

虚拟图书馆信息搜集、组织方式及研究方向^{*}

徐 晓 万跃华

(浙江省社会科学院) (浙江工业大学)

摘 要 讨论了数字图书馆中虚拟图书馆信息搜集与组织方法;介绍了浙江省社会科学院人文、社会科学虚拟图书馆的设计与构建,指出了目前虚拟图书馆研究中存在的主要问题和研究方向。

关键词 虚拟图书馆 信息搜集 网络信息组织 研究方向

在数字图书馆中,信息将以多媒体形式表达,读者将从文本、图形、图像、声音、视频等多种媒体所表达的信息中获取知识。虚拟图书馆是根据特定的目标、特定的用户群,选定信息资源的学科领域,针对网上已经存在的信息资源(包括文本、图形、图像、声音、视频等多种媒体)进行搜索和收集,加以鉴定核实,并对核实后的网址进行合理组织、筛选、分类、标引、注解信息资源,使之能够提供检索、浏览和链接的信息集合,使用户随时选择调用,主要作用是实现网上信息资源的共享。构建虚拟图书馆目前已经成为图书情报领域主要热点问题,许多研究者对此倾注了极大的研究热情和兴趣。

虚拟图书馆的主要功能是在庞大的因特网信息群中搜索某一学科、某一领域的相关信息,然后将这些信息加以组织,供用户浏览或者查询。对于虚拟图书馆管理人员而言,其完成的工作可以概括为两点:信息搜集和信息组织。

1 虚拟图书馆信息搜集方式

1.1 人工搜集 人工搜集通常有两种渠道:正式渠道和非正式渠道。前者包括查阅大量的书目、期刊、报纸以及其它有关的工具书,浏览因特网,利用搜索引擎采用人工检索有用的网络信息资源;后者则包括与他人口头交流、书信来往等等。通过正式或者非正式渠道,我们可以获得如下信息:选定学科领域有关研究机构、实验室的 URL,电子书籍的 URL,学术期刊的 URL,学术活动(如会议、展览、讲座等)的 URLs,网上学术论坛(讨论组)的 URL,有关专家学者的 URLs 以及其他相关 URL。这种搜集方式的优点在于链接站点经人工筛选,相关度较高。其不利之处是人工搜集效率较低,难以较完整地搜集到相关的 URL,因而虚拟图书馆的功能在很大程度上依赖于其策划者所投入的精力。但由于人的精力有限,因而很难彻底地搜集到相关信息,从而影响到虚拟图书馆搜集资料的全面性。另外,虚拟图书馆的管理人员还需周期性地核实原有链接,及时排除“死链”。

1.2 自动化搜集 这种方式是指因特网上相关信息的搜集工作由专用的搜索软件完成,如通常采用一种称作 robot(或 spider、crawler 等)的软件技术自动跟踪浏览 Web 信息,然后将浏览结果进行加工、处理,形成一个庞大的数据库,每次跟踪浏览的结果可自动追加进数据库,此数据库可由网上任一站点通过输入提问的

方式进行访问。其工作原理通常是从一个或一组 URL 开始,访问该 URL 并进行本地索引,同时记录该 URL 所指 HTML 文件中所有新的 URL 锚链(anchor);然后再以这些新的 URL 为起点,继续进行本地索引,直到没有满足条件的 URL 为止。在记录新的 URL 时,可以进行分析和判断,去掉不需要或不想要的 URL。这种搜索方式与人工搜集方式相比,节省了大量时间,因其面向整个因特网进行搜索,因而具有较强的全面性。但这种搜集方式含有很高的人工智能成分,对软件的编写提出了较高的要求。

实际上,在搜索选定学科或领域的相关信息的过程中,往往是这两种方式被同时采用。

2 虚拟图书馆信息组织

信息组织是指采用一定的方式,将某一方面的大量的、分散的、杂乱的信息经过整序、优化,形成一个便于有效利用的系统的过程,其目的是将无序信息组织成有序信息,便于人们利用信息和有效地传递信息。相对于传统图书馆的文献信息而言,网络信息资源的组织整理也有不同的方法和手段。

2.1 主题树方式 所谓主题树组织方式,就是将虚拟图书馆所含这一学科的所有已获得的信息资源按照某种事先确定的概念体系结构,分门别类地逐层加以组织,用户通过浏览的方式逐层加以选择,层层遍历,直到找到所需要的信息线索(即相关站点链接),继而通过信息线索直接找到相应的网络信息资源。

这种组织方式的优点是简单易用,屏蔽了网络信息资源系统对于用户的复杂性,提供了一种基于树型浏览方式浏览网络信息的界面;信息检索按照一定的体系结构,逐次查看,因而对于用户而言,目的性强,查准率高。其缺点是体系的结构不能过于复杂,每一类目下的信息索引条目也不宜过多,这无疑限制了其所能容纳的网络信息资源的数量。因而,这种组织方式对于小型的虚拟图书馆而言是可行的,但倘若虚拟图书馆包括大量复杂的某一领域的网络信息资源,这时就需依靠数据库组织方式。

2.2 数据库组织方式 所谓数据库组织方式,就是将虚拟图书馆所包含的这一学科的所有已获得网络信息资源以固定的记录格式存储,并提供一些检索入口,用户通过关键词及其主配查询,就可以找到所需要的信息线索(即相关站点链接),继而通过信息线索直接连接到相应的网络信息资源。

这种组织方式利用数据模型对信息进行规范化处理,利用关系代数理论进行数据查询的优化,从而大大提高了数据操作的灵活性,因而成为广泛的网络信息资源组织方式。它对用户也提出了较高的要求,要求用户掌握一定的检索技巧,包括关键词及其组配方法的选择等。同时,在数据库服务端,对于如何进行数据库的自动扩充,如何提供良好的人机交互,也提出了较高的要求。

3 浙江省社科院人文、社会科学虚拟图书馆的设计与构建

作为中文专业性的人文、社会科学虚拟图书馆国内目前还很少做过,美国加州大学河滨分校(University of California at Riverside)图书馆建设了一个专门为教师、学生和研究人员服务的虚拟图书馆 Imfomine: Social Sciences & Humanities(<http://lib.www.ucr.edu/>),包括资源网站、数据库、电子期刊、电子书、电子公告板、邮件服务器、联机图书馆目录、研究文献和研究人员指南等。美国 University of Evansville 开发的历史联盟资源网站 Argos(<http://argos.evansville.edu/>)主要收集了经过选择的包含大量关于远古和中世纪历史研究方面的信息资源。

对于浙江省而言,目前已经形成对信息资源的迫切急需,图书馆如何为浙江的经济建设提供信息服务保障体系将成为未来的工作重点。浙江省人文、社会科学研究资源非常丰富,如何使浙江省的人文、社会科学研究走向世界,发展电子化、数字化的虚拟馆藏信息资源,提供远端的信息资源检索服务,使用户不受时空、馆藏限制,信息资源以超媒体(hypermedia)展现,是我们必将解决的问题。浙江省社会科学院于 1999 年开展“网络环境下人文、社会科学虚拟图书馆研究”课题研究,并受到院重点基金项目资助。其虚拟图书馆建设研究的主要目的是方便浙江省社会科学研究人员利用因特网上丰富的信息资源开展科学研究,以使浙江省社会科学研究实现在世界范围内的信息资源共享,这对于充分开发和利用人文、社会科学信息资源,让浙江省的人文、社会科学走向世界具有十分重要的学术意义和经济意义。对于经费拮据、出版物价格上涨而本身馆藏信息资源又不丰富的社会科学院图书馆而言,选择搜集网络信息资源来补充其本身比较贫乏的馆藏信息资源显得尤为重要。

该虚拟图书馆是根据浙江省社会科学院的学科建设情况组织的一个网络信息资源库,是为了方便广大研究人员进行科学研究而设立的。课题组利用网络检索工具——各类搜索引擎,通过对网上信息的研究、选择、筛选与组织,并对所搜集的网络信息资源的内容进行介绍,采用超文本技术建立的人文、社会科学虚拟图书馆,可将用户直接“引入”所要查寻的专业和主题,使用户的检索需求得以实现。该虚拟图书馆搜集了因特网上与本院有关的(如哲学、法律、经济学、历史、海外中国学、心理学、妇女研究等)学科,构建的虚拟图书馆包括该学科的组织机构(大学、研究机构)、学会、协会、电子期刊、国际会议、数据库、参考工具书、投稿指南、著名学者、计算机软件等。在近一年的时间里,我们从网上搜集了近 4 000 多个有关人文、社会科学学术网站的网址,通过该虚拟图书馆可以十分便捷地了解国际学术动态。

我们在虚拟图书馆中设置了各种研究方向,各位来访者可以方便而迅速地查阅许多专业方向的重要信息。

4 虚拟图书馆的研究方向

4.1 研究制定馆藏信息资源发展政策 图书馆组织搜集、整

理、分类、标引、注解网络信息资源的范围应根据各馆的馆藏信息资源发展政策确定,馆藏发展政策是规范馆藏发展和具体运作的纲领性文件。我们目前所言的馆藏发展政策主要是针对传统的印刷型文献信息资源,绝大多数图书馆都还没有制定有关网络信息资源的馆藏发展政策。要系统、高效地对网络信息资源进行搜集、整理、分类、标引、注解,必须依据网络信息资源的馆藏发展政策所确定的原则、标准来选择网络信息资源。

4.2 网络信息资源的编目 网络信息资源的组织管理是一项相当复杂艰巨的工作,而解决复杂问题的最简单有效的办法便是传统方法的延伸和扩充。图书馆编目技术用于网络信息资源组织即是其中一例。

美国国会图书馆(LC)、美国图书馆学会(ALA)、OCLC 和一些大学图书馆都提出了网络信息资源组织整理计划,这些计划倾向于用传统的机读目录格式来组织整理网络信息资源,最典型的例子是 OCLC 因特网编目计划 InterCat。尽管这种编目方式对于数量庞大的网络信息资源不仅成本昂贵,而且根本赶不上网络信息资源的发展速度,但其意义是不可低估的。

网络信息资源编目方面取得的重大进展是 USMARC 一体化编目格式的正式实行。图书馆界用 USMARC 格式进行的网络资源编目实际上是属于受控编目,编目数据质量高,但编目速度慢、成本高,对数量如此庞大的网络信息资源深感力不从心。在信息技术界,许多研究者设计了多种元数据(metadata)格式来描述电子信息资源。元数据就是数据的数据,元数据与网络信息资源之间的关系如同书目数据与图书之间的关系一样。Dublin Core 就是一种典型的描述网络信息资源的元数据格式,它有 15 个元素。图书馆界正在密切关注元数据记录能否用于图书馆编目记录,以及这些元数据能否帮助图书馆实现网络信息资源的高效编目。尽管元数据和书目数据在结构、数据表示、数据元素和使用规则等方面差异明显,但元数据在书目数据上的应用从元素和形式上都是可以接受的,转换语法也是可以实现的,但需要对元数据加以语义和内容控制,才能很好地适应书目数据的要求。已有研究者将 Dublin Core 的 15 个元素与 UNIMARC 格式的元素做了映射。虚拟图书馆的信息资源以元数据编目是今后的研究方向。

4.3 虚拟图书馆链接的电子文献以元数据编目加入 OPAC 的问题 目前网络信息技术的一个发展方向是将各种传统和非传统的信息整合起来,构成一个不断完善、不断充实的信息体系,使用统一的界面为读者提供服务。对虚拟图书馆建设来说,可以进一步考虑把虚拟图书馆站点链接的电子资源以元数据编目集成到图书馆原有的基于 MARC 的 OPAC 系统中,通过 OPAC 揭示虚拟图书馆的信息资源。

4.4 网络目录 分类法是图书馆界在文献资源组织整理方面使用最早、应用最广的方法,而今它已被广泛应用于网络信息资源的组织,一些著名的搜索引擎一般都向用户提供关键词检索和分类浏览两种信息查询途径。搜索引擎中使用图书分类法只是方法移植。图书馆使用图书分类法建立网络目录(Webdirectory)是图书馆进行网络信息资源组织管理的积极措施。图书馆所使用的一些主要分类法几乎都已应用于网络信息资源的分类组织,如 DDC(杜威十进分类法)、UDC(国际十进分类法)、LCC(国会图书馆分类法)和中国法等。采用图书分类法组织网络信息资源的成功实例为数不少,虚拟图书馆应该采用分类法来组织目录。

4.5 网上自动跟踪和自动漫游技术将逐步完善 (下转第 20 页)

3.2 人际网络在竞争情报收集中的作用 人际网络在企业竞争情报系统中的作用最突出的还是在情报收集阶段,日本和美国在竞争情报利用方面可以说是卓有成就,日本以其在竞争情报的收集方面举世闻名,美国的竞争情报代表了世界上竞争情报理论和实践的最高水平。美国的 FORTUNE 杂志上曾刊登过收集竞争情报方法以及日本竞争情报收集的方法,将其归纳如下:从潜在的企业成员中收集;从参加各种会议中搜集;从竞争对手中挖走关键人物收集;通过咨询人员间接访问竞争对手;通过设计顾问收集;询问竞争对手的前职员;与竞争对手的基本客户交谈;通过物料供应商侧面了解;向商业经营管理部门渗透收集;分析报刊上的征求广告;分析劳动雇佣合同;研究空中摄影照片;支持信息部门查阅文件和档案,从中收集竞争对手有关新产品情报;查阅对手商业贷款记录;衡量专用运输线上铁轨的锈渍程度;以假设的身份参观竞争对手的工厂;反求工程,分析对手产品进行工艺还原;购买对手的工业垃圾。

以上是美国、日本搜集竞争情报的方法,真可以说是“天网恢恢,疏而不漏”(尽管有些方法属于违规的竞争情报收集方法)。如果我们与前面的人际网络的构成图对照一下,就可以知道以上的方法大部分都是利用了人际网络来收集情报的。

有时候一些不经意的发现也会成为企业的重要情报,例如,美国一位加油站工作人员发现一辆满载油管的火车向南方驶去,当即把这个信息反馈给情报人员,情报人员经调查发现竞争对手正在一个新的区域铺设天然气管道,从而掌握了竞争对手的动向。这是一个充分利用了企业内部人际网络的例子,如果企业的每一个员工都有这种意识,那么企业的情报收集将会更全面。有效促进员工对于企业内部有价值信息的交流和挖掘,能使整个企业保持一种对于外界环境的敏感,不断地从外界吸取可能对提高

企业竞争力有重要影响的消息,并将其转化为企业内部的东西。

3.3 人际网络在竞争情报分析阶段的作用 竞争情报分析阶段主要是在企业内部进行的,首先要获得高层领导的大力支持以及其他部门的配合,然后是企业的竞争情报人员对所获情报进行分析。在这个阶段,人际网络也同样重要。

在分析阶段,最重要的还是内部分析人员的互相配合、互相协调。竞争情报分析人员是一些很有思想、很有主见的人,如果他们各自固守自己的观点,而不注意与其他成员相互配合,那么再好的分析人员也不会很好地完成任务。只有全体成员齐心协力,充分发挥团队精神,先各抒己见,再取长补短,才有可能使他们出色地完成分析任务。

在分析阶段也不可忽视高层领导的支持以及其他部门的配合,如果竞争情报部门在以上两方面能够做得很好,那么就会使分析阶段的一些内部问题尽快解决。

4 结束语

随着社会的发展,企业的竞争将越来越激烈,但是这些企业越来越意识到必须采取竞争策略才能更好地生存和发展,因此人际网络也将会在企业竞争情报系统中发挥着越来越重要的作用。

参考文献

- 1 包昌火,谢新洲等. 企业竞争情报系统. 北京:华夏出版社,2001
- 2 陈峰,梁战平. 信息化概念及其对竞争情报工作的指导意义. 情报学报, 2003;(3)
- 3 包昌火,谢新洲等. 人际网络分析. 情报学报. 2003;(3)
- 4 企业竞争情报系统中的知识管理思想. <http://www.chinalibs.net/>
- 5 学问好不如人面好, <http://msyw.myrice.com/shjw-1.htm>

(责编:愚王京)

(上接第17页)利用和开发网络搜索软件自动搜索指定站点的网页和多媒体数据,可以从主页和任一页面开始搜索,遍历所有网页。自动跟踪网页资源变化,自动删除重复网页和无效连接,通过分析页与页的连接顺序查找新地址,做到实时更新。

4.6 对网络信息的组织采取更先进的技术手段 利用和开发网上自动分类标引、自动文摘技术,拷贝各种网站有关资源和内容尤其是大型或专业门户导航网站相应学科分类目录下的站点条目的内容,然后组织建立索引文档,形成数据库,并有相应的搜索引擎对索引数据库进行检索。利用和开发定量分析软件,自动跟踪、统计用户访问的学科站点情况。各站点被访问次数、统计各站点有关内容被下载次数,从而做出客观、公正的评价。自动确定用户常用的学科站点和用户经常访问的信息,然后在虚拟图书馆中作为推荐站点并生成相关信息。

4.7 利用数据挖掘组织虚拟图书馆 数据挖掘也称为数据库中的知识发现,是近几年来随着数据库和人工智能发展起来的一门新兴的数据库技术,数据挖掘是按照既定的业务目标从海量数据中提取出潜在、事先未知的而又可能极为有效并能被人理解的高级处理过程,帮助人们从庞大的目标数据集中智能地、自动地抽取一些有价值的知识模式,以满足人们不同的应用需要。对于因特网上超巨量的数据,自然让人们想到应用数据挖掘技术来提高在因特网上检索信息、利用信息的效率。如何从浩如烟海的

数据中发现隐藏的有用知识,创造更大的效益,是虚拟图书馆迫切需要研究的课题。

5 结束语

需要指出的是,虚拟图书馆的建设是一项复杂的系统工程,它的发展与完善有赖于现代信息技术本身的发展以及现代信息技术在信息管理领域的应用水平。同时,人们对因特网环境下虚拟图书馆这一信息组织形式的理解与认识,以及知识产权保护状况等人文因素也会对虚拟图书馆的建设和发展产生直接或间接的影响。另外数字化信息的学科专业分布,特别是馆藏文献的数字化程度也必将影响虚拟图书馆利用的深度和广度。

参考文献

- 1 Jan T. Orick. The Virtual Library: Changing Roles and Ethical Challenges for Librarians. The International Information & Library Review, 2000;(3-4)
- 2 Bob Travica. Organizational Aspects of the Virtual Library: A Survey of Academic Libraries. Library & Information Science Research, 1999;(2)
- 3 万跃华,徐晓. 构想社会科学虚拟图书馆. 观察与思考,2001;(11)
- 4 王桂平. 创新重点学科虚拟图书馆的建设. 中国图书馆学报,2002;(4)
- 5 臧国全. 虚拟图书馆中网页的自动分类研究. 现代图书情报技术,2002;(3)
- 6 刘兹桓,张久珍. 国内外虚拟图书馆研究综述. 中国图书馆学报,2000;(3)

(责编:愚加勃)