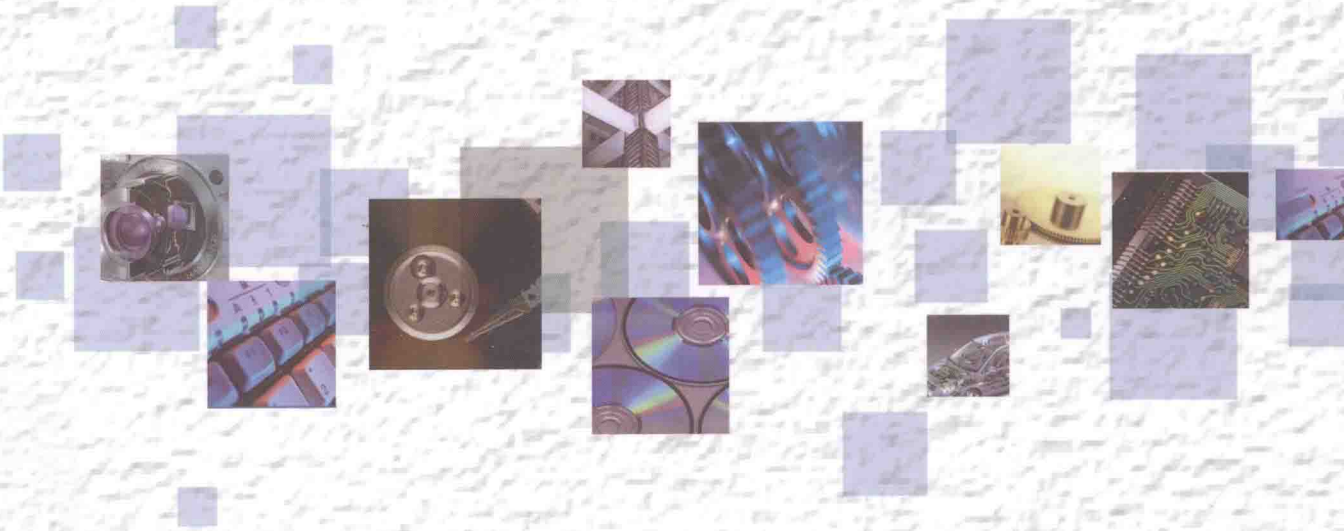
网络资源采集与数字资源长期保存
学术研讨会论文集

本书编委会 编



國家圖書館出版社
National Library of China Publishing House

网络资源采集与数字资源长期保存学术研讨会论文集

本书编委会　编

國家圖書館出版社
National Library of China Publishing House

图书在版编目(CIP)数据

网络资源采集与数字资源长期保存学术研讨会论文集/本书编委会编. --北京:国家图书馆出版社,2013.12

ISBN 978-7-5013-5239-5

Ⅰ.①网… Ⅱ.①本… Ⅲ.①计算机网络—信息资源—采集—学术会议—中国—文集 ②数字技术—信息资源—信息存贮—学术会议—中国—文集 Ⅳ.①G250.73-53 ②G203-53

中国版本图书馆 CIP 数据核字(2013)第 287439 号

书　　名　网络资源采集与数字资源长期保存学术研讨会论文集
著　　者　本书编委会　编
责任编辑　金丽萍

出　　版　国家图书馆出版社(100034　北京市西城区文津街 7 号)
(原书目文献出版社　北京图书馆出版社)
发　　行　010-66114536　66126153　66151313　66175620
66121706(传真),66126156(门市部)
E-mail　btsfxb@nlc.gov.cn(邮购)
Website　www.nlcpress.com ——→投稿中心
经　　销　新华书店
印　　装　北京科信印刷有限公司
版　　次　2013 年 12 月第 1 版　2013 年 12 月第 1 次印刷

开　　本　787×1092(毫米)　1/16
印　　张　12.5
字　　数　280千字

书　　号　ISBN 978-7-5013-5239-5
定　　价　50.00 元

论文评审专家组成员(按音序排序):

姜爱蓉　李春明　李晓明　李志尧

吕淑萍　申晓娟　孙一刚　王乐春

魏大威　邢　军　张振海　张智雄

编辑说明

随着数字图书馆建设的不断深入，网络资源采集与数字资源长期保存的工作在我国图书馆界及相关学术机构迅速展开并进入快速发展期。为促进该项工作的进一步发展，国家图书馆主办了“网络资源采集与数字资源长期保存研讨会”，会议前期向业界广泛进行了征文，同时邀请国内外专家学者、业界同仁围绕网络资源采集与数字资源长期保存的实际战略、实现模式、问题经验、规划展望等内容进行了交流探讨，取得了丰硕的成果。

现将此次研讨会的优秀论文整理成册，结集出版。希望能够通过这些文章分享业界同行的优秀实践经验，引他山之石，形成自己的创新与思考，共同推进网络资源采集与数字资源长期保存理性、实效和可持续发展。

本书编委会

2013 年 11 月

目　录

Web Infomall：一个大规模的 Web 存档系统①

闫宏飞　黄连恩　谢正茂　李晓明
（北京大学网络与分布式系统研究所）

摘　要：随着时间的流逝，中国互联网上出现过的信息资源会成为一笔宝贵的财富，会为各领域和各行业提供持续价值。为了使历史网页不会随时间的流逝而改变，并且每日不断记录网站的变化，不更改以前的保存结果，需要一个适合于 Web 规模的存档系统。在本文中，我们提出 Web Infomall 系统，它是专为搜集、组织与服务大量的网页而设计的，该系统从 2001 年以来收录了约 85 亿网页，每天还以约 100 万到 200 万网页的数量增加。在保存下来的网页集合中，可以通过时间和空间 URL 两个维度来定位一个网页。因此对于一个 URL，可能有一组在不同时间抓取的网页与其对应。在系统中，排好序的网页是依照一定的时空粒度放在一起的。这样的好处是，指定 URL 和时间，用户能够有效地检索到相应网页；或者根据 URL 范围和时间范围，获得某些批次的网页。

关键词：Web 存档　Web 存储系统　Web Infomall

互联网上的网页是不稳定的资源，这意味着它们有可能会消失。与传统媒体如报纸相比，网页的寿命很短。考虑到网络信息在现代社会中扮演着重要的角色，为子孙后代保留当前的网页信息是我们刻不容缓的责任。幸运的是，从 20 世纪 90 年代以来，保存互联网信息的意义已经被大家普遍接受，很多机构已经开始存档网页工作[1]。据 IIPC[2]，超过 30 个机构已参与到网页存档中来。

鉴于互联网的规模，而且会有越来越多的网页产生，很难想象如何来存档互联网的网页。其中一个巨大的挑战是如何有效地处理数量如此庞大的网页，目前很少有资料讨论互联网信息存档技术。

本文提出一个大型存储系统 Web Infomall，旨在为互联网信息存档。我们从 2011 年开始，一直利用这个系统在存档中国互联网上的网页。目前我们的系统中有 85 亿网页。

1　Web Infomall 的数据和服务

从 2002 年 1 月 18 日上线运行至今（2013 年 9 月），已逾 10 年。Web Infomall 保存网页 85 亿，占用磁盘空间为 73TB ×2（双份备份在线），另有一份线下备份。机器共计 18 台，其中访问服务器放在高性能机房，是 2004 年购买的 2 台机器；储存服务器是 2010 年购买的 8 台机器（机器内存为 32GB）和 2012 年购买的机器（2 个 CPU，64GB 内存，2TB ×6 磁盘容量）。线下备份机器配置为单 CPU，4GB 内存，1T ×12 磁盘容量。

① 国家自然科学基金（60933004，61073082，61272340）支持。

Web Infomall 目前主要基于网页信息库对外提供历史网页查询服务。即用户给出 URL，系统提供该 URL 的所有历史网页供用户查询。系统采用 Web 方式对外提供服务（http://www. infomall. cn），每天接受数万次访问请求，包含数百个独立 IP。系统还提供网页数据批量访问服务（http://data. infomall. cn）。该服务用户数量很少，只有少数内部用户使用。

除了上述基础服务之外，还存在两个基于 Web Infomall 信息库的应用系统，分别为事件搜索系统（http://sewm. pku. edu. cn/eventsearch）和历史事件追踪系统（http://hist. infomall. cn）。

Web Infomall 的使用方法

（1）访问：在浏览器地址栏输入“http://www. infomall. cn”，访问到 Web Infomall 系统主页上（如图 1 所示）。

（2）查询：在查询输入框输入要查询的历史网页 URL 并回车（Enter）（或者点击“开始浏览”按钮），就可以开始浏览中国 Web 在某一时间的“快照”，即得到相关历史网页存档。例如，想查北京大学的历史网页，只需在图 1 的搜索框中输入：www. infomall. cn。

（3）阅读查询结果。在图 2 中：

a. 统计栏，包括用户输入的 URL 和有关查询结果的统计数字；

b. 查询结果，包括存档网页的年代，该网页网址存档时间记录。选择任何一条记录进入，就可以浏览历史网页了。之后，系统返回的都是和该版本同时保存（同一版本）的链接网页，从而实现历史回放。

图 1　中国 Web 信息博物馆（Web Infomall）

图2　中国 Web 信息博物馆查询结果

2　Web Infomall 系统与技术

2.1　网页存储和回放设计目标

Web Infomall 系统采用分布式体系结构，基于普通服务器构建，具有良好的系统可扩展性。系统软件经过多年的建设与完善，目前运行比较稳定，在无硬件故障的情况下，可以 7 × 24 小时不间断运行半年，无须人工干预。

网页存储设计目标是力图实现对中国所有 Web 网页进行定期镜像和整理，以尽可能短的工作周期把网上的信息保存下来，主要是指所有的静态网页。

回放设计目标首要功能就是历史网页回放。即用户给出一个初始 URL 和指定时间后可以在该时段的 Web 上漫游。

Web Infomall 的发展目标是：成为中国最大的互联网信息档案馆和博物馆，为国家的政治、经济、文化和社会发展提供有力的信息支撑。Web Infomall 系统根植于一个基本认识：随着时间的流逝，中国互联网上出现过的信息资源会成为一笔宝贵的财富，会为各领域和各行业提供持续价值。因此，Web Infomall 的设计思想是：尽可能全面及时地搜集中国互联网上出现的重要信息资源，稳定可靠地进行存储。同时，基于这些信息资源，一方面开发实用有效的应用以服务于广大终端用户，另一方面建立数据访问开放平台和云计算平台供各领域二次研发人员使用，共同打造围绕 Web Infomall 的应用生态链。

2.2 系统结构

2.2.1 存储模型

Web Infomall 存储系统是被设计用于处理大量的 Web 页面，并且已经持续收集超过 10 年时间，而且未来还将不断地收集和存档互联网页信息。可以想象的是，收集的网页数量将不断扩张，所以重要的是采用一种有序的且可扩展的方式存储网页。图 3 展示了 Web Infomall 的存储模型。系统的输入是从 Web 上抓取的网页，每个网页都有对应抓取到的时间，因此 Web Infomall 里的网页时间跨度是很长且连续的。图中显示了存储在 Web Infomall 中的一个 2008 年 6 月至 8 月片段。

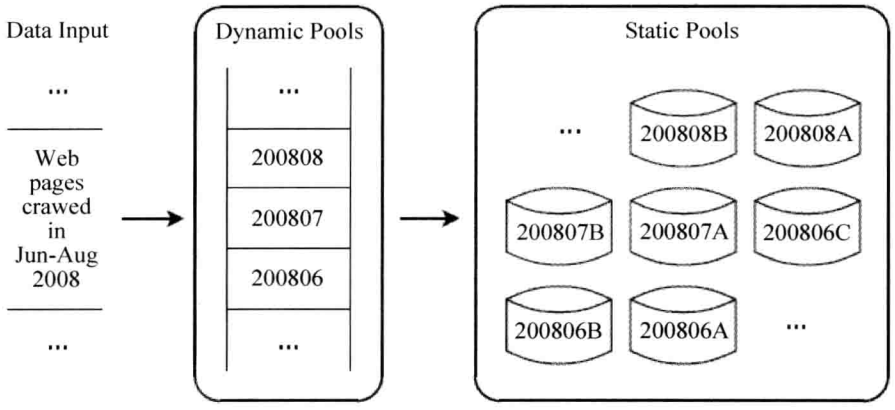


图 3 Web Infomall 存储系统

我们使用的基本存储单元是池(pool)。一个 pool 本质上是一个数据库，包含了很多 db-files 文件，每一个 db-file 都包含了大量的网页。此外，要求一个 pool 中的网页在一个时间段内，可能的话，也要求在一定 URL 范围内。这意味着仅当抓取网页的时间跨度和 URL 范围满足要求，一个网页才被收录到该 pool 中。在我们的系统中，使用一个自然月时间单位跨度来形成一个 pool。例如“200806”pool，“200806A”pool，根据 pool 的名字可以知道包含的是 2008 年 6 月 1 日到 30 日抓取的网页。

从图 3 中可以看出，在我们的系统中有两类 pool，动态 dynamic pool 和静态 static pool。它们之间的一个主要区别是，dynamic pool 中每个 db - file 中的网页是按照 URL 排序的，而在 static pool 中不止是每个 db-file 中的网页是按照 URL 排序，db-file 文件之间也是排序的。或者我们可以这样说，dynamic pool 是部分有序的，而 static pool 是全局有序的。其他的主要区别在于，dynamic pool 是可变的，这意味着网页可以插入或删除，而 static pool 一旦创建是不变的。

2.2.2 系统结构

我们的系统被设计为运行在普通 Linux 群集服务器上的，与通常配置的服务器相比，其中每个节点有一个比较大的存储容量。鉴于服务器通常比个人电脑更稳定，所以像 GFS 这类提供高可靠性的中间件系统软件[3]不是必须的。在我们的设计中可维护性和性能是两个关键问题，同时系统可靠性也需要考虑。

Web Infomall 的体系结构如图 4 所示。从图中可以看出，有两种类型的服务器：元数据

服务器和存储服务器。元数据服务器管理元数据信息,定期从存储服务器收集和提供给用户存取数据的服务。在存储服务器中含有大量的 pool 和用以提高数据访问性能的索引。存储服务器有两种类型对应到 dynamic pool 和 static pool。它是自解释的,动态存储服务器包含 dynamic pool,而静态存储服务器包含 static pool。除了这种差异,两种类型的存储服务器的工作方式非常相似,所以在图中我们只显示简单的静态存储服务器。实际上,存储服务器可以被配置为一个动态的或者静态的。

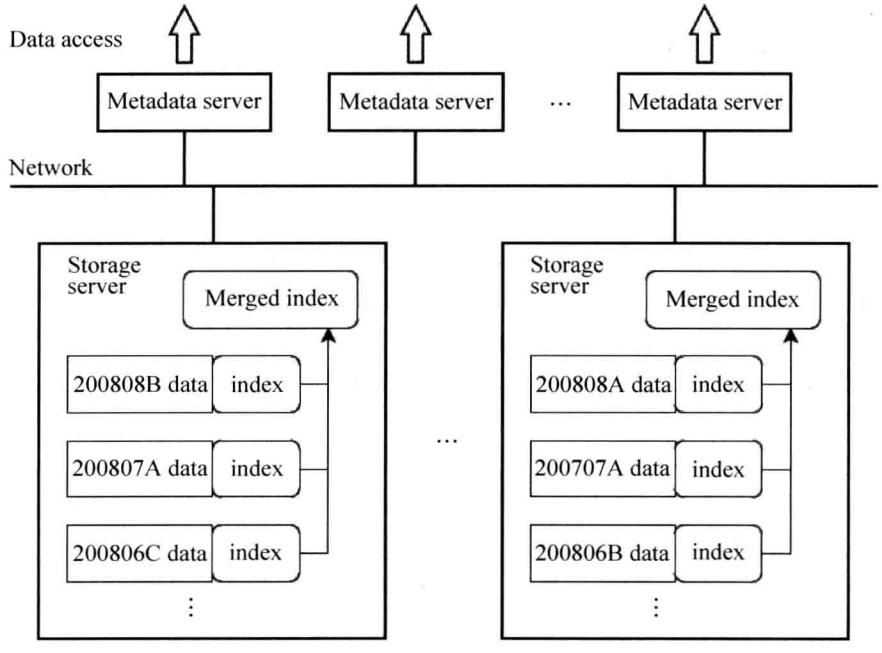


图 4　Web Infomall 体系结构

存储服务器中的 pool 是可以有多副本存在的。例如,图 4 中显示两个相同的 pool,名为“200807A”。重复的 pool 位于不同的服务器上,这样,如果一台服务器上的数据损坏仍然可以从其他服务器恢复。这个机制在很大程度上确保了我们已经收集的超过 10 年的网页数据的安全。

与 BigTable[4] 不同,它使用一个主节点来存储元数据信息,在我们的系统中没有主节点或控制中心。每个服务器由一个 IP 地址(加上一个端口)定位,它还保存了系统中其他服务器的地址列表,这个列表是可以定制的。即使有一些存储服务器宕机,其他有效的服务器还可以继续提供服务。元数据服务器定期检测存储服务器是否有效。当一个查询到达,元数据服务器将解析查询,将查询发送给相关的存储服务器,合并查询返回结果,发送给用户。

为了提高根据 URL 检索网页的性能,我们建立了全局索引,它是合并每一个存储器中的 pool 的索引生成的。当在存储服务器中创建一个新 pool 或者移除一个 pool 的时候,后台有一个进程专门来完成这个任务。

2.3　网页信息存储的天网格式

在网页的搜集过程中,我们采用增量方式抓取网页。对网页链接和主体段落进行 Di-

gest-MD5(消息摘要),利用 Bloom filter 保存摘要,保存新增、更新网页。对 Web 主机进行流量控制,包括遵守 Robots 协议,模仿人的浏览行为进行抓取。对 Web 主机进行任务分配,采用自适应的抓取频率。进行抓取状态的实时监控、展示,即通过多播把情况汇总到监控节点,在 Web 页面上展示。

抓取系统有 7 个蜘蛛节点,单节点每次启动 400—800 个抓取进程。每天抓取 5 万个网站,收获 5M—20M 个有效网页。蜘蛛节点性能是双路四核 CPU,8GB MEM,10Krpm SAS 硬盘 6 ×300GB。

将获取网页信息保存在磁盘中,需要按照规定的格式保存,便于后续处理和提供服务。下面介绍天网格式存储方案。注意这种方案只是顺序保存网页信息,没有索引文件。

原始网页信息的存储格式应当设计为适合长期保存并易于处理,可以作为终端产品提供给用户使用。考虑到终端产品使用的便利性,要求原始网页库的存储格式具备简单性的特点。

存储介质都是有寿命的,所以应当考虑当存储介质损坏时数据的可恢复性。例如:磁盘的某个扇区损坏,导致部分数据不能读出,如果剩下的数据仍然可以使用,就能将损失降到最少。对于海量数据来说,在存储和传输的过程中,由于硬件和软件问题导致数据错误是不可避免的。因此,原始网页的存储格式还应当具备容错性的特点。

1. 天网存储格式定义

根据以上考虑,天网存储格式定义如下:

(1)一个原始网页库(RAW_DB)由若干记录组成,每个记录(RECORD)包含一个网页的原始数据,记录的存放是顺序追加的,记录之间没有分隔符;

(2)一个记录由头部(HEAD)、数据(DATA)和空行(BLANK_LINE)组成,顺序是:头部 + 空行 + 数据 + 空行;

(3)一个头部由若干属性组成,每个属性(PROPERTY)是一个非空的行,头部内部不允许出现空行;

(4)一个属性包含属性名(NAME)和属性值(VALUE),并由冒号":"隔开,顺序是:属性名 + 冒号 + 属性值;

(5)头部的第一个属性必须是版本属性,属性名为 version,例如:version: 1.0,该属性表明记录的版本号;

(6)头部的最后一个属性必须是数据长度属性,属性名为 length,例如:length: 1800,该属性值必须是数据(DATA)的长度(字节数),不包括空行的长度;

(7)为简化起见,属性名必须是小写的字符串。

注:一个空行(BLANK_LINE)仅由一个换行符(line feed,LF,即 C 语言中的"\n")组成,在 UNIX 系统的显示中表现为一个空行,所以称为空行。Microsoft Windows 系统和 UNIX 系统在空行机制上有所区别:在 Windows 系统下,纯文本显示中的一个空行由一个回车符(carriage return,CR)和一个换行符组成(即 C 语言中的"\r\n");而在 UNIX 系统中一个空行仅由一个换行符组成。

2. 当前存储格式版本描述

存储格式允许有多个版本,以满足将来进行扩展的需要。

当前存储格式的版本属性为 1.0。一个记录的存储格式如下(//后为注释):

```
version: 1.0                                    // 版本号
url: http://www.pku.edu.cn/                     // URL
origin: http://www.somewhere.cn/                    // 原来的 URL
date: Tue,15 Apr 2003 08:13:06 GMT                  // 抓取时间
ip: 162.105.129.12                              // IP 地址
unzip-length: 30233                             // 如果数据经过压缩,则需有此属性
length: 18133                                   // 数据长度
                                                // 空行
XXXXXXXX                                        // 以下为数据
XXXXXXXX
...
XXXXXXXX                                        // 数据结束
                                                // 最后再插入一个空行
```

各属性说明:

version 属性为版本号,以下说明适用版本号为 1.0 的情况。

url 指该网页的 URL,如果因为网页头信息中包含 Location 字段而产生网页转向时,该 URL 为最后实际抓取的 URL 地址。该属性是必需的。

origin 指该网页的原始 URL。该属性仅在 HTTP HEAD 中包含 Location 字段而产生网页转向时存在,指向最原始的 URL。

date 属性为该网页的保存时间,保存格式为 RFC822 所制定的格式。该属性是必需的。

ip 属性为该网页所在服务器的 IP 地址。

unzip-length 属性仅在数据经过压缩时存在,记录数据未压缩时的原始长度。

length 属性记录数据长度。

若存在其他未加说明的属性,应用程序可以简单地忽略。

关于数据是否压缩的问题:天网格式并不指定数据是否必须经过压缩。但是压缩的数据必须包含 unzip-length 属性而未压缩的数据不能包含该属性。该属性同时也是解压缩所必需的。如果数据经过压缩,还应附带说明压缩算法,必要时附带压缩函数库及源代码。

3. 数据的可恢复性分析

假设由于数据遭到破坏,只得到其中一个残存的片段。则可按以下步骤找出该残存片段中所有完整的记录:

(1)特定字符串"version:",除非没有一个完整的记录,该字符串肯定能找到。记录该字符串的位置 POS。

(2)找到该字符串后,判断其后的数据是否满足存储格式(2)、(3)、(4)、(6)、(7)条件。如果任何一个条件不满足,返回 1,从记录的位置 POS 开始继续查找下一个特定字符串"version"。

(3)当满足条件 2 时,假定这是一个正确的记录,则下一个记录也必定是一个正确的记录。检查该记录满足天网存储格式(2)、(3)、(4)、(5)、(6)、(7)条件,如果任何一个条件不满足,说明原先的假定错误,返回 1,从记录的位置 POS 开始,继续查找下一个特定字符串"version"。如果条件都满足,则继续检查下一个记录是否正确。

(4)如果连续3个记录都是正确的,则认为(1)所找到的"version"是一个正确的记录的开始,可以依此提取出全部正确的原始网页。

由于原始网页是随机的,而存储格式是严格的,因此经过上述方法得到的记录为错误记录的可能性极小,是完全可以接受的。

4. 其他问题

在实际应用天网存储格式时,应该注意下面两个方面。

(1)文件打开模式。用标准IO库打开文件时有两种模式:文本(text)模式和二进制(binary)模式。在UNIX下这两种模式并没有区别,但在Windows下如果用文本模式打开,"\r\n"会被当成一个字符,而天网格式中的"length"域表示的是实际字节数,这就可能引起错误。因此,在用标准IO读取文件时,为了兼容性最好用二进制模式打开,例如FILE * fp = fopen("filename","rb")。

(2)FTP传输。FTP传输也有两种传输模式:文本(text)模式和二进制(binary)模式。传输原始网页库文件时,应以二进制模式进行传输。如果以文本模式传输,可能会出现"\r\n"被替换为"\n"或"\n"被替换为"\r\n"的现象,导致数据错误。

3 衍生数据服务和系统

基于Web Infomall系统不断增长的网页信息,我们一方面通过Web Infomall通用公共许可证来免费分发数据给研究机构,另一方面,我们也展开了数据挖掘工作。

3.1 中国事件检索与发现系统

我们实现了中国事件检索与发现系统EventSearch,该系统的数据来自4种数据源,包括网页新闻、《人民日报》、中央电视台新闻联播、微博。其中网页新闻提取自Web Infomall,包含有从2001年到2011年的1100万个网页。报纸和电视新闻视频也跨越2001至2011年。对于一个查询,系统会返回一个事件的摘要列表和与查询相关的事件分布情况(时间与地点),可以按照规模、时新性和相关性排序来展示。在事件的检测中,我们使用了一种新的基于burst词检测的方法。同时,我们还实现了一种在线的事件检测方法来提升系统的效果。

对于一个查询系统通常会返回多个事件的摘要信息和对这个查询相关的热门地点和相关事件的整体分布。每个事件摘要信息则包含:事件的起始时间,结束时间,相关地点,事件的关键词,各个源的相关报道(包括网页、报纸、微博和视频)。通过点击相关的报道和视频就可以查询相关的内容。另外,用户可以在事件的整个分布图中点击每个峰值来获取属于这个点的事件,也可以在地国中点击相关的地点来过滤事件。对每个事件的摘要,我们提供了一个事件的详细信息展示页。

如图5所示,以查询"奥运会"为例,我们可以看到在2004年8月和2008年8月有一个显著相关事件的数量激增。主要的原因是在这两个阶段分别举行了雅典和北京奥运会,这段时间内有很多与奥运会相关的报道。与这个奥运会相关的热门地点在地图标为北京。通过看第一个返回的事件摘要的关键词云,我们基本可以了解这个关于北京奥运会的开幕相关的报道。通过点击详细事件链接,我们就可以进行事件的详细信息页面。

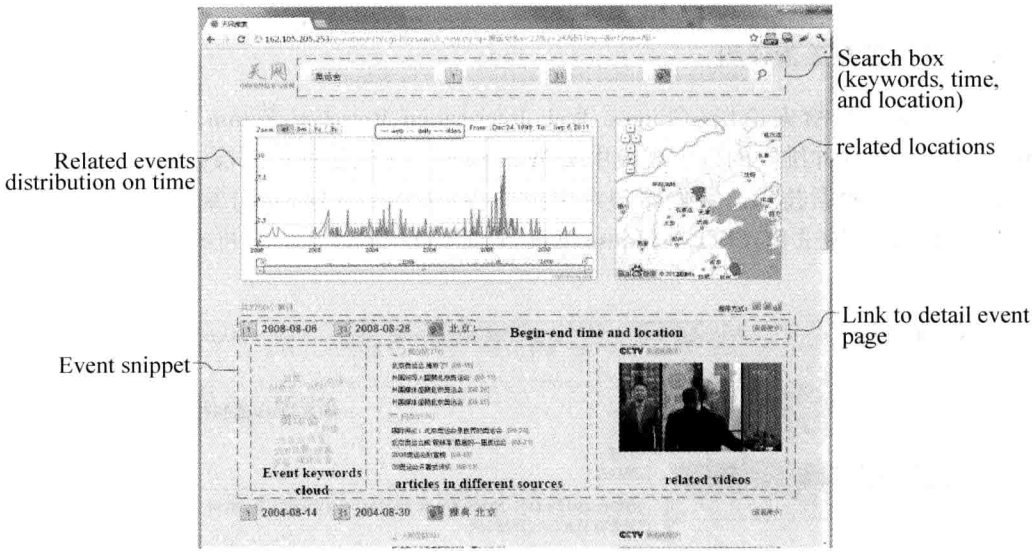


图5　事件检索系统查询结果展示页面(奥运会)

3.2　历史事件追踪

HisTrace 是一个基于 Web Infomall 的历史事件检索系统,您可以用它来回顾在多年前发生过的历史事件,寻找它们曾经在互联网上留下的印记。如图6所示。

图6　历史事件追踪系统

3.3 中文信息检索评测

中文 Web 信息检索论坛(Chinese Web Information Retrieval Forum, CWIRF)是我们从2004 年 6 月起建立并维护的以大规模中文 Web 信息为测试集的信息检索研究论坛,其目标是推动中文信息检索技术。我们希望在国内外各个研究小组的共同参与下建立并完善以中文为主的 Web 测试集 CWT(Chinese Web Test collection),一起推动中文检索技术的发展。

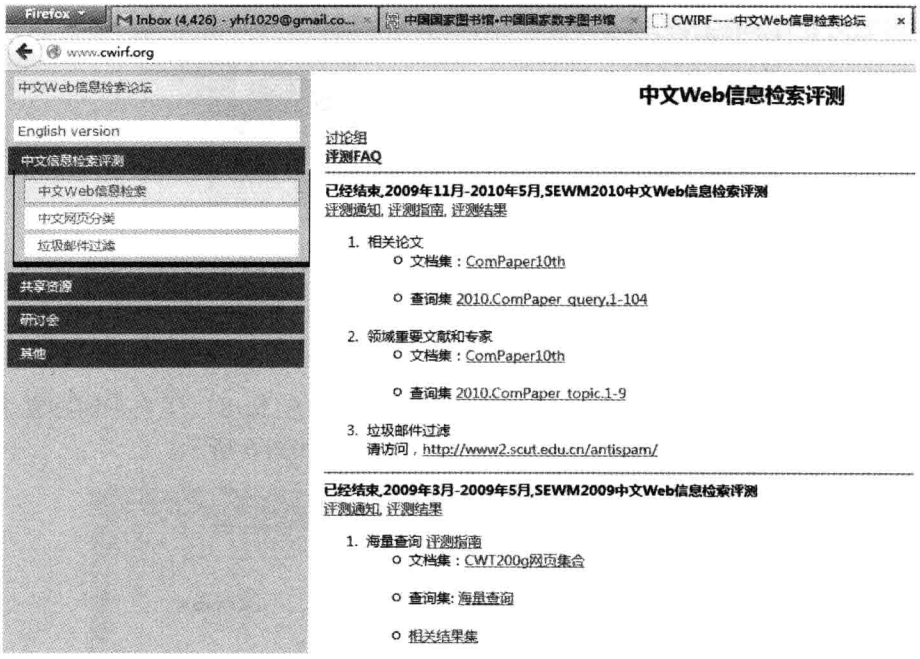


图 7 中文 Web 信息检索论坛(CWIRF)

3.4 中国互联网数字资源财富库藏(CDAL)

2003 年北京大学网络实验室启动了中国互联网数字资源财富库藏(CDAL, Chinese Digital Assets Library)的工作。包括 7.5TB 的 1.63 万资源。

这一工作的两个目的是:①收藏,不仅包括热门类别网络资源,还注重用户自己创建、组织的内容;②研究,关注网络资源从无序的原始状态到达有序组织过程中,有哪些环节可以用何种省力、省事的方式去完成。

CDAL 提供的服务主要有 3 种方式(见图 8):基于分类体系的浏览(分类树见左侧框)、基于名字关键词的检索(检索结果见中间框)和基于专题的收藏(见右侧框)。

由于每个资源有自身的内部构成,一些资源是某种作品集,可能包括关于一个事件的多个图片,一个作者的多部作品,或者一个专辑的多首歌曲,所以 7.5TB 的 1.63 万资源事实上包括了更多的单个实体。我们选出 13 个常见类别,其资源量和字节数分布如表 1 所示。