

信息基地的构架和建设模型¹⁾

苏贵洋 王永成 马颖华

(上海交通大学计算机系, 上海 200030)

摘要 用户在过载的 Internet 信息检索中, 往往不能查询到自己希望获得的内容。为此, 本文提出了信息基地的概念, 它是对特定领域内因特网信息资源进行“精”加工处理和长期保存的“小型数字图书馆”, 将在特定领域内最大限度的便利用户的检索需求。本文对信息基地的构架和建设中涉及的关键技术进行了初步的剖析, 还介绍了上海交通大学 OA 实验室正在建设中的“中文信息处理信息基地”系统流程。

关键词 信息基地 数字图书馆 中文信息处理

Architecture of InfoBase

Su Guiyang, Wang Yongcheng and Ma Yinghua

(Department of Computer Science, Shanghai Jiao Tong University, Shanghai 200030)

Abstract In the overload Internet information searching, user always can not get the content what he/she really want. In this paper, we propose to use InfoBase—a special kind of Digital Library which collect, purify and save Internet information on specific researching area, in order to give maximum convenience to user's information retrieval. The InfoBase's structure is put forward and some of its key techniques are described in detail. Also a system which is being constructed by OA Lab of Shanghai Jiao Tong University is introduced.

Keywords InfoBase, digital library, Chinese information control.

1 前 言

Internet 的发展速度是惊人的, 随着网站数目的迅速增长, 网上的信息量更是以几何级数剧增。如今世界第一搜索引擎 Google 宣称它索引到的页面已经达二十余亿(2,073,418,204)张网页^[1]。如果根据 Lawrence 的理论^[2], 目前搜索引擎中网页索引量最大的搜索引擎也不超过整个 Internet 网页的 1/6, 那么可以推算如今 Internet 上至少存在 120 亿张网页。

所有这些信都散布在无数的服务器上, 使人

们无法收集甚至无法发现它们。为了有效地利用数量如此庞大的信息资源, 搜索引擎(Search Engine)成为当前解决这一问题的有效工具。中国互联网络发展状况统计报告 2002 年 1 月的最新统计资料表明^[3], 有 76.3% 的人是通过搜索引擎发现网站的, 用户第二种经常使用的网络服务(62.7%)是搜索引擎。

然而, 即便有最强大的搜索引擎(例如 Google), 当用户面对一个具体问题去查询时, 面对的常常是杂乱无章的返回结果。用户往往在各个网站或者网页之间跳来跳去地查找自己所需的内容, 却也未必

收稿日期: 2002 年 5 月 3 日

作者简介: 苏贵洋, 男, 1974 年生, 上海交通大学计算机系博士研究生, 专业研究方向为网络信息智能处理。王永成, 男, 1939 年生, 上海交通大学计算机系教授, 博士生导师, 主要从事中文信息处理领域的研究工作。马颖华, 女, 1973 年生, 上海交通大学计算机系博士研究生, 专业研究方向为文本自动处理和自动标引。

1) 国家自然科学基金资助项目(60082003)

得到他希望得到的结果。对于用户共同关心的话题或者内容,如果存在一个权威的、全面的信息集散基地,囊括了绝大部分相关资料,这将会是用户非常乐意使用的。所以,在搜索引擎日益扩大自己索引页面——即追求“广”的同时,用户同样迫切需要另外一种“专”的服务,即针对用户固定的关心的内容,来建设一个个的信息基地(InfoBase)。

对此,本文提出了信息基地的概念。所谓信息基地,狭义上讲,可以理解为是对特定领域中因特网信息资源进行“精”加工处理和长期保存的“小型数字图书馆”。它集中在特定领域,可以迅速的结合动态网络信息,真正在特定领域内做到最大程度地满足用户需求。信息基地从某种程度上属于广义的数字图书馆的一部分,它所拥有信息的领域可大可小,而且更贴近群体用户的需求。但由于信息基地的信息更多来源于因特网,而因特网信息发布的随意性、自由性较大,信息基地的建设和组织与基于正式出版的文献资源的数字图书馆存在很大差异,主要表现在信息的精处理加工过程以及动态信息的保存维护过程。

一个个狭义的信息基地,就可以组成全国乃至全球的信息库。广义上讲,一个个狭义的信息基地,即一个个特定领域的信息集合,就可以组成为大型的、综合的信息基地,为各种用户提供最好的服务。

本文所提及的信息基地,属于狭义信息基地范畴。也只有研究好狭义的信息基地,才能更好的建设广义的信息基地。

2 信息基地应该具备的功能

在设计信息基地之前,有一个良好的规划,确定它们需要具备的功能是非常关键的问题。为了回答这个问题,首先应该考虑用户在网上浏览查询过程中遇到的问题和困难:对于用户来讲,信息过载——过多的信息等待用户去提取是用户面对的最大障碍。既然信息基地的建设就是为了从一定的方面帮助用户最快获得网络信息、最好的管理网络信息,它们至少应该具备以下功能:

- 信息基地应该是某一特定领域内最权威的信息集散地。建设信息基地的一个根本目标,就是为了信息的“精”;同时,也应该实现在特定领域内信息的“广”,即最大限度收集整理在某一特定领域的相关内容。

- 信息基地应该快速反映某一特定领域内信息

的动态变化。对于网络信息的日新月异,信息基地一个重要的功能就是应该可以快速获取最新的信息并进行分类、整理,使得用户无论在某一特定领域内查找历史、现状还是最新内容,都无需再自己进行各种搜索,只需要在信息基地查询就可以及时获得。为此,信息基地也应该拥有自己的机器人(Robot),不断地进行信息搜集和整理。

- 用户可以方便快捷地使用信息基地。搜集某一特定领域的各种相关内容仅仅是建设信息基地的前期工作,而如何组织这些内容,供用户最便捷地获得自己的需要是建设信息基地的另一个关键问题。从这一点上来讲,信息基地应该为用户提供良好的人机界面,允许用户反馈对信息基地使用情况和满意度。信息基地还应该给管理员一个方便的管理界面,最大限度地提高管理的效率和效果。

3 信息基地的构架和建设模型

根据目前信息基地的特点,信息基地的建设可以分解为如下四项工作:

- (1)信息的搜集。
- (2)信息的整理加工。
- (3)信息的检索。
- (4)信息的发布和传输。

信息基地的组成如图 1 所示。

3.1 信息的收集

信息收集的目的是针对特定领域,对领域内的各种主题进行网络信息收集,以此建设信息仓库与专题库。收集的办法可以分为人工搜集和自动搜集。

在人工搜集,可以给相关信息点(例如网站)签订协议,由对方提供信息,这称之为协议搜集。Yahoo 的信息搜集,就有大量信息是通过这种方式提供的。对于特定信息,还可以根据需求进行搜集。

应用面更广的是自动搜集,自动搜集可以由机器人(Robot)或者间谍(Spy)来完成。机器人可以不加任何区分地向回搜集信息(全面搜集),可以对目前比较热门的话题进行搜集(热点搜集),也可以对信息进行更加深入的搜集(深入搜集)。目前最受用户欢迎的当属“定点”、“定题”和“定向”的常规搜集。

自动搜集过程的关键技术除了搜索 Agent 的设计外,还包含一个领域内主题描述、更新问题。领域

内的主题可以形成领域内概念空间,如何维护该概念空间,以及如何从概念空间出发进行搜索是研究的主要内容。

3.2 信息的整理加工

将信息收集回来之后,就需要对信息进行整理加工,这一步工作,将建成信息库和索引库,这里包含了“精”加工、著录标引、分类、聚类等处理过程。

“精”加工过程包括“去糟”(去除反动、黄色等信息)、“去重”(去掉重复的 URL、内容等),以及信息的过滤与压缩工作等。“精”加工过程最为重要的工作是对每个页面进行评价,评价的算法很多^[4,5],常见的是使用 PageRank 的方法^[6]。评价的目的是对用户的检索结果进行排序,以期提供给用户信息量大、普遍评价高、更符合用户需求的网页。

著录的目的是对网页的来源、作者等信息进行标注。标引的目的是形成供检索命令查询的索引库,加快文档的检索和处理速度。标引又可以分为文标引和库标引。其中文标引可以分为:特征标引、分类标引、主题标引、字串标引、概念标引和热点标引。

最后的加工工作是对信息进行聚类和分类储存。分类和聚类的目的在于网页存储的有序化,加快和便于用户有针对性的查找。目前存在多种固定分类法,在固定类别下的网页数据不断增加后,还有必要对类内网页进行一定的聚类。聚类在一定条件下可以转换为分类法的一部分。在分类存储时,由于狭义的信息基地往往针对特定领域进行,所以对特定领域的信息,可以特色库、精品库、档案库等来分类。

3.3 信息的检索

如何进行“多快好省”的检索是信息科学研究的重要领域。检索可以分为广义的检索——分布式全球式的检索,也可以是狭义的对本地信息库的检索。后者的检索方法相对简单,可以应用传统的搜索方法,例如关键词匹配、案例匹配等。也可以使用更先进的方法:自然语言搜索,例如问句式提问搜索。

这一部分的主要内容是提供对各种检索方法的支持,并根据不同的检索方法形成信息的检索元数据库,以加快检索速度。由于信息基地通过对领域内各个主题信息的收集和整理,信息基地的检索过程可以主要依靠本地信息库进行。信息基地建设过程中分离信息的收集过程和信息的检索过程能在一

定程度上降低系统的复杂度。另外,用于检索的元数据库的建设是和信息整理加工部分的元数据库建设分离的,这有助于在支持不同检索方法的同时,建设稳定的信息资源存储和处理系统。针对分布式全球式的检索也是用户检索需求的一部分,也是信息基地建设不可忽视的一部分。

3.4 信息的发布

有了先前工作的基础,最后就需要对信息进行发布。信息的发布是信息基地最终的应用界面,主要内容是用户如何来使用信息基地。发布可以分为主动发布和被动发布,还可以对信息进行变形,如机器翻译就是最常见的机器变形。

所谓被动发布即等待用户主动来查询和检索。最常见的方式有网页浏览(HTTP)、文件下载(FTP)、专题公告(NNTP、BBS)等。

而更富有个性化的趋势是主动发布,即将不同用户各自不同的兴趣保存下来,等有了相关信息之后,主动发布给各个用户或用户群^[7]。这方面早期常用的有推技术,主动的把信息推给用户。具体的方法有对频道进行预定,或者通过传统的邮件列表等来实现。目前的潮流则是对不同的用户进行用户信息定制,这其中包含了用户需求挖掘、需求表达、需求解析、需求转达和需求满足等多方面的技术^[8,9,10],也是信息科学正在攻克的难关。

4 信息基地的试验

信息基地的建设内容涉及面极广,工作量巨大。目前由上海交通大学计算机系 OA 实验室构建的“中文信息处理”信息基地,已经对中文信息处理领域的各种信息、资料进行了初步的收集和整理。在初步建立信息库之后,用户界面及信息的自动加工过程将是下一步信息基地建设重点。本文将对这两方面问题进行较为详细的讨论。

4.1 信息加工处理过程

正在建设中的“中文信息处理”信息基地采用关键词形成领域内的概念空间,所以关键词库是概念空间的数据存储部分,其中保存了关键词以及关键词之间的关系数据。

如图 2 所示,在信息基地的管理维护中,首先就是要对关键词库进行维护和扩充,这个过程是用户以及管理员共同完成的。用户在检索过程中可以提

交他所感兴趣的关键词,补充到基地关键词库中。关键词添加、更改、删除过程都属于管理员的可控范围。管理员还可以人工干涉用户提交关键词的过程。另外,网上信息被收集并存入信息基地后,需要经过“新文献高频抽词”和“样本库高频抽词”,以期能够自动发现新关键词,通过适当的审核过程添加到关键词库中。其中,审核过程应该是在管理员半人工干预下,经过关键词同现运算来自动进行。

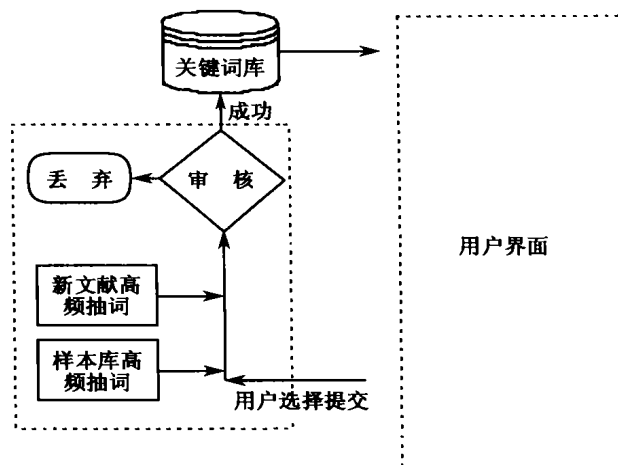


图2 关键词管理流程

信息基地不断派出机器人对网络信息进行搜索和更新,搜索返回的信息需要经过处理并存入基地信息库。基地的信息库有两大组成部分,一是文献文件库,保存有搜索返回的信息;二是文献信息库,保存文献的著录、分类、标引信息,也就是关于信息的信息(通常也称为二次信息)。这些存储处理过程大部分将是自动进行,但是同样处于管理员的控制之下,如图3所示。

4.2 库内检索过程

库内信息的检索过程依照库内数据的整理处理过程,同样是依赖形成概念空间的关键词进行的。其他非关键词检索的检索类型,例如案例检索、提问式检索,将最终转换为关键词检索并得以执行。下面是用户检索流程,如图4所示。

用户以关键词进行内容的查询,用户提交的关键词将在关键词库中进行匹配,如果匹配成功,则输出相应文献信息,并询问用户的满意度;如果关键词在关键词库中不能匹配到,在进一步的模糊匹配中也不能命中,或者用户对输出文献的满意度低,则都进入用户提交关键词界面,允许用户手工输入关键词。

词。这些关键词将由管理员确认后,进入关键词库。

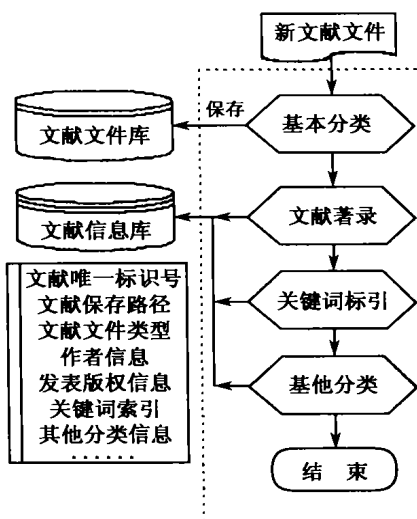


图3 管理员对文献的管理

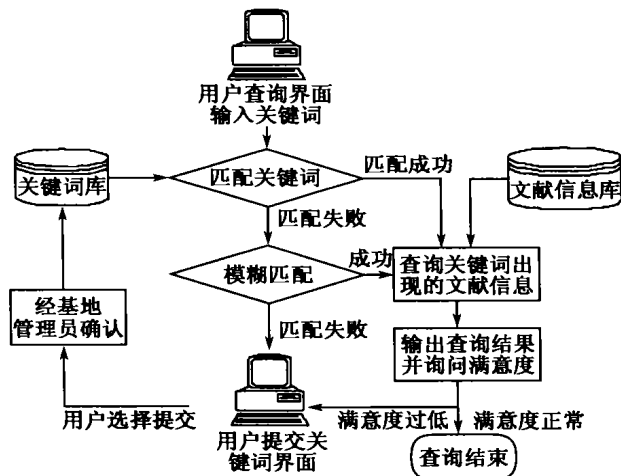


图4 用户检索流程

5 结论和展望

上海交通大学计算机系 OA 实验室根据上述架构和建设模型,正在进行“中文信息处理”信息基地的建设,希望能够在该领域内进行有益的研究和探索。

通过建设一个个的信息基地,就可以对用户比较关心的特定领域进行全面的、规范化的整理,大大方便了用户在特定领域的信息获取。信息基地可以在实践生活中得到非常好的应用,例如针对用户最常见的电脑故障问题,可以建立相关的信息基地,则用户可以对电脑故障的原因、维修方法等各种最新

信息、各种历史资料在很短的时间内获取。各种各样的信息基地建设,将是对 Internet 信息建设的一种良好的补充和扩展。

参 考 文 献

- 1 <http://www.google.com>
- 2 S. Lawrence, C. Lee Giles. Accessibility of information on the Web. *Nature*, 400, 1999
- 3 <http://www.cnnic.net.cn/develst/2002-1/4.shtml>
- 4 Kleinberg JM. Authoritative sources in a hyperlinked environment[C]. [s.l.]: Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms. 1998. 668 ~ 677
- 5 Carrière J, Kazman R. Web Query: searching and visualizing the Web through connectivity[OL]. <http://www.cgl.uwa-terloo.ca/Projects/Vanish/webquery-1.html>, 1997
- 6 Page L. The Page Rank Citation Ranking: Bring Order to the Web[OL]. Stanford Digital Libraries Working Paper, <http://www.diglib.stanford.edu>, 1999
- 7 Hsien-Chang Tu, Jieh Hsiang. An architecture and category knowledge for intelligent information retrieval agents. *Decision Support Systems*, 28(2000)255 ~ 268
- 8 E. S. Lee, R. Okada and I. G. Jeon. Agent-based Support for Personalized Information with Web Search Engines. *International Conference on HCI*, Aug. 1997
- 9 Mitch Cherniack. Expressing User Profiles for Data Recharging. *IEEE Personal Communications*, August 2001, 32 ~ 38
- 10 R. Barrett, P. Maglio, and D. Kellem. How to personalize the Web. In *Proc. ACM CHI'97*, Atlanta, USA, 1997

(责任编辑 马兰)