中国图书馆学报(双月刊)2000 年第 6 期
ZHONGGUO TUSHUGUANXUEBAO

●毕 强 温 平

网络环境下信息资源的获取策略研究

摘 要 分析了网络信息资源的分布特点,探讨了由这些特点所造成的信息获取障碍,阐述了选择网站的基本策略。参考文献 5。

关键词 因特网 网络信息资源 网站 信息获取
分类号 G352

ABSTRACT In this paper, the authors analyze characteristics of the distribution of network information resources and barriers to information acquisition caused by these characteristics, and propose strategies for the selection of Web sites. 5 refs.

KEY WORDS Internet. Network information resources. Web sites. Information acquisition.
CLASS NUMBER G352

出于互联网络具备拥有信息丰富、传输信息方便和提供信息多样等特点,登网信息消费已成为一种便捷有效的信息消费方式而大为时人推崇,并逐渐成为新的社会消费热点。但也应看到,网上信息资源分布的离散性、动态性、不均匀性、潜在性和模糊性,极大地影响了人们准确高效地获取网上信息。探讨网络环境下信息资源的获取策略,无疑对开发与利用网络信息资源具有重要意义。

1 网络信息资源及其分布特点

从信息传播或交流的角度而言,信息源是指拥有相对信息优势(即信息势)的信息媒体,包括善于积累和贮存信息的人以及生产,制作,贮存和传播信息的机构、网站等。例如,在世界范围内,发达国家更多地处于信息源的位置,发展中国家则以吸收信息为主,因特网就是以美国为主要信息源的一种洲际信息网。

信息资源的分布不同于信息资源的布局,信息资源的分布是一种客观现象,信息资源布局则是一种主观行为。在以信息为基础的因特网世界里,信息源的分布主要以网站为单位,数以千计的网站在因特网上形成一个错综复杂的广义的信息资源库。一般来说,网络环境信息资源的分布具有 5 个特点。

(1)离散性。即因特网上网站信息资源不受限制,自由发展,形式多种多样,内容千变万化,从一个网站资源可链接到更多相关或相似的网站资源,同样,这个资源也可能是另外一个资源链接而来。我们知道,WWW 是受 TCP/ IP 支持的应用协议集。它采用 URL,HTTP,HTM 和 CGI 等协议和标准进行信息的定位,存取和显示。WWW 可被静态地看作是一个全球范围的相互铰链的多媒体信息网,也可被逻辑地视为一个用 HTML 书写的巨大的分布式超文本结构。该超文本结构是由分布在许多物理站点上的超文本文档通过热键(因特网地址指针)互连而成的。一个文档任何一处都可以有一个指向另一文档或者其自身某处的热键。这种前所未有的自由度使得信息的交换和分享的潜力无穷。

(2)动态性。即指因特网上的环境是变化的。任何网站资源都有可能在短时间内建立、更新、更换地址或消失的可能,使得网上的信息资源瞬息万变。

(3)不均匀性。这有两层含义。一是质量不均匀。这是网络信息资源与传统信息资源(图书、期刊等印刷品资源)的本质区别。在我们周围大部分现实的文献信息资源都是经过不同职业的人所过滤过:书籍、期刊、杂志里面的信息经编辑把关;书目、索引等信息库由专家和学者筛选整理过;报纸、电视或广播里不论新闻、娱乐性节目,甚至广告都经过制作人员和编辑人员的过滤整理。因特网是无人管理的"网中之网",网上的资源并不像图书和期刊那样经过编辑和出版部门的权威审核,而且网站资源的提供不受任何组织或制度的控制,导致了网上资源质量的良莠不齐。二是分布的不均匀。因特网协会最近展开的一次调查显示,美国大约有 340 万台电脑与网络连接,西欧有 50 万台,

而全非洲仅有 2.71 万台,中美洲和南美洲有 1.6 万台,中东有 1.3 万台。科学技术的差距扩大了贫富差距,也使信息源分布的不均衡更加明显。互联网络是信息的载体,网络环境下信息源分布的不均衡带来了新的信息霸权。如西方发达国家对非英语国家在网上的文化渗透,在互联网上能时时感到美国文化独揽一切的味道。美国正在把它的有关法律、技术标准贴上国际化标签,通过互联网络强加给别国。目前,互联网上的英语内容占 90 %,法语只有 5 %,处于第 3 位的西班牙语约占 2 %。难怪法国司法部长雅克邦认为以英语占主导地位的互联网络是一种"新形式的殖民主义"。网络上中文信息的贫乏,也已成为一个现实问题。

(4)潜在性。因特网是完全开放的网络(并不属于任何组织或机构的管理范围),人人都有上网的自由。一方面,政府、公司、组织、团体甚至个人都可以成为因特网上信息资源的接受者;另一方面,政府、公司、组织、团体甚至个人在网上建立网站,这些使用者都有可能从单向的信息接受者转变成网站信息资源的提供者。

(5)模糊性。因特网的逐渐商业化将自己与商业广告紧密联系在一起。以前,不论是书、期刊或视听资料,都容易辨别到底是广告或非商业资料。但现在,许多企业在网上提供相关资料的同时,还在网页上向因特网用户提示该公司服务、产品介绍等。这种信息本身与商业广告相结合,就使得商业与非商业信息的界限逐渐模糊。

2　网络信息资源分布所造成的信息获取障碍

网络信息资源分布的上述特点,给我们获取网上信息带来困难。其主要原因是,WWW(网上获取信息逐渐向 WWW 方式统一)既不"理解"它所处理的信息,也无法对其内容作自动处理,从而增加了客户端用户手工操作的负担。网上信息获取是在导航机制的协助下,依照信息地址实现的。显然,在已知信息地址的情况下这种方式是快捷的。另外,文件级的文档是网上获取信息的基本单位。这些特点在信息定位和定界上引发出 3 个问题。

(1)单步信息定位。导航只允许沿着热键一步一步地追踪信息。在这个追踪过程中客户机接受了许多最终被排除的信息,从而造成冗余信息传输的浪费。例如,像 Yahoo lnfoseek 等提供了优秀的搜索引擎的大量的信息资源,但仍需根据检索到的地址信息,按照屏幕页面的导航提示一步一步地向下追踪,如果查到的地址有几百条、几千条甚至更多,要确定哪个地址符合要求就很困难。

(2)非即席信息查询。只有有限的服务器提供信息查询服务或代理这种服务。查询需要用户在服务器间来回切换。这也是造成迷失和冗余信息传输的一个原因。

(3)偏差信息定位。客户机在得到不相关信息的同时只得到所需的一部分相关信息。其原因在于 WWW 既不能排除一个文档中对用户无用的部分,也不能同时从不同的文档中获取信息。WWW 目前还不能定制信息。由于文档是由提供者按自身设想制作的,并没有一个规范,所以这种偏差往往不能控制,而且很严重。

解决网上信息资源分布所产生的问题,需要采取有效的策略。

3　网络信息资源的获取策略

网上信息获取的基本前提是网站的选择。这里阐述选择网站的几个基本策略。

3.1　界定访问者经常访问的站点

这是选择站点的首要原则。很多站点的信息发布带有一定的综合性,很可能覆盖某个行业或特定阶层的群体。对这样的网站,就要审查该站点的信息内容、信息范围、信息可用性、可得性和可浏览性等。

网页内容的评价可包括 6 个方面:(1)正确性。网页内容的正确与否直接影响到网页的使用价值。(2)权威性。包括:网页的作者、提供者、维护者是否是网页内容主题的专家或专门机构,他们的可信度如何,网页是否具有这些人或机构的联系方式,该联系方式的畅通与否等。(3)独特性。网页内容是否具有自己的特色?有哪些特别可取的信息?是否有从别的资源中无法获得的信息?等。(4)内容更新速度。网页内容的新颖性与其内容质量有密切联系。随时更新是网络资源的优势。如果一个网页自建立以来没有更新或更新频率低,则极有可能失去实效性。(5)目的及目的用户。(6)文字表达。网上的信息内容大多是靠文字传递,文字表达的质量对信息的传输有重要的影响。

范围包括广度、深度、时效及格式等。"广度"指是否包括特定主题的所有概念。"深度"指关于某一主题信息的详细程度。"时间"指信息是否限定在特定的时间段内。"格式"指其他一些类型的因特网资源(Telnet, Gopher, FTP)是否排除在外。

好的网站的信息的获得不仅容易,且能被用户有

效利用。网站的可用性和可获得性包括:(1)链接。链接主要包括两个方面:一是从该网站到别的网站的链接,二是从别的网站连到该网站。我们可从3个指标来评价网站的链接情况。首先是网站的稳定性。网站所在的地址(URL)是否稳定且容易记,将影响用户是否顺利地找到该网站。此外网站是否提供全天服务也影响用户是否能正常链接网站。其次是链接信息的正确性。网站的维护人员应定期检查网站链接的信息是否有变动,如有变动,应根据实际情况对其更新。第三,被其他网站链接的对象,从某种程度上证明该网站信息具有特定价值。(2)电脑环境需求。如果网站对电脑环境的要求高,那么用户的数量就会相对减少,一般来说,基本电脑环境下信息资源的利用率和可得性将提高。网站选择与评价应结合这一点。(3)传输速度。随着网上用户不断增加,因特网整体的速度越来越慢,一个网页即使有独特的内容,美观的设计,但如果传输速度太慢,往往也会使用户倒胃口,使信息变得毫无价值。(4)检索功能。在一个内容丰富,设计完整的网页里,以该网页内容为检索范围或具备检索功能是必不可少的。此外,还应注意检索方式及检索界面。(5)可浏览性和组织。资源是否以一定的逻辑方式进行组织使其易于查找,组织方式是否合理。

3.2　考察所选择的站点本身的经营策略、经营方法和效果

一般来说,所选择的站点应该是信息量比较大,信息的准确性比较高,信息定期更新和补充,栏目设置条理清晰而且丰富,栏目中的文字简洁、主题鲜明、重点突出,主页设计与制作比较精良。在经营方面,应有:(1)服务承诺。它是目前发达国家和地区普遍实行的一种制度。这种制度要求一切承担社会服务职能的组织和个人,按照行业要求,把服务内容、服务标准、服务程序、服务时间和服务责任向社会作出承诺。用户有理由要求ISP作出承诺,包括计时准确、收费合理、使用方便、安全、保密性好等。(2)合理收费。价格差异是用户挑选ISP的决定因素之一。ISP要求用户支付的基本费用有开户费和使用费两种。开户费一般是一次性的。使用费一般有两种收费方式:一种是按照实际用时计费,一种是固定的月租费。

随着因特网的持续发展,网站可能会具有一些新特点。我们应关注因特网,适应网站资源的变化。

依托网络技术自动获取网上信息资源,可采用的策略有:

(1)在分布式的环境下,采用"客户机一服务器组"计算模型,即在客户机提出一个查询请求后,将会有一组服务器协调求解,联合应答,构成服务器组的服务器不是一成不变的。这样通过查找热键的支持,实现在现行信息环境中的即席查询。

(2)通过继承,使客户机不经过中间过程即能得到目标信息,实现信息的直接定位。

(3)采用增量式逐步求精的求解策略。获取未知信息除了其细节是未知的以外,其数量和结构也是未知的。采用增量式逐步求精的策略,逐步由粗到精地在深度上和由少到多地在广度上完善应答。应注意的是:深度与广度是多个层次的相对概念,因而这种策略有好的并发性。

(4)深度搜索和广度归并。为获取未知信息,可从当前已知信息出发,逐步向深度(包括元素的嵌套深度和热键连接深度)推进。每一步推理都引起搜索历史的变迁和当前目标的变化,并且从当前元素开始依据当前历史记录和目标寻求一些能使目标范围缩小的语义连接或嵌套指示的路径继续进行搜索,直到目标被尽可能多地满足。在搜索过程中可能会有多个相同元素出现(例如有多个元素满足查找连接),这时搜索过程会按照知识的结构建立归并框架,将它们归并为一个对象,然后再对它展开搜索。以元素为查询单位可以最大限度地排除冗余信息:通过目标被尽可能多地满足搜索,又使客户端能得到完整信息。这样通过在深度上的信息缩写和广度上的综合,提供定制信息。

总之,以需求为导向,以网络技术为依托,合理运用获取策略对于网络信息资源的开发利用具有重要意义。

参考文献

1　毕强. 网络信息资源管理. 长春:吉林科技出版社,1999
2　李立新. 对象语义网络WWW信息结构模型. 计算机科学,1999(11)
3　Kuokka D. Harada L. Matchmarking for Information Agents. IJCAI-95
4　阚劲松等. 一种从因特网上自动获取行业信息的方法. 微型机与应用,1999(9)
5　Hyach inch S. Nwana. Software Agent: An Overview Knowledge Engineering Review,1996

毕　强　吉林工业大学经济管理学院信息管理系教授。通讯地址:长春市人民大街142号。邮编130022。

温　平　长春广播电视大学副教授。通讯地址:长春市。邮编130022。

(来稿时间:2000-04-10)