过对信息的收集、整理、加工、处理、存储、传递、利用等环节的控制,使信息系统在整体上处于最优状态,都能够改善和优化服务效果。以上我们只是从信息源的控制与选择、信息资源的组织与管理和信息服务方式等 3 个方面论述了优化网上信息服务效果的一些技术。实际上,每一方面都存在许多急需研究的相关问题。例如,网上信息主动服务模式 Push 技术。尽管国内外一些

计算机和图书情报专家围绕着这些方面做了许多开创性的研究,但网上信息资源的挖掘、过滤、筛选、智能搜索、智能代理以及 Push 服务器等,都是有待进一步研究重要课题。

参考文献

- 1,10 张福学,时永梅.图书馆 Web 站点服务的实践研究.图书情报工作,2000(9):35~39
- 2 肖珑.元数据格式在数字图书馆中的应用.大学图书馆学报,1999(4):18~24
- 3 卢秀菊.英美编目规则原则之探讨.图书与资讯学刊,2000(2):16~39
- 4 沈红芳.论因特网资源选择及相关问题.情报理论与实践,2000(2):146~148
- 5 索传军.论网络环境下图书馆虚拟馆藏的建设.中国图书馆学报,2000(1)
- 6 尹俊涛,徐如镜.全球信息网络环境下信息资源开发利用技术初探.情报学报,1997(2):100~103
- 7 邹涛.基于 WWW 的信息搜集系统 IDGS 的实现.计算机世界产品与技术专辑,1999-04-19(C10)
- 8 陈光祚.论“图书情报学虚拟图书馆”的建设.中国图书馆学报,2000(1)
- 9 晏凌,王应解. Internet 信息资源的组织策略——建立基于联机分析挖掘的导航系统.图书情报工作,2000(5):46~49
- 11 Styczynski J. Make a Web site sing with free push technology. Information Outlook. 37 Jul 1999
- 12 Pedley P. Intranets and push technology: creating an information environment. Managing Information. 6(6) Jul/ Aug 1999
- 13 Tso S K;Laur H C W; Ip R W L. Development of a fuzzy push delivery scheme for Internet sites. Expert-Systems. 16(2) May 1999
- 14 Solomon M. When Push comes to Pull: serving current awareness applications in your company's news cafeteria. Searcher. 7(6) Jun 1999
- 15 A·贺福纳等.网络环境中学术图书馆一对一服务模式策略之探讨.图书情报工作,1999(12)

注 本文是 2000 年国家自然科学基金项目:70073022 研究成果之一。

索传军 武汉大学信息管理学院在职博士生,郑州大学北校区图书馆副馆长、副研究馆员。通讯地址:武汉。邮编 430072。

焦玉英 武汉大学信息管理学院教授,博士生导师。

(来稿时间:2000-12-25)

关于召开计算机数字图书出版技术研讨会的通知

各有关单位:

随着互联网的快速发展,传统的传媒载体已逐渐由单纯的纸张印刷为电子图书、电子视频、电子声像等技术所取代。如何构筑网络数字图书出版系统和建设数字图书馆是有关主管领导和专业人员极为关切的问题,也是当今计算机业界和图书馆界普遍关注的问题。为了开展这一技术领域的研讨,提高我国出版界、图书馆等行业的计算机应用水平,中国计算机用户协会微机系统分会、中国图书馆学报联

合在哈尔滨市举办计算机数字图书出版技术研讨会,将邀请国内外计算机业界专家做主题演讲。会议事项如下:

(1) 会议日期:2001 年 8 月 15 日—18 日。(2) 会议地点:哈尔滨市道里区友谊路 276 号石化大厦;(3) 联系人:张鸿飞,徐苇。电话:010—68573311-86309,010—68415566-5141。传真:010—66122820。(4) 联系地址:北京市西城区月坛南街 75 号国家统计局计算中心。邮政编码:100826。电子信箱:Zhanghf@hotmail.com