Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

# 基于Web Archive的网页重现方法及应用研究*

□ 向菁 吴振新 孙志茹 / 中国科学院国家科学图书馆 北京 100190

摘要：网页重现是利用相关技术方法、工具来恢复网页原貌。文章结合Warrick和Past Web Browser等实际案例分析了基于网络资源长期保存的网页重现技术在网站恢复、网站重建、历史页面重现应用的方法、过程、效果，为相关研究提供了参考价值。该文为2009年第七期“网络信息资源保存”专题文章之一。

关键词：网页重现，网站恢复，网站重建，网络资源长期保存
DOI：10.3772/j.issn.1673—2286.2009.07.005

## 1 引言

网络资源长期保存（Web archive，简称WA）的主要目的在于选择有价值的Web资源予以保存并提供访问。在经过十余年发展研究的基础上，目前WA领域已经存档了大量的Web资源，世界上最大的网络信息资源仓储IA[1]（Internet Archive）自1996年成立至今已保存了1500亿的网页资源。

如何利用相关技术方法、工具来恢复网页“原貌”（look and feel），为用户呈现已存档的web资源是保存资源应用中的一个重要问题。另外，由于WA的累积性保存，网页重现（Web representation）技术能真实呈现页面在不同时间点的版本，帮助用户了解网页的历史实况和发展演进过程，也因此为社会科学研究提供了大量的原始数据。

在实际保存活动中，基于WA的网页重现面临诸多的困难和挑战。由于缺少有效的网页更新、监测的管理机制来及时定位网页的演变过程，以及普通的网络爬虫无法发现深层的网络资源等原因，导致网页历史数据保存的不完整；网页超链接情景变更控制、各个时间结点管理等技术方面的局限性也是阻碍重现效果不佳的因素。

## 2 基于Web Archive网页重现发展现状

基于WA的网页重现技术就是将网络仓储中存储的网站内容以其原有的样貌展现给用户，主要应用于网站恢复、网站重建和历史页面重现等方面。

用户点击遭受意外损坏的网页，该网页（网页快照）无法呈现，此时网站服务器将用户指向WA浏览器（Web archive viewer）或再现工具（如Wayback[2]、Warrick[3]），浏览器或再现工具从WA网络仓储（web archive repository）中获取页面并将这个页面尽可能忠实地呈现，让用户感觉就像是在访问原始网站一样。这些页面与平时浏览页面在形式上虽没有区别，但在网站的恢复、重建的过程中，需要采取一定的存储格式、组织形式等技术手段对网页进行处理，将采集的网页存储在专门的网络仓储中，因此，这些页面的存储位置、后台存储格式、组织形式都有所不同。

在历史页面重现过程中涉及到的最主要问题就是超链接情景变更和各时间点网页版本的管理。出于WA存储的需要，存储过程中需要

* 本文系国家社会科学基金项目“网络信息资源保存的理论与方法研究”（项目编号：06BTQ025）的研究成果之一。

对网站原有的超链接形式和结构进行相应的改变。在历史页面重现过程中，为了保持网页原有的“look and feel”，需要恢复原有的超链接情景。另外，由于WA网络仓储会保存同一页面不同时间采集的版本，历史页面重现时需要确定网页具体的时间点，需要利用“URL+时间戳”的访问方式来定位具体的再现内容。值得注意的是，由于WA采集周期和网站更新频率的不同，可能会出现某一时间点的网页内容并没有被采集的情况，就需要回溯利用该时间点之前的内容。因此，网站重建和历史页面重现都不能准确重现某些时间点上的全部原始网页内容。

目前，搜索、浏览工具的不断改进，在很大程度上增强了网站恢复、网站重建和历史页面重现的能力。Google Sitemap Protocol[4]、Webmaster Tools[4]、Yahoo Site Explorer[5]能发现搜索引擎机器人无法发现的深层网络资源；MSN、Google、Yahoo以直接或间接的方式支持OAI-PMH协议的资源发现；在浏览器上安装Alex工具条[6]，可以显示网站的历史档案，也能让IA采集并存储web 资源。在WA研究领域，美国弗吉尼亚Old Dominion大学计算机学院利用lazy preservation的二次采集方法[7]利用网络仓储的存储进行网站恢复、网站重建；针对网页404错误采取的Just-In-Time方法[8]进行网站恢复与重建，并开发出网站恢复、网站重建的工具—Warrick；日本东京大学开发的历史页面重现浏览器——Past Web Browser[9]都是对网页重现技术在WA领域应用的探索实践，为今后的研究奠定了良好的基础。

## 3 基于Web Archive的网页重现方法

### 3.1 Just-In-Time—利用Opal框架恢复网站

美国弗吉尼亚Old Dominion大学计算机学院开发了针对404错误网页（网页不存在）进行网站恢复的应用框架，采用opal服务器重定位网页，将采集的URL存储在专门的URL数据库中，使用缓存资源中的词法签名（lexical signature）在URL数据库中检索是否存在404错误的URL，生成链接由旧版本网站指向网站的修复版本。

Opal是一个灵活、轻量级框架，可以定位404错误的类似网页。Opal服务器的配置简单，只需在定制的404错误页面安装JavaScript的标签页。如图1（左），Opal服务器的具体处理步骤是[8]：（1）服务器中未找到用户请求的URL，向客户端返回404 错误的页面；（2）定制的404错误网页和JavaScript的网页标签（pagetag）返回到客户端；（3）网页标签将用户重指向Opal服务器；（4）Opal服务器在Web仓储（如IA）或搜索引擎临时仓储（如Google、Yahoo！）中来检索网页缓存的各个版本，如果没有缓存资源的网页版本，生成词法签名（lexical signature），并在URL数据库中检索是否存在404错误的相似URL，这些相似URL按相似度由高到低排序并生成描述元数据；（5）最后，将相似的URL返回给用户，由用户选择网站重建的方式。如图1（右），Opal呈现给用户的界面，列举了404错误的URL、相似页面、缓存版本，用户可选择哪个搜索引擎临时仓储（Google、Yahoo、IA）作为网站恢复的数据来源。


Opal服务器架构　　利用Opal框架恢复 www.cs.odu.edu/_tharriso/

图1 Opal服务器架构及利用Opal框架进行网站恢复[8]

该项目发现：web资源从发现到被缓存时间为10-103天，该网页删除后在缓存中保留的时间为7-51天，由于网络仓储对流行程度高网页缓存数据较为完整，因此Opal对流行程度高的404网页恢复效果较好，但随着网页流行程度的逐步降低，风险也随之增加。

### 3.2 Lazy preservation—二次采集方法重建网站

目前进行网站、网页的保存工具较多，如InfoMonitor[10]可保存网站的文件系统并可进行远程存档；TTApache[11]用来保存由网页服务器发出请求的网页；iPROXY[12]代理服务器保存不同服务器请求的网页；但这些工具都不能用于第三方进行网站重建。

由于各种原因，一些运行网站遭受严重损坏而网站没有备份，Lazy preservation方法可以利用网络仓储采集爬虫（web repository crawler）采集相关网页进行网站的恢复和重建。因为受损站点已不存在，爬虫无法直接从web下载资源，而是利用网页仓储或搜索引擎的临时缓存库中保存的Web资源来重建网站，所以特别适合第三方使用，无须内容制造商提供数据，但也无法保证重建质量。

Lazy Preservation在网站重建过程中首先会利用爬虫根据重建网站的URL从网络仓储中下载资源，从中抽取出更多相关URL，放入需继续采集的种子站点（Seed URLs）队列中（Frontier），同时标注已访问URL列表（visited URLs），将下载资源存入Repo；然后，循环从队列中提取种子URl进行下载处理。下载过程中由网络仓储爬虫决定选择同一URL的哪一版本的资源，通常情况下是采集最新版本的缓存资源。

为提高Lazy Preservation的效率，美国弗吉尼亚Old Dominion大学计算机学院“Tools for a Preservation-Ready Web”项目[14]开发了网络仓储采集工具Warrick[3]。在实验中，设定Warrick优先抓取IA缓存资源，如果IA存档的网页快照没有MIME类型text/html，Warrick不再继续发送请求，仅保存HTML资源的标准版本；如果IA存档的网页快照有text/html的MIME类型，再使用API向Google、MSN 、Yahoo发送访问请求，比较各自快照缓存的日期，选择最新缓存的资源进行网站重建。Warrick还限制每日向网络仓储发送请求的数量，减少服务器负担，保证大型网站在一定时间间隔内进行网站的重建。

Warrick使用简单，用户只需在Warrick中输入需重建网站URL、选择网络仓储数据来源、限定网站日期等，即可开始网站重现的工作，网页重现的工作完成后Warrick会邮件通知用户。对于无法找到链接而无法重建的网站，Warrick将检查除IA、Google、MSN 、Yahoo之外的网络仓储看其是否存储；对于无法访问、域名已发生改变的网站，Warrick会在采集的网站列表中增加新的域名，用于网站重建。

该项目在Warrick开发过程中通过重现300个网站对网站格式、顶级域名、网站变化、网站年限等进行测试比较的一系列实验发现[15]：Warrick能识别77%的HTML和75%的文本资源，但仅能识别42%图像和32%的MIME类型的其他资源，HTML是网站重建成功率最高的格式，Google缓存进行网站重建的效果最好；网站的年限与重现效果之间的关系发现网站年限与网站重现成功与否没有必然的直接联系。

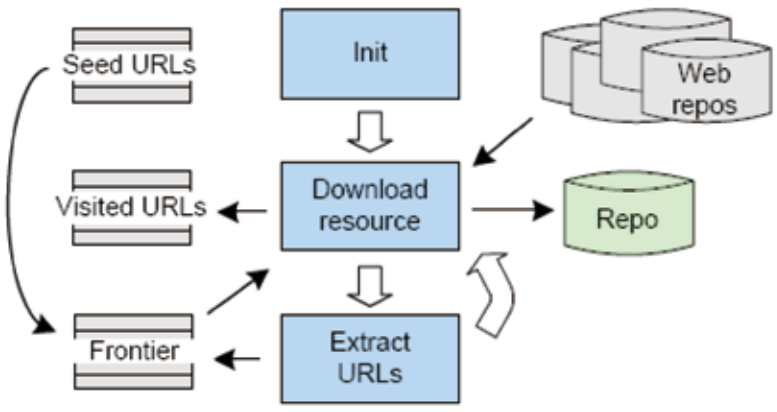


图2 Lazy Preservation采集流程[13]

### 3.3 Past Web Browser——历史网页重现

Wayback、WERA是WA领域的访问工具，对于存储量小的Web资源浏览效果较好，但在网页容量大时会由于网页快照的不完整、WA无法采集、网页文件变化的情况，浏览效果受到一定的限制。为了解决此问题，日本京都大学开发了基于WA数据抽取对历史页面重现的浏览器Past Web Browser[16]，对

历史网页进行可视化导航呈现，帮助用户更好了解网页的演变进程。此框架不限制同一时间点WA仓储的存储数量，通过呈现网页变化实现页面的自动浏览，对于网页数据量大、内容更新变化较少的页面浏览效果有极大提升。

与Warrick只采集最新版本快照不同，Past Web Browser从不同的网络仓储中（Internet Archive、Google、MSN）检索和采集同一URL的不同时间结点的网页版本，其中以IA为主要的来源，通过各个仓储的meta-archive 合作形式联合访问历史页面来采集和比较网页版本变化。

Past Web Browser首先访问网页的参考链接，获取其中提供的网页快照的时间戳信息，据此来判断内容更新与否；然后对已更新的网页进行数据准备，通过自主研发的检索结果可信度估算系统来比较各仓储的同一网页不同版本间的可信度，选择可信度最高的网页供用户浏览；并使用负载均衡的方法来保证资源不过载和数据的快速传递，从而提高历史页面重现的效率。

该浏览器作为仓储与本地系统的代理服务器中介，对下载的网页快照进行保存来解决快照保存数量少的问题；根据网页上下文语境分析网页变化解决URL经常变换但网页内容保持不变的问题；对于没有时间戳的数据（如Yahoo!快照），在没有检查此内容与其他有时间戳内容的相似性的情况下，会提前下载好网页快照。

Past Web Browser是一个可下载的客户端系统，可根据用户需要浏览网页的各种版本，拥有导航、观察网页更新变化、过滤、可视化发现网页年限等功能。导航方面该浏览器呈现离原始页面最近的时间点的页面；网页更新变化查看功能允许用户输入新的日期或点击时间线上的任意时间点，呈现网页更新变化的情况；在用户不明确网页的时间版本时，过滤功能可以加快浏览效率和检索效率，并可预测网页不同部分出现的内容类型，分析主题块内容的相似性，解决本地化浏览由于网页结构变化过大带来的浏览失败的问题。在可视化发现网页年限方面，JavaScript对各部分网页内容在红框的右下角标注释网页创建日期；对于没有创建日期的内容部分用“before t_oldest_snapshot”进行标注。浏览器页面的顶端显示已下载网页快照的时间线，在资源列表的右侧显示年限最长的页面创建日期和整个页面的平均创建日期（图3）。但当网页内容年限较长、网页快照数量较多时，该浏览器反应时间就会过长。为了解决此问题，该项目下一步准备增加动态可视化的模块，在检索过程中实现逐渐可视化注释，并计划增加用户查看网页内容特定区域年限的功能模块，方便用户查看。

该项目结合实际实验，通过比较past Web browser与Wayback在网页重现的完整程度、花费时间、Keystrokes、点击率等方面指标后发现[17]：past Web browser在各个指标方面的网页重现效果都优于Wayback。通过用户使用情况调查也显示：由于past Web browser的自动重现和便于查看网页演进变化的功能，相比Wayback而言，用户更喜欢使用past Web browser。

## 4 结语

网站恢复、网站重建以及历史页面重现的效果取决于多方面因素影响：网站外部链接、内部链接、Google PageRank、主页面跳转次数（Hops from root page）、路径深度、MIME类型、查询字符串参数、网站年限、资源创建的时间、顶级域名、网站大小、资源容量等。通过Warrick实验发现[17]：Google PageRank、跳转主页面的次数、网站年限是决定网站重建成功机率大小的最为重要的三个因素。

网页重现技术应用效果的关键


图3　网站各个部分的创建日期显示（http://www.delaware.gov/）[17]

在于网络仓储保存的网页快照的数量和质量，因此未来需要越来越多像IA那样保存网页资源各个时间点版本的网络仓储。此外，对于链接情景变更与恢复、网页变化监测、深层网资源重建等都是网页重现技术下一步研究的重点和方向。

**参考文献**

[1] Internet Archive[EB/OL]. [2009-04-09]. http://www.archive.org/index.php.
[2] Wayback [EB/OL]. [2009-04-12].http://www.archive.org/web/web.php.
[3] Warrick[EB/OL]. [2009-04-09]. http://warrick.cs.odu.edu/.
[4] Google Sitemap Protocol [EB/OL]. [2009-04-09]. https://www.google.com/webmasters/tools/docs/en/protocol.html.
[5] Yahoo Site Explorer [EB/OL]. http://siteexplorer.search.yahoo.com/ .[2009-04-09].
[6] Alex Toolbar [EB/OL]. [2009-04-09]. http://www.alexa.com/toolbar/.
[7] McCOWN F, SMITH J A, NELSON M L. Lazy Preservation: Reconstructing Websites by Crawling the Crawlers [EB/OL]. [2009-04-09]. http://www.cs.odu.edu/~fmccown/pubs/lazyp-widm06.pdf.
[8] TERRY L, HARRISON M. JustInTime Recovery of Missing Web Pages[EB/OL]. [2009-04-09]. http://www.cs.odu.edu/~mln/pubs/ht06/ht10-harrison.pdf.
[9] ADAM J, YUKIKO K, SATOSHI N. A Browser for Browsing the Past Web[EB/OL]. [2009-04-09].http://portal.acm.org/citation.cfm?id=1135777.1135923.
[10] InfoMonitor [EB/OL]. [2009-04-09]. http://www.infomonitor.pl/u235/navi/31657.
[11] Managing versions of web documents in a transaction-time web server [EB/OL]. [2009-04-09]. http://delivery.acm.org/10.1145/990000/988730/p422-dyreson.pdf?key1=988730&key2=5637990421&coll=GUIDE&dl=GUIDE&CFID=33391758&CFTOKEN=47581819.
[12] A Proxy-Based Personal Web Archiving Service[EB/OL]. [2009-04-09].
http://delivery.acm.org/10.1145/380000/371462/p61-rao.pdf?key1=371462&key2=8262701421&coll=GUIDE&dl=GUIDE&CFID=33593129&CFTOKEN=22948765.
[13] McCOWN F, BENJELLOUN A, NELSON M L. Brass: A Queueing Manager for Warrick [EB/OL]. [2009-04-09].http://www.cs.odu.edu/~mln/pubs/iwaw-07/brass-iwaw07.pdf.
[14] Tools for a Preservation-Ready Web [EB/OL]. [2009-04-09]. http://www.digitalpreservation.gov/news/events/ndiipp_meetings/ndiipp08/docs/session9_nelson.ppt.
[15] ADAM J, YUKIKO K, SATOSHI N. A Browser for Browsing the Past Web [EB/OL]. [2009-04-09].http://portal.acm.org/citation.cfm?id=1135777.1135923.
[16] ADAM J, YUKIKO K. Journey to the Past: Proposal of a Framework for Past Web Browser [EB/OL] [2009-04-09]. http://www.dl.kuis.kyoto-u.ac.jp/~adam/ht06.pdf.
[17] ADAM J, YUKIKO K, KATSUMI T. Detecting Age of Page Content[EB/OL]. [2009-04-09].
http://delivery.acm.org/10.1145/1320000/1316925/p137-jatowt.pdf?key1=1316925&key2=6368701421&coll=GUIDE&dl=GUIDE&CFID=33610283&CFTOKEN=46663952.

**作者简介**

向菁（1984-），中国科学院国家科学图书馆2007级硕士研究生，研究方向：信息检索与技术。通讯地址：北京市北四环西路33号，中国科学院国家科学图书馆615，100190。E-mail: xiangj@mail.las.ac.cn
吴振新，中国科学院国家科学图书馆副研究馆员。通讯地址：北京市北四环西路33号，中国科学院国家科学图书馆，100190。E-mail: wuzx@mail.las.ac.cn
孙志茹，中国科学院国家科学图书馆兰州分馆2006级博士生。

## Research on Web Reconstruction Based on Web Archive

Xiang Jing, Wu Zhenxin, Sun Zhiru / National Science Library, Beijing, 100190

Abstract: Web representation is to recover web page without missing its original visage by using some technologies and tools.This paper summarizes web representation technologies based on web archive , and how to apply to web recovery, web reconstruction and past web representation, wishes to provide some references for related research and practices.

Keywords: Web representation, Web recovery, Web reconstruction, Web archive

（收稿日期：2009-05-15；责任编辑：贾延霞）