基于元数据与 Z39.50 的分布协作式 Web 信息检索*

王继成, 杨晓江, 潘金贵, 张福炎

(南京大学 计算机软件新技术国家重点实验室,江苏 南京 210093);
(南京大学 计算机科学与技术系,江苏 南京 210093)
E-mail: wjc@graphics.nju.edu.cn
http://cs.nju.edu.cn

摘要: Web 上大量的异质、分布、动态的信息造成了"信息过载". 如何有效地为用户提供 Web 信息检索已经成为一项重要的研究课题. Web 搜索引擎部分地解决了信息检索问题,然而其效果却远远不能令人满意. 提出了 Web 信息检索的分布协作策略以取代传统的集中式信息检索方式;给出了一种新的 Web 信息检索系统模型,该模型支持对 Web 文档的元数据进行检索,并采用 Z39.50 协议作为接口标准,以克服不同信息检索系统之间的访问异构性. 在此基础上,设计了一个分布协作式 Web 信息检索框架,用以帮助用户有效地进行 Web 信息检索.

关键词: Web;信息检索;搜索引擎;协作;元数据;Z39.50
中图法分类号: TP393 **文献标识码**: A

Web 从 1991 年出现以来,经过短短几年已经发展成为一个巨大的全球化信息空间. 有数据[1]表明,在 1999 年 2 月,Web 上大约有 2.8×10^6 台服务器,存储了 8×10^9 个页面,信息量高达 15TB. Web 信息的大容量、异构性、分布性、动态性等特点造成了"信息过载",如何有效地为用户提供 Web 信息检索已经成为一项重要的研究课题.

20 世纪 60 年代以来,信息检索领域取得了许多研究成果. 这些成果被成功地应用在 Web 上,产生了搜索引擎,例如 Altavista,Yahoo 等. 搜索引擎部分地解决了信息检索问题,但其效果还远远不能令人满意. 我们注意到,搜索引擎通常采用的是典型的集中方式,它们都试图遍历整个 Web,对其上所有的文档生成巨大、集成的全文索引,以供用户检索. 这种集中方式带来了一些严重的弊端,主要表现在:(1) 资源消耗太大:包括占用大量网络带宽,搜索引擎自身昂贵的硬件设施等;(2) 覆盖度有限:Lawrence 等人 1998 年在 Science 杂志上发表的一份研究报告[2]表明,任何一个搜索引擎索引的 Web 页面都不到页面总数的 1/3;(3) 维护困难:搜索引擎索引数据库的更新频率有限,往往会产生索引失效[3]. 元搜索引擎,如 MetaCrawler 等,通过综合多个搜索引擎的结果,在一定程度上扩大了覆盖度. 但是,元搜索引擎对搜索引擎的依赖,使得上述问题无法从根本上解决. 随着 Web 的迅速发展,集中方式已经不能适应信息检索服务的需要. 一方面,需要管理的信息资源极其巨大,任何一个集中式系统都无法完全满足需求;另一方面,各个集中式系统各行其事,重复建设.

在本文中,我们首先提出了 Web 信息检索的分布协作策略,以取代搜索引擎所采用的集中方式;然后给出了一种新的 Web 信息检索系统模型,该模型支持对 Web 文档的元数据进行检索,并

* 收稿日期: 1999-08-23; 修改日期: 2000-01-20
基金项目: 国家自然科学基金资助项目(60073030);江苏省科委"九五"科技重点攻关资助项目(BE96017)
作者简介: 王继成(1973-),男,江苏镇江人,博士,讲师,主要研究领域为信息检索与挖掘;杨晓江(1965-),男,江苏南通人,博士,副教授,主要研究领域为中文信息处理,网络信息服务;潘金贵(1952-),男,江苏镇江人,教授,博士生导师,主要研究领域为中间件,Agent 技术;张福炎(1939-),男,江苏常州人,教授,博士生导师,主要研究领域为数字化图书馆,多媒体技术.

采用Z39.50协议作为接口标准,以克服不同系统之间的访问异构性.在此基础上,我们设计了一个分布协作式Web信息检索框架,并对其进行了初步实验.结果表明,本文的方法具有良好的透明性、可扩充性和互操作性,能够同时提高检索速度和精度.

1 Web信息检索的分布协作策略

作为CERN内部的协作环境而诞生的Web,目前已经成为全球信息共享的基础结构,人们可以不受时间和空间限制进行信息的发布和浏览[4].集中式信息检索与Web的分布协作本质之间存在着无法避免的失配问题.为此,我们提出Web信息检索的分布协作策略,并认为这是Web信息检索未来的发展方向.

1.1 概念与定义

下面,我们给出一些概念及相应的定义,以便对分布协作式Web信息检索系统地进行分析.

定义1. Web信息空间R,是指Web上可检索的文档信息的全体集合.需要说明的是,有些文档,例如,由Web服务器端动态生成的页面,是无法直接访问或检索的,因此不包含在R中.

定义2. Web信息检索,是指从Web信息空间R中找到与给定的查询请求q相关的、恰当数目(记为n)的文档子集C.Web信息检索的过程对应于一个映射$\xi:(R,q,n)\rightarrow C$.

定义3. Web信息空间R的一个划分,是指集合$S=\{S_1,\ldots,S_n\}$,其中$S_i\subseteq R, S_i\neq\varnothing\ (i=1,\ldots,n)$,且$\bigcup_{i=1}^{n}S_i=R$.

定义4. Web信息空间R的一个划分S是相容的,当且仅当$\exists i\exists j,(i\neq j)\wedge(1\leqslant i,j\leqslant n)\wedge(S_i\cap S_j\neq\varnothing)$.反之,$S$是不相容的.

定义5. S中的每个元素S_i称为Web信息子空间;可以对各个S_i作进一步的分割,从而得到R的层次式划分$S=\{\{S_{1,1},\ldots,S_{1,j}\},\ldots,\{S_{n,1},\ldots,S_{n,j}\}\}$.

定义6. Web信息空间R的划分子∇,是从R到S的一个映射,即$\nabla:R\rightarrow S$.例如,当∇_1为工业领域,∇_2为地理区域时,可以对R作如下简单的不相容划分:$R\xrightarrow{\nabla_1}\{\langle.\text{com}\rangle,\ldots,\langle.\text{edu}\rangle\}\xrightarrow{\nabla_2}\{\{\langle.\text{com.cn}\rangle,\ldots,\langle.\text{com.fr}\rangle\},\ldots,\{\langle.\text{edu.cn}\rangle,\ldots,\langle\text{edu.fr}\rangle\}\}$.此外,还可以将学科主题等因素作为划分了,得到$R$的较为复杂的相容划分.

1.2 分布协作式Web信息检索

Web信息检索的分布协作策略,是指使用划分子∇对Web信息空间R进行分割,得到R的一个划分S.对于每个Web信息子空间S_i,建立一个信息检索系统IRS$_i$(information retrieval system)以提供对S_i的检索服务.这些系统分布在Web上,构成了一个协作检索群体IRC(information retrieval community),如图1所示.用户可以根据自己的需要向一个特定的IRS提出检索请求.当用户的检索请求q涉及到多个(记为k)IRS$_i$时,需要将q分解为对各个IRS$_i$的子查询,即$(R,n,q)\Rightarrow\overset{k}{\underset{i=1}{\Omega}}(S_i,n_i,q_i)$,其中$\Omega$为布尔操作;各个IRS$_i$返回结果的组合构成了检索结果集,

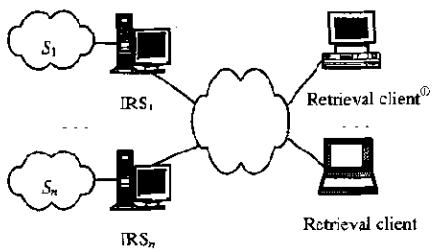


Fig. 1 Distributed and cooperative Web information retrieval community

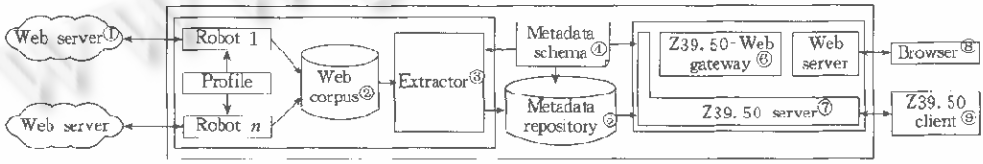
图1 分布协作式Web信息检索群体

即 $\overset{k}{\underset{i=1}{\Psi}} C_i \Rightarrow C$,其中 Ψ 为集合操作.

采用上述策略,各个 IRS 所要管理的 Web 信息量相对缩小,可以降低消耗,便于维护;同时,各个 IRS 之间通过相互协作,扩大了覆盖度.此外,当某个 IRS 出现故障时,其他的 IRS 仍然可用.因此,这种策略可以有效地克服集中方式的不足,提高信息检索的质量.

2 Web 信息检索系统模型 IRSM

IRS 作为 IRC 的基本组件,其设计的好坏对 IRC 的整体服务质量有着重要影响.常用搜索引擎的检索手段和展现手段比较单一.用户有时希望利用作者、主题、关键字等元数据来检索 Web 信息,同时也希望检索结果中包含这些元数据.但搜索引擎一般不支持基于元数据的检索,仅提供全文检索,检索结果也只是一串顺序固定的 Web 文档列表.此外,不同搜索引擎的检索接口是异构的,这给利用多个搜索引擎进行检索带来了困难.下面,我们给出一种新的 Web 信息检索系统模型 IRSM(information retrieval system model),该模型支持对 Web 文档的元数据进行检索,并采用 Z39.50 协议作为接口标准以克服各个系统检索接口之间的异构性,如图 2 所示.



①Web服务器,②Web文档库,③抽取器,④元数据模式,⑤元数据库,
⑥Z39.50-Web网关,⑦Z39.50服务器,⑧浏览器,⑨Z39.50客户端.

Fig. 2 Web information retrieval system model
图2 Web信息检索系统模型

2.1 Web 文档元数据

Web 文档的元数据是关于文档属性的数据,可以分为描述性和语义性两类.描述性元数据包括文档的名称、日期、大小、类型等属性;语义性元数据包括文档的作者、机构、内容、主题等.Web 文档的元数据在很大程度上反映了 Web 文档的特征.因此,它既可以作为一种检索手段来提高信息检索的准确度,又可以作为一种表现手段来改善检索结果的可视化.Dublin Core[5]作为一个描述 Web 文档元数据的国际标准,定义了 15 个核心的基本元素.使用 HTML 来表示 Dublin Core 的方案[6]也已发布.同时,W3C 近来制定的 XML[7]和 RDF[8,9]等规范提供了对 Web 文档资源进行描述的更为丰富的语言和框架.因此,利用元数据来支持 Web 信息检索的基础已经具备.

IRS 不可能也没有必要收集 Web 上的所有文档,它只负责管理相应的 Web 信息子空间.因此,可以通过指定起始 URL(uniform resource locators)列表以减少文档收集的盲目性.更为灵活的方法是给出与划分子相吻合的收集策略,例如,有关特定主题或特定网络域的文档,这些策略保存在 profile 中,由系统自动生成待收集的 URL.将分布在多个 Web 服务器上的文档收集到本地后,我们采用抽取器来从 Web 文档中抽取/创建元数据.其中,描述性元数据的抽取比较容易,而语义性元数据的自动抽取则比较困难,需要利用关键词自动提取、自动分类、自动摘要等技术.这些工作均在 IRS 的后台进行,如图 2 中的左半部分所示.为了适应 Web 的动态性,上述过程要周期性地重复,并需要对数据库进行递增式的更新.

2.2 基于 Z39.50 的检索接口

Z39.50 作为一个网络信息检索标准[10],规定了客户机查询服务器以及提取结果记录等过程中

所涉及的数据结构和数据交换规则.这些规则与信息资源以及检索系统的具体实现无关,使得用户可以使用统一的检索接口去检索多个远程数据库,提取满足条件的部分或者全部记录.Z39.50 目前主要应用于图书馆,我们已经成功地开发出国内第一套基于 Z39.50 的图书馆书目信息检索系统[11].在实践中我们发现,Z39.50 具有很好的灵活性和可扩充性,因此可以采用它作为 Web 信息检索的前台接口标准,从而克服不同检索系统的异构性问题.如图 2 中的右半部分所示.

Table 1 An instance of the mapping between Web document metadata and Bib-1/Use

表 1 Web 文档元数据与 Bib-1/Use 属性之间的映射示例

HTML meta tag①	Dublin Core	Bib-1 attribute②
meta name="dc.title" lang="zh" content="基于...的分布协作式 Web 信息检索">	Title③	att 4 Title
<meta name="dc.subject" lang="zh" scheme="中图法分类号" content="TP391">	Subject④	att 21 subject-heading
<meta name="dc.date" scheme="ISO 8601" content="1999-8-15">	Date⑤	att 30 date
meta name="dc.description" lang="zh" content="Web 上大量、异质..."	Description⑥	att 62 Abstract
<meta name="dc.creator" lang="zh" content="王继成">	Author⑦	att 1003 Author
<meta name="dc.identifier" scheme="URL" content="http://dlib.nju.edu.cn/...">	Identifier⑧	att 1007 identifier-standard
<meta name="dc.format" scheme="MIME" content="text/html">	Format⑨	att 1013 format-id
<meta name="dc.type" scheme="DCRT" content="text/paper">	Type⑩	att 1034 content-type
...	...	...

①标记,②属性,③标题,④主题,⑤日期,⑥摘要,⑦作者,⑧标识符,⑨格式,⑩类型.

通过建立映射表,我们可以用书目检索属性集 Bib-1 中的 Use 类型属性来表达对 Web 文档元数据的检索,表 1 中给出了二者之间的映射关系.检索结果的返回则采用一般记录语法 GRS-1.为了能够表达更为丰富的文档检索语义,我们还对 Z39.50 作了如下扩充:

(1) 在 Bib-1/Use 中增加新的属性项,包括文档的得分 Score、文档的序号 Rank、文档使用的语言 Language 等.

(2) 在 Bib-1 中增加新的属性类型 Modification,用于说明文档检索时对大小写敏感(中文简、繁体敏感)CaseSensitive、同义词 Thesaurus、禁用词 StopWord 等选项的处理要求.

(3) 在 Bib-1 中增加新的属性类型 Item,用于说明查询项的特征,如查询项语言 Language、权重 Weight、出现次数 Count 等.

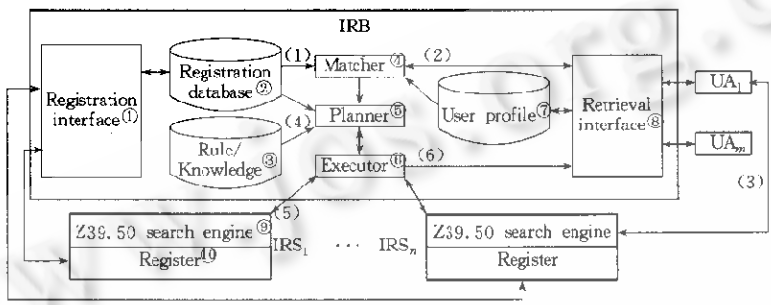
在使用 Z39.50 协议提供 Web 信息检索服务时,客户端要使用支持 Z39.50 协议的检索工具.考虑到 Web 浏览器使用的广泛性,我们在 IRSM 中加入了一个 Z39.50-Web 网关.该网关能够将用户通过 HTML 表单提交的检索请求转化为 Z39.50 服务器支持的格式,并将 Z39.50 服务器返回的结果转化为 HTML 格式.这样,用户可以通过浏览器对 IRS 进行访问.

3 分布协作式 Web 信息检索框架

在分布协作式 Web 信息检索过程中,存在两种协作关系:IRC 和各个 IRS 之间的任务/子任务关系以及用户代理(普通浏览器或智能化代理)和各个 IRS 之间的供应商/消费者关系.在 Web 这个开放环境中,协作关系具有很强的动态性,各个 IRS 可以动态地加入或者退出 IRC,用户代理也需要动态地定位具有特定功能的 IRS.为此,我们设计了一个分布协作式 Web 信息检索框架 DCIRF(distributed cooperative information retrieval framework),以帮助用户有效地进行 Web 信息检索.

3.1 DCIRF 组件

DCIRF 中包含 3 类组件:分布在 Web 上的基于 IRSM 的信息检索系统 IRS、用户代理 UA (user agent)以及 Web 信息检索中间商 IRB(information retrieval broker),如图 3 所示. 当 UA 有明确的检索目标 IRS 时,可以直接向该 IRS 提出检索请求. 但是,在 IRC 中 IRS 的加入和退出是动态的,而且在大多数情况下 UA 并没有明确的目标 IRS. 此时,IRB 作为 IRS 与 UA 之间的智能化中间件,能够管理 IRS 的动态加入和退出,并对 UA 提出的需求和 IRS 提供的服务进行动态匹配和监视.



①注册接口,②注册数据库,③规则知识库,④匹配器,⑤计划器,⑥执行器,
⑦用户profile,⑧检索接口,⑨Z39.50检索引擎,⑩注册器.

Fig. 3 Distributed cooperative information retrieval framework
图3 分布协作式Web信息检索框架

3.2 服务元数据注册

IRB 内置的注册数据库用于存放 IRS 的元数据,即对 IRS 所提供的服务进行描述. 每个 IRS 的元数据可以表示为一个五元组:(id,handle,Content,Capability,Constraint). 其中,id 为 IRS 的惟一性标识符;handle 为 IRS 的句柄(网络地址等);Content 为服务内容,包括该 IRS 所管理的 Web 信息子空间的类型、领域等;Capability 为服务能力,包括该 IRS 所支持检索的 Web 文档元数据、关系操作(截词、邻近)、布尔操作等;Constraint 为服务约束,包括该 IRS 对检索语法以及返回结果数目等的限制.

IRB 的注册接口提供 3 条原语供 IRS 进行注册,包括:① Register:IRS 向 IRB 登记自己的服务信息,从而加入 IRC;② Unregister:IRS 从 IRB 中删除自己的服务信息,从而退出 IRC;③ Modify:IRS 向 IRB 更新自己的服务信息.

3.3 UA 检索请求

UA 通过 IRB 检索接口与 IRB 相交互. 按照作用时效,可以将 UA 检索请求分为一次性请求和持续性请求. 持续性请求保存在用户 profile 中,由 IRB 负责监视并定期向 UA 推送新的检索结果. 按照结果类型,可以将 UA 检索请求分为目录性请求和细节性请求. 通过提交目录性请求,UA 从 IRB 获得目标 IRS 的句柄以用于后续检索过程. 通过提交细节性请求,UA 从 IRB 直接获得 Web 信息检索结果(对目标 IRS 的检索由 IRB 完成,该过程对 UA 而言是透明的). 针对以上 4 种请求类型,IRB 检索接口提供 Match,Get,Monitor,Subscribe 四条原语供 UA 使用. 见表 2.

Table 2 Request type and primitive

表 2 请求类型与请求原语

	One Off query①	Durative query②
Directory query③	Match	Monitor
Detail query④	Get	Subscribe

①一次性请求,②持续性请求,③目录性请求.④细节性请求.

4 实例与分析

经过两年的研制和开发,我们成功地实现了以 IRSM 为核心的 IRS 原型系统.通过配置,IRS 可以为各种 Web 信息子空间提供有针对性的信息检索服务.在此基础上,我们对 DCIRF 进行了实验.在实验中,将关于"计算机"主题的 Web 信息子空间进一步划分为"中国计算机类主要报纸"、"中国计算机类重要学术期刊"和"国内著名高校计算机系",并设置了相应的 3 个 IRS_1,IRS_2,IRS_3.此外,还建立了一个 IRB,以管理各个 IRS,并处理用户请求.DCIRF 实验系统一方面为验证"分布协作式 Web 信息检索"的可行性和有效性提供了实验环境,另一方面也为计算机领域的教学、科研人员提供了专业检索服务.

4.1 检索实例

下面,我们以一个简单的例子来说明分布协作检索流程.假定 UA 提出以下问题:查找在 1998 年度公开发表的有关"多媒体"的中文学术论文,则 IRB 对 UA 检索请求的处理过程如下(图 3 中标出了各个步骤的序号):

(1) 匹配器从注册数据库中寻找对"论文"类型文档提供检索服务的 IRS.在实验系统中,符合该条件的 IRS 为 IRS_2.当 R 的划分是相容划分时,匹配器可能会得到多个符合要求的目标 IRS.

(2) 如果 UA 提出的是目录性请求,则匹配器将目标 IRS 的句柄返回给 UA,并转步骤(3);对于 UA 的细节性请求,则转步骤(4).

(3) UA 对各个目标 IRS 进行检索,得到检索结果,处理过程结束.

(4) 计划器利用注册数据库中有关目标 IRS 的能力、约束信息以及规则知识库中有关网络路由等知识生成对 IRS 进行检索的计划,并提交给执行器执行.在有多个目标 IRS 时,计划器还要考虑资源消耗、连接代价等因素,从这些 IRS 中选择出部分或者全部来生成检索计划.

(5) 执行器向目标 IRS 的 Z39.50 检索引擎提交元数据检索请求:标题、关键字等属性中包含有"多媒体",日期为"1998",且作者单位为"高校"的文档.

(6) 执行器将从目标 IRS 得到的检索结果返回给 UA.在有多个目标 IRS 时,执行器还要对各个 IRS 的检索结果进行综合,并进行消除冗余等后处理.

4.2 分析与评价

Web 信息检索的常用评价标准包括:精度、速度、易用性等方面.Web 的大容量和动态性等特点使得召回率难以估测,因而一般不采用[12].

首先比较集中方式和分布协作方式对检索速度的影响.由于用户在检索时的等待时间受网络流量的影响很大,因此不宜作为速度的度量指标.在本文中,我们考虑在元数据检索时所需查找的数据库记录数目 Nmeta 以及在全文检索(矢量空间模型)时所需匹配的文档数目 Ndoc,它们在很大程度上决定了系统的检索速度.以上面的检索请求为例,在集中方式和分布协作方式下,元数据检索和全文检索时的速度指标见表 3.结果显示,分布协作方式通过对 Web 信息空间进行划分,检

索速度一般比集中方式有所提高.

Table 3 Speed evaluation

表 3 速度评估

	Metadata retrieval①	Full-Text retrieval②
Distributed cooperative③	$N\text{meta}(\text{IRB})+N\text{meta}(\text{IRS}_i)=3+7590=7593$	$N\text{meta}(\text{IRB})+N\text{doc}(\text{IRS}_2)-3+7590=7593$
Centralized④	$\sum(N\text{meta}(\text{IRS}_i))=22879+7590+2442=32911$	$\sum(N\text{doc}(\text{IRS}_i))=22879+7590+2442=32911$

①元数据检索,②全文检索,③分布协作,④集中.

在 Web 环境下,通常使用前 n 个检索结果中包含正确结果的比例来度量系统的精度(记为 $P@n$)[12]. 若正确的结果数目为 m,则 $P@n=m/n$. 对于上述检索请求,我们分别考察 $n=10,20$ 以及取所有结果时的精度,见表 4. 可以看出,随着考察结果数目的增加,全文检索的总体精度迅速下降;元数据检索则比全文检索有显著改善.

Table 4 Precision evaluation

表 4 精度评估

	$P@10$	$P@20$	$P@ALL$
Metadata retrieval①(%)	10/10=100.0	18/20=90.0	23/26=88.5
Full-Text retrieval②(%)	9/10=90.0	15/20=75.0	27/89=30.3

①元数据检索,②全文检索.

信息检索系统的易用性缺乏定量的评价指标,主要取决于用户的主观评判. 从系统检索服务的最终用户来看,DCIRF 实验系统表现出了良好的透明性,借助于 IRB,UA 不需要了解 IRS 的具体内容就可以检索到所需信息. 对于系统的开发和维护人员而言,IRS 可以利用注册机制来加入和退出协作检索群体 IRC,方便了系统的扩充;同时,Z39.50 的采用有助于实现 IRB,UA 与各个 IRS 之间的互操作.

5 相关工作

目前,对 Web 信息检索的研究工作[3,13]普遍集中于如何提高单个搜索引擎的服务质量上,但迄今为止收效不大,其根本原因在于未能突破传统的集中式检索方式,且缺乏对 Web 文档中元数据的支持.

有些研究人员也曾经对多个检索系统之间的协作进行了尝试,例如,Anders Ardö 等人开发的北欧 Web 索引服务系统 NWI(nordic Web index)[14]. NWI 由分布于丹麦、芬兰、冰岛、挪威和瑞典的 5 个检索了系统构成,用户通过其中的任何一个都可以对其余的子系统进行检索. 但是,在 NWI 等系统中,用户必须显式地参与检索系统之间的协作,例如,指定待检索的目标子系统. 此外,参与协作的子系统数目固定,缺乏可扩充性,也不支持子系统的动态加入和退出. 这些缺陷使得 NWI 等系统不便于在 Web 环境下扩充.

6 结束语

在信息充斥的情况下,Web 信息检索是一个具有极大潜力的研究方向. 本文的主要贡献在于,提出了一种适用于 Web 环境的信息检索框架,该框架以分布协作作为新的切入点,以 Web 元数据和标准化检索接口为支撑,较好地解决了 Web 信息检索现有的不足. 我们下一步要进行的工作是将基于 TCP/IP 的实验系统推广到更大范围,并尝试在 DCOM/CORBA 等互操作模型上实现 DCIRF. 同时,我们还希望在以下方面作进一步的研究和探讨,包括:设计和选择合适的 Web 信息空间划分子;利用 mobile agent 技术改进 Web 信息的收集与处理等工作.

References:

[1] Lawrence, S., Giles, C. Lee. Accessibility and distribution of information on the Web. Nature, 1999,400:107~109.
[2] Lawrence, S., Giles, C. Lee. Searching the World Wide Web. Science, 1998,280(5360):98~100.
[3] Lawrence, S., Giles, C. Lee. Context and page analysis for improved Web search. IEEE Internet Computing, 1998,2(4):38~46.
[4] Berners-Lee, T. The World Wide Web: past, present and future. 1996. http://www.w3.org/People/Berners-Lee/1996/ppf.html.
[5] Weibel, S. L., Kunze, J. A., Lagozes, C., *et al*. Dublin Core metadata for resource discovery. Internet RFC 2413, 1998.
[6] Kunze, J. A. Encoding Dublin Core metadata in HTML. Internet Draft, 1999.
[7] Bray, T., Paoli, J., Sperberg-McQueen, C. M. Extensible markup language (XML) 1.0 specification. World Wide Web Consortium Recommendation, 1998.
[8] Brickley, D., Guha, R. V. Resource description framework (RDF) schemas. World Wide Web Consortium Proposed Recommendation, 1999.
[9] Lassila, O., Swick, R. R. Resource description framework (RDF) model and syntax specification. World Wide Web Consortium Recommendation, 1999.
[10] ANSI/NISO Z39.50~1995, ANSI Z39.50: Information Retrieval Service and Protocol, 1995.
[11] Yang, Xiao-jiang, Zhang, Fu-yan. Online bibliographic retrieval service based on Z39.50. Journal of Software, 1999,10(8):824~828 (in Chinese).
[12] Hawking, D., Craswell, N., Harman, D. Results and challenges in Web search evaluation. In: Mendelzon, A., ed. Proceedings of the 8th International World Wide Web Conference. 1999. http://www8.org/w8-papers/2c-search-discover/results/results.html.
[13] Brin, S., Page, L. The anatomy of large-scale hypertextual Web search engine. In: Ashman, H., ed. Proceedings of the 7th International World Wide Web Conference. 1998. http://decweb.ethz.ch/WWW7/1921/com1921.htm.
[14] Ardöa, A., Lundbergb, S. A regional distributed WWW search and indexing service——the DESIRE way. In: Ashman, H., ed. Proceedings of the 7th International World Wide Web Conference. 1998. http://decweb.ethz.ch/WWW7/1900/com1900.htm.

附中文参考文献:

[11] 杨晓江,张福炎.基于 Z39.50 的联机书目检索服务.软件学报,1999,10(8),824~828.

A Distributed and Cooperative Approach to Web Information Retrieval Using Metadata and Z39.50*

WANG Ji-cheng, YANG Xiao-jiang, PAN Jin-gui, ZHANG Fu-yan

(*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210093, *China*);
(*Department of Computer Science and Technology, Nanjing University, Nanjing* 210093, *China*)
E-mail: wjc@graphics.nju.edu.cn
http://cs.nju.edu.cn

Abstract: A mass of heterogeneous, distributed and dynamic information on the Web has resulted in "information overload". It's an important and urgent research issue to provide users with effective information retrieval service on the Web. Web search engines attempt to solve this problem, yet their effect is far from satisfying. In this paper, a distributed and cooperative strategy for Web information retrieval is proposed to substitute the centralized mode adopted by the current search engines. Then a new information retrieval system model (IRSM) is presented, which supports the retrieval of metadata about Web documents and uses Z39.50 standard protocol to unify the heterogeneous interfaces of different systems. Based on them, a distributed and cooperative information retrieval framework (DCIRF) is designed to help users search the Web effectively.

Key words: Web; information retrieval; search engine; cooperative; metadata; Z39.50

* Received August 23, 1999; accepted January 20, 2000
Supported by the National Natural Science Foundation of China under Grant No. 60073030; the Sci-Tech Project of the 'Ninth Five-Year-Plan' of Science Commission of Jiangsu Province of China under Grant No. BE96017