●相丽玲 曹 平

互联网上灰色信息的挖掘与利用

摘 要 开发和利用互联网上的灰色信息资源,在企业参与市场竞争的活动中占有非常重要的地位。挖掘网上灰色信息资源的主要方式有:利用各种类型的搜索引擎,建立灰色文献虚拟数据库,使用专门的信息收集系统,开发数据信息挖掘技术等。参考文献4。

关键词 因特网 灰色信息 信息挖掘 信息利用
分类号 G253

ABSTRACT The development and utilization of gray information resources on the Internet play a very important role in an enterprise's participation in market competition. Major methods for the mining of gray information resources on the Internet include using various search engines, creating virtual databases of gray information, using special information harvesting systems, developing information mining technologies, etc. 4 refs.

KEY WORDS Internet. Gray information. Information mining. Information utilization.
CLASS NIMBER G253

1 互联网上灰色信息的含义及特征

互联网上灰色信息是指在互联网上存在的,非常规发行、并且允许用户免费或在一定范围内收集、整理和利用的信息资源。其涵盖面非常广泛,包括网站的商业广告、会议文献、个人网页等[1],同传统意义上的灰色信息一样,互联网上的灰色信息也是国内外图书情报界公认的重要情报源。

互联网上的灰色信息资源具有以下明显特征:

(1)范围的模糊性。灰色文献是一种过渡性质的文献。一般来说,网上公开发行的电子期刊、电子书籍,不管是免费使用还是有偿使用,均属于白色文献。在网上发布的各类没有公开发行版权的电子信息资料,归于网上灰色文献信息。由于网上灰色文献信息范围越来越广,且与其他文献信息的分界线越来越模糊,因此,网上灰色信息资源范围更难确定。

(2)发布的高自由度。从文献信息控制的角度看,灰色文献信息是处于受控边缘的文献形式。信息社会到来会加剧信息的无序状态,灰色文献的自由性和失控性,使其在互联网上的发布具有更高的自由度,即使采用有效措施,也不可能回到传统信息文献的受控状态。

(3)数量的无限增长趋势。互联网上的信息资源数量极大,作为网上信息资源组成部分的灰色文献信息,涉及人类生活的各个方面,深入到经济、政治、文化、科技、军事等各个领域,网上灰色信息正朝着无限量的方向发展。

(4)出版的时效性。各类网站发布的灰色信息具有极强的针对性和实用性。对这些信息的更新,少则一两天,多则一个星期或一个月,与传统的纸质文献相比,时效性更强。

(5)收集的便利性。传统的纸质灰色文献多为内部出版发行,印刷数量有限,报道范围狭窄,加之受保密制度和专业的限制,给灰色文献收集带来了很大困难,而网上灰色文献以光盘、硬盘等介质进行存储,利用互联网可以快速传送,只要供方愿意,使用方即可随意复制使用,不受时空限制。

网络环境中的信息传播只是非网络环境中信息传播功能的一种延伸和发展。它所要实现的基本目的和功能在本质上和非网络环境中的信息传播是一致的,只是实现的环境、手段和数量不同。和非网络环境中的信息传播一样,网络环境中也存在通过使用共同的软件进行会话、交谈、会议、信件往来等纯粹私人的或集团性的传播方式,也存在着由社会集团所控制的和各类社会组织所提供的比较制度化的传播方式。网络灰色信息资源区别于非网络灰色文献的一个重要特点,就是网络灰色信息资源存取和利用的多样性。

2 互联网上灰色信息资源的挖掘与利用

2.1 互联网上灰色信息资源的挖掘与利用渠道

2.1.1 互联网上灰色文献的信息源

(1)单位概况。网上信息发布的基本场所是散布在世界各地的网站,而每一个网站都分属于不同的单位和机构,几乎所有的网站都免不了介绍其管理机构的基本状况,以扩大本单位的知名度和影响力。

(2)动态报道。这类信息的时效性最强,更新速度最快。不同网站对其标识也不尽相同。主要包括网上发布的新闻报道、新闻追踪分析等,比如新浪网对国内外新闻、体育、娱乐、政治等各方面重大时事报道信息非常丰富和全面,而企业网站中的动态报道集中在公告和关于企业最新动态的栏目中。

(3)网站广告。网站广告在网络世界中占有极其重要的位置,它不受空间范围的限制,可以产生世界性广告效应,

并且具有广告效益的可准确计量特征。对于商家来说，网站广告有着报刊电视广告无法比拟的优越性，广告收入也是各商业网站得以生存发展的经济支柱之一。商家的青睐和网站的生存发展需要使网上广告所占幅面不断增大，使广告信息遍布于互联网上。

(4)用户信息。网站与用户的相互交流，依靠网站提供的公共界面来实现。用户要访问网站的资源，根据访问内容的不同会受到不同的制约。比如，新用户要申请免费电子信箱或进入聊天室，一般会被要求进行注册，老用户则被要求输入注册号或密码。通过这种方式，网站可以掌握用户的个人主页、注册、电子邮件和聊天等大量信息，这些信息也为用户之间的相互交流提供了保障[2]。

(5)索引数据库。很多网站在网上发布诸如专题导航之类的索引型信息，用户可以依据索引找到相关资源。索引信息数据库是网上灰色文献信息资源最重要的二级信息源，它通过对信息的再次加工整理，提供最快的检索通道，增强了原始信息源的利用效率。

2.1.2　互联网上灰色信息资源的挖掘与利用方式

信息挖掘和信息收集是不同概念。信息收集是指通过各种方式获取所需要的信息；而信息挖掘指从各种各样的信息源中，抽取先前未知的、完整的信息，来做关键的业务决策。信息挖掘主要利用了数据挖掘技术，从大型数据库的数据中提取人们感兴趣的知识，这些知识是隐含的、事先未知的、潜在的有用信息。信息挖掘是基于信息收集基础之上的。

(1)广泛利用各种类型的搜索引擎，挖掘网上的灰色信息。搜索引擎是针对网上信息爆炸，为解决用户的查询而设计的，主要有两类，即分类目录式和主题检索式。用户利用哪种搜索工具，这取决于所要查询的具体问题。利用搜索引擎收集灰色信息，要注意避免由于搜索引擎本身的技术问题带来的不利影响，比如"关键词"问题，很多搜索引擎都屏蔽一些本身缺乏实际意义或使用过于广泛的所谓的"关键词"。

网络信息挖掘技术在搜索引擎上的应用很多，比如Google搜索的最大特色就体现在它所采用的对网页Links信息挖掘技术上。网络信息挖掘是目前网络信息检索发展的一个关键，如通过对网页内容挖掘，可以实现对网页的聚类、分类，实现网络信息的分类浏览与检索；通过对用户所使用的提问式(query)的历史记录分析，可以有效地进行提问扩展(query expansion)，提高查全率和查准率；可以运用网络内容挖掘技术改进关键词加权算法，提高网络信息的标引准确度，从而改善检索效果。灰色信息在网上过于分散，缺乏特色主题，只有科学地使用各种不同类型的搜索引擎工具，才能有效地开展挖掘工作。

(2)建立灰色文献虚拟数据库。网上灰色文献信息只有经过系统并且有序地处理，才可能得到高效率的利用，而运用虚拟数据库技术，建立虚拟的灰色文献数据库是极为实用的方法。虚拟数据库是将各类型数据转变为以关系数据库为统一界面的系统。在网络数据源中，数据的组织形式、检索词和存取机制各不相同，它不支持统一的查询操作，要求利用虚拟数据库技术，为用户提供友好通用的人机界面。

现在分类技术与虚拟数据库相互结合，就是所谓虚拟数据分类技术，它以优良的检索词组配方式，为信息类型、著作、书名等确立搭配使用的窗口。分类检索和主题检索可以相互转换，并增加自然语言查询方式，从而增强对资源的选择功能与查询功能。利用这一技术构建灰色文献虚拟数据库，可以方便地与相关站点链接，使各个检索系统的协调更加便利。

虽然大多数数据库生产者还没有收集灰色文献信息的手段，但某些数据库生产者(如工程情报公司)已经在致力于灰色文献的收集工作。一些数据库的用户有时也就可能是灰色文献的生产者。例如，AGRIS(国际农业科学技术情报系统)与几个国家的全国中心合作，参与数据库的建设，自己就成为灰色文献的生产者[3]。

特别要引起重视的是某些国际性实验室和研究中心正在进行中的研究项目也被编入数据库，如NTIS(National Technical Information Service，美国国家技术情报数据库)。日本和欧洲也在生产这种数据库。另一种不应当忽视的灰色情报源是新闻单位数据库，这类数据库可用于查找公司企业感兴趣的事务和领域。公司企业出版的关于本公司企业的年度报告和财务报告等，过去往往很难得到，现在这些企业也愿意提供了，他们不仅提供报告目录，而且提供报告的仿真本。

从总体上看，网络上的文献信息资源是动态的，作为其传递系统的虚拟数据库应具有完全开放和动态性特点。为了使用户能方便及时地查找最新的灰色文献信息，必须不断跟踪网上站点及其内容的变化，随时增加新的指导性信息，更改和删除过时的相关信息。

(3)使用专门的信息收集系统。专门的信息收集系统是指使用专门的信息收集软件系统来获取网上潜藏的灰色信息资源。近年来，我国的软件企业也推出了简单易用的信息系统软件产品，如天下互联中国网络情报中心开发的企业情报门户系统软件(CIPS)，已经成为企业情报人员的好帮手。中国网络情报中心的CIPS系统，是要为企业建立个性化信息需求的"企业的情报门户"。它是区别于大众门户网站和行业门户网站的智能互联网门户网站，是企业的门户网站。CIPS系统的最大特色，是作为企业情报门户的功能，它不是简单的竞争情报系统(CIS)，更不仅是企业内部知识管理(KM)，主要研究的是企业的门户(Portal)，CIPS系统是对CIS、KM、Portal的有效整合。

(4)开发数据信息挖掘技术。运用网络数据挖掘技术能够从服务器以及浏览器端日志记录中发现隐藏在数据中的模式信息，了解系统的访问模式以及用户的行为模式，从而作出预测性分析。例如通过评价用户对某一信息资源浏览所花的时间，可以判断出用户对资源兴趣如何；对日志文件所收集到的域名数据，根据国家或类型(.com,.edu,.gov

等)进行分类分析;应用聚类分析来识别用户的访问动机和访问趋势等,这项技术已经有效地运用在电子商务中。通过对网站内容的挖掘,主要是对文本内容的挖掘,可以有效地组织网站信息,例如采用自动归类技术实现网站信息的层次性(hierarchy)组织;同时可以结合对用户访问日志记录信息的挖掘,把握用户的兴趣,从而有助于开展网站信息推送服务以及个人信息的定制服务。目前,PDA(Personal Digital Assistant,个人数字助理)和蜂窝移动电话都已经可以直接接受网络信息服务。这些设备的显示界面较小,因而网站面向这些设备的设计就应该突出精品化、个性化的特点,这类特色推送服务就必须采用网络信息挖掘技术。

网络灰色信息的应用正在变得越来越广泛,用户对高品质、个性化的信息需求也将进一步推动学术界与实业界的研究开发工作。

(5)注重日常收集整理,建设相关馆藏。在日常工作中,应重视收集网上更新速度快的灰色文献信息资源,如动态报道,其中包含了很多具有重大信息价值的内容。信息工作人员日积月累,将这些信息收入现实馆藏。

2.2 互联网上灰色信息资源挖掘利用的模式

通过以上的论述,我们设计出一个互联网上灰色信息资源挖掘利用的模式,这个模式可以分为4个步骤。

(1)资源发现。即检索所需的网络文档。首先要确定所遇到的问题,然后主要利用搜索引擎之类的搜索工具进行查找、检索。

(2)信息选择和预处理。即从检索到的网络资源中自动挑选和预先处理得到专门的信息,主要利用数据挖掘工具来进行信息的深度挖掘。

(3)概括化。即从单个的Web站点以及多个站点之间发现普遍的模式。

(4)分析。对挖掘出的信息进行确认、解释,进行结果评价,可以用可视化的工具呈现数据,目的是便于整理挖掘到的信息。

经过以上几个步骤,我们就可以将散落于互联网上灰色信息作出系统整理,得到自己所需、有利于决策的有用信息。在某个信息挖掘的过程中,有时需要重复以上的某些步骤。

3 对互联网上灰色信息挖掘与利用时应注意的问题

灰色信息的挖掘与利用属于竞争情报研究的一部分,"合法性"是其活动的坚实基础,收集网上灰色信息时必须遵循职业道德,主要通过正式渠道收集公开发表的信息(即通过大众媒体公开传播的信息),而且强调在大量公开信息分析的基础上,获得所需的信息。但是,非公开发表信息的收集与分析同样是竞争情报的有机组成部分,因此,在挖掘与利用互联网上灰色信息时应注意以下两点:

(1)商业秘密的正确判断与保护。商业秘密是在生产经营中使用的,不为他人所知并可在同行竞争中占有优势的一种生产经营手段。美国学者Dennis Vakovic将构成商业秘密的法律要素归结为信息的新颖性(Novelty)、价值性(Value)和保密性(Secrecy)3个要素[4]。我国《反不正当竞争法》第10条规定,商业秘密是指不为公众所知悉、能为权利人带来经济利益、具有实用性并经权利人采取保密措施的技术信息和经营信息。从法律意义上说,权利人是否采取保密措施,通常是认定某项信息能否成为商业秘密的条件,也是寻求法律保护的前提。

信息工作中应根据法律,正确判断所需信息是否属于商业秘密。商业秘密的另一个重要特征就是权利主体的多元化。由于商业秘密的持有人对其持有的技术信息、经营信息没有法定的专有权,因而不能排除他人通过正当的独立研究掌握相似的或相同的技术或经营信息,权利主体的多元性使商业秘密的持有人在竞争中处于不稳定状态,这给灰色信息的挖掘与利用创造了有利条件。

(2)正确判断信息的可靠性和时效性。工作中对信息挖掘目标的不明确、对信息分析深度不够、过分注重信息的形式和细节,容易忽视有实质意义的建议,从而降低信息来源的权威性。只有从用户的角度出发来进行关键信息因素挖掘,增加最终挖掘灰色信息中有实质意义的信息含量,这样才能有效、合法地收集与应用这些灰色信息资源。专利和商业秘密都有一定的期限保护,过了这个保护期会变成公开信息使广大公众得以使用。

灰色信息对于许多企业确实有很大的诱惑力,以至于一些企业混淆了灰色信息与商业机密的界限,采取了一些不正当的手段如偷窃、贿赂、威胁、间谍、黑客入侵数据库等去获取,引起法律纠纷。在互联网上采集灰色信息时,必须自觉控制自己获取信息的正当行为,判断灰色信息的可靠性与时效性,才能有效、合法地使用灰色信息。

参考文献

1 费愉庆. 论灰色文献与网络信息资源. 苏州大学学报(工科版),2003(4)
2 莫泽瑞. 网上灰色文献及其收集利用. 大学图书馆学报,2001(6)
3 赵武. 灰色信息的情报价值及其开发利用. 徐州建筑职业技术学院学报,2002(2)
4 唐纳德·A. 马灿德,托马斯·H. 达文波特,提姆·迪克森编;吕传俊,周光尚,魏颖译. 信息管理 = Information management:信息管理领域最全面的MBA指南. 北京:中国社会科学出版社,2002

相丽玲 山西大学管理学院教授。通信地址:太原市。邮编030006。

曹 平 山西大学管理学院2004级研究生。通信地址同上。

(来稿时间:2004-09-20)