国家档案局科技项目（项目编号：2011-X-26）

《网络信息资源归档与利用平台建设的研究》

研　究　报　告

辽 宁 省 档 案 局

目　录

（八）辽宁省档案信息网信息资源分析

档案网站是各级各类档案部门通过现代化技术向社会提供档案信息利用服务的主要途径和手段。从信息集合的角度来讲，档案网站是由档案部门建立的一个信息集合，此信息集合全部是以数字代码的形式存在，并按照不同的分类、排序组织起来，以各种美工方式进行设计从而搭建起来的档案信息资源集合。它包含有文字、数字、图表、图形、图片、声音、动画、音像等所有多媒体信息。这就好比是高速公路上来往的车辆所载的货物，货物数量越多、种类越多越能体现高速公路的利用价值。

档案网站是众多网站的一种，是向社会提供档案信息查询服务的电子信息集合。能够反映档案部门政务公开、档案工作动态以及馆藏档案信息的内容。是各级档案局（馆）、档案学会、协会、学院等档案机构、组织为介绍档案机构、档案馆藏资源、提供档案信息服务、促进专业信息交流而在互联网上建立的专业网站。

档案网站中的信息资源记录了不同时期档案部门的机构概况、重大活动、业务工作、通知公告等重要信息，一些信息是网站上独有的，不同于纸质档案和电子文件，是通过各种档案资源优化重组而成的，（如网上档案展览动画）。可以说不同时期的档案网站都记录与见证了档案事业的发展历史和进程。这些资源弥足珍贵，进行归档十分必要。下面

以辽宁省档案信息网为例，分析档案网站的一般框架及信息资源构成。

一级栏目	二级栏目	三级栏目
局馆介绍	基本职能	
	部门设置	
	主要领导	
	岗位职责	
	馆区平面图	
工作动态	局（馆）动态	
	省内新闻	
	国内新闻	
	国际新闻	
政策法规	法律法规	
	业务标准	
	规章制度	
	法律咨询	
机关工作	业务指导	业务咨询
		业务考核
		经验交流
	科研管理	科研工作
		立项与评奖
		科研成果
	教育培训	教育计划
		培训纲要
		岗位培训
		网上课堂
	人事管理	政策标准
		职称评定
		专业考核
信息化建设	信息化标准体系	
	信息化成果	
档案管理	档案接收	
	网上征集	
	库房管理	
	档案保护	
公共服务	利用查询	档案查询
		现行文件查询
	档案馆指南	
	馆藏简介	
	查档流程	

	预约调卷	
	论坛或留言板	
档案展示	档案博览	
	珍档荟萃	
	网上展览	
	名人介绍	
	城市变迁	
	历史掌故	
档案文库	档案著作	
	编研成果	
	期刊查阅	
	优秀文摘	
	论文评选	

表 3　档案信息网框架构成

可以从上表中看到，档案网站上信息资源丰富，有经过一定组织加工馆藏数字化档案信息，有通过有效组织编排的结构化和非结构化信息（如网上展厅、珍档荟萃、档案博览等），其中档案许多都是档案部门工作的重点，包括：现行文件、特色档案、编研成果、信息化建设、人事管理等信息资源。

八、目前已解决的技术难题

（一）解决了海量文件的存储和提取的问题

归档网络资源的有效存储，即将爬虫采集下来的各类网络资源文件以符合国际标准的 WARC 格式保存，解决了海量文件的存储和提取的问题。并且存储格式符合国际标准，有利于系统的扩展和归档文件的与其他系统共享与利用。

ISO 28500:2009《信息与文档-WARC 格式》所定义的格式（即 WebARChive），WARC 格式是在 ARC 格式的基础上发展

起来的。从1966年推出以来，WARC一直被国际上许多档案保存机构用于“网络爬虫”来抓取网络信息资源。可以把多种数据对象整合于一个大文档中，而这种格式可用于信息内容的采集、管理、存取和交换等多种应用，能够存储与其他已存数据相关联的任意元数据（如主题、分类，编码等）。WARC格式保障了这些海量信息得以实现有效的管理、机构化和存储，并且能够脱离采集平台，采用autoCAD、unARC等软件可以直接打开，从而保证了信息的长期保存。

通过网络爬虫程序，将指定网站网页、图片等资源文件抓取下来，以WARC文档格式保存存于本地的文件系统。通过网站回放程序，使用户可以通过普通的浏览器，浏览查看保存于WARC文档中的过去的网站内容。下图显示了一个典型的WARC 文件结构。

图12 WARC文件结构

WARC 文件可以将多个 Web 文档文件的原始内容保持于一个大的文件中。一个 WARC 文件由若干个存档记录构成。每一存档记录包含了一个 Web 文档的所有归档信息，其中包括元数据、网页请求头、网页应答头，和网页应答体。

元数据用于保存与这个 Web 文档相关的各种元数据信息，如归档人姓名、部门等。

网页请求头是客户端程序（如浏览器和采集器）向 Web 服务器请求这个 Web 文档时所发送的请求信息，其中包括这个 Web 文档的 URL 地址。下图为一个典型的网页请求头的内容。

Request Header	Value
(Request-Line)	GET /lnsdaj/api/find.jsp HTTP/1.1
Host	www.lndangan.gov.cn
User-Agent	Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US; rv:1.9.2.8) Gecko/20100722 Firefox/3.6....
Accept	text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language	en-us,en;q=0.5
Accept-Encoding	gzip,deflate
Accept-Charset	ISO-8859-1,utf-8;q=0.7,*;q=0.7
Keep-Alive	115
Connection	keep-alive
Referer	http://www.lndangan.gov.cn/lnsdaj/xwzx/gzyw/content/ff8080812b31fbb3012be7adfb9e...
Cookie	Hm_lvt_1abcba225075217797f05ad81bdb802e=1288608474770; Hm_lpvt_1abcba22507521...

图 13 网页请求头的内容

网页应答头为 Web 服务器收到网页请求后，像客户端请求程序发送的应答内容的头部信息。包括应答产生时间、Web 文档最后修改时间、Web 文档大小等信息。下图为一个典型的网页应答头的内容。

Response Header	Value
(Status-Line)	HTTP/1.1 200 OK
Date	Mon, 01 Nov 2010 10:51:14 GMT
Content-Length	2136
Content Type	text/html; charset=utf 8
X-Powered-By	Servlet/2.4 JSP/2.0
Keep-Alive	timeout=5, max=98
Connection	Keep-Alive

图 14 网页应答头的内容

网页应答体为Web服务器收到网页请求后，像客户端请求程序发送的应答内容的正文信息。例如对于一个HTML网页来说，网页应答体就是这个网页的HTML源码。对于一个图片文件来说，这个应答体就是这个图片文件的二进制数据。下图为一个典型的HTML网页应答体的内容。

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<base href="http://www.lndangan.gov.cn/lnsdaj/">
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<title>辽宁省档案信息网</title>
<link href="css/style.css" rel="stylesheet" type="text/css" />
<script>
function test_null(){
var submit_value = "0";
            var key = search.key.value;
           if(search.key.value==""){

                alert("查询关键字不能为空...");

                search.key.focus();

                return false;

           }

           if(search.key.value=="请输入关键字"){

                alert("请填写您需要查询的关键字...");

                search.key.focus();
```

图 15 HTML 网页应答体的内容

每一次网络信息采集任务会产生一组WARC文件，每个WARC文件可以包含几千至几万个Web存档记录，每个Web存档记录会保存一个Web文档的完整的信息。每一次网络信息采集任务所产生的一组WARC文件存储在一个单独的目录中。这个目录按如下规则创建：

...\webarchives\[站点网址]\[采集年份]\[采集日期]

这样，网站回放程序可以根据要回放的网站地址和回放日期，快速找到所对应的Web存档文件。

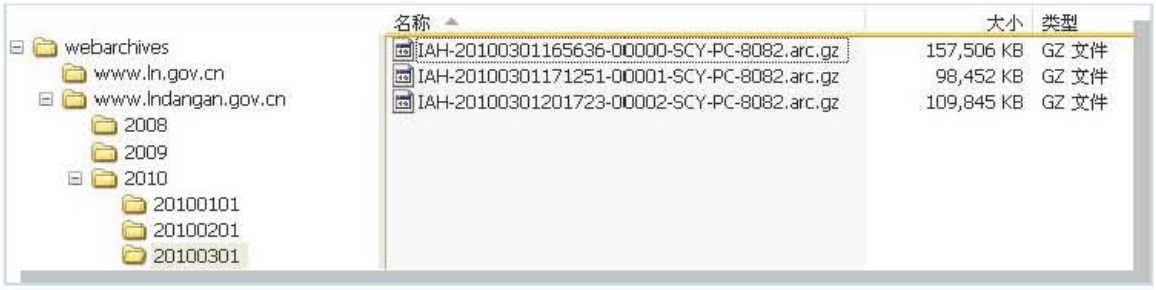

图 16 Web 存档文件

（二）解决了对网络资源进行增量采集的问题

所谓增量式 Web 信息采集，即是在执行采集任务时，对上次采集任务执行完成后，没有发生变化的 Web 文档不再采集，而只对目标网站上发生变化或新增的文档进行采集。增量式 Web 信息采集有两个目的：一是节省 Web 存档文件所占用的磁盘空间；二是节省后续采集任务的执行时间。

Web 信息采集器第一次采集某个网站信息时，会将这个网站上所有的 Web 文档都采集下来，保存在 Web 存档文件中。在后续的采集的任务中，采集器会对在该网站上的发现 Web 文档与上次采集下来的 Web 文档存档进行比较，如果上次没有采集或内容发生了变化，就直接将 Web 文档写入 Web 归档文件。如果现在在线的 Web 文档与 Web 存档是完全一样的，就只在 Web 存档文件的存档记录中写入一个指针描述，说明这个 Web 文档的完整内容可以在上一次采集任务生成的 Web 存档文件中找到，而不再将这个 Web 文档的完整内容重新写入这次采集任务所生成的 Web 存档文件中。

例如：对于一个 Web 文档（比如百度上的一张图片 http://www.baidu.com/pic.jpg），分别在 9 月 28 日和 10

月 6 日进行了两次修改。如果采集器分别在 10 月 1 日、10 月 3 日、10 月 5 日和 10 月 7 日对这个 Web 文档进行了采集，那么在 10 月 1 日采集的 Web 存档文件中，保存的就是在 9 月 28 日修改后的这个 Web 文档的完整内容。在 10 月 3 日和 10 月 5 日采集的 Web 存档文件中，保存的是指向 10 月 1 日采集的 Web 存档文件的指针描述。而在 10 月 7 日采集的 Web 存档文件中，保存的是 10 月 6 日这个 Web 文档修改后的完整内容。由于在 10 月 3 日和 10 月 5 日采集的 Web 存档文件中没有保存 Web 文档的完整内容，而只是保存了指针描述，因此会节省大量的磁盘的储存空间。

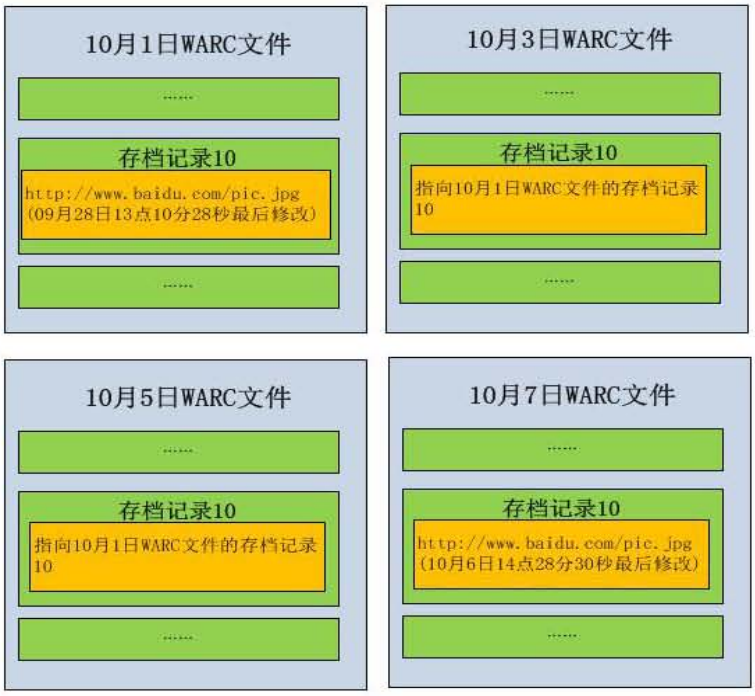

图 17 增量采集示意图

当网站回放程序回放这个 Web 文档时，回放程序首先根据网站网址和回放时间找到对应的 Web 存文件。然后根据 Web 文档的 URL 地址，找到对应的存档记录。如果存档记录中保

存的是该 Web 文档的完整信息，回放程序只需将 Web 文档的内容从存档记录中读出，返还给浏览器就可以了。如果存档记录中保存是一个指针描述，回放程序将根据指针描述，在指针所指向的 Web 存档文件中找到该 Web 文档的完整内容，再返还给浏览器。

除了节省 Web 归档文件所占用的磁盘空间，增量式采集的另一个目标是是节省后续采集任务的执行时间。这个目标是通过 HTTP/1.1 协议提供的 HTTP 缓存机制来实现的。其基本方法是：要采集一个 Web 文档时，先在最近的一个 Web 存档文件中找到这个 Web 文档上次采集时保留的 Web 存档将记录。然后查看 Web 存档记录的应答头中是否包含 Etag 验证器和最后修改日期验证器。

如果包含，将这两个验证器加入到网页请求，发送到 Web 服务器。如果 Web 服务器返回代码 304（Not modified），说明 Web 服务器上的文档版本与上次采集的 Web 文档版本相同，没有发生变化。采集器直接在 Web 存档记录中写入指针描述，而非完整内容。这种情况下，由于 Web 服务器只返回代码 304，而没有返回 Web 文档的完整内容，这将极大地节省了传输 Web 文档所需的带宽，从而缩短了采集时间。

对于没有验证器的网页，采集器将先下载在线的 Web 文档，然后与 Web 存档进行比较，判断采集的 Web 文档是否发生了变化。

关于验证器：

当原始Web服务器产生完整的网页响应时，它在网页应答头上附加了一些验证器。这些验证器会被采集器保存到Web存档文件中。当采集器重新采集这个Web文档时，它将Web存档文件中保存的验证器附加的网页请求头中。然后服务器会将Web请求中的验证器与现在在线的Web文档中的验证器进行比较。如果匹配，Web服务器只返回304代码，而不返回整个Web文档内容。如果不匹配，则返回Web文档的完整内容。

下图为验证Web文档是否发生变化的流程图：

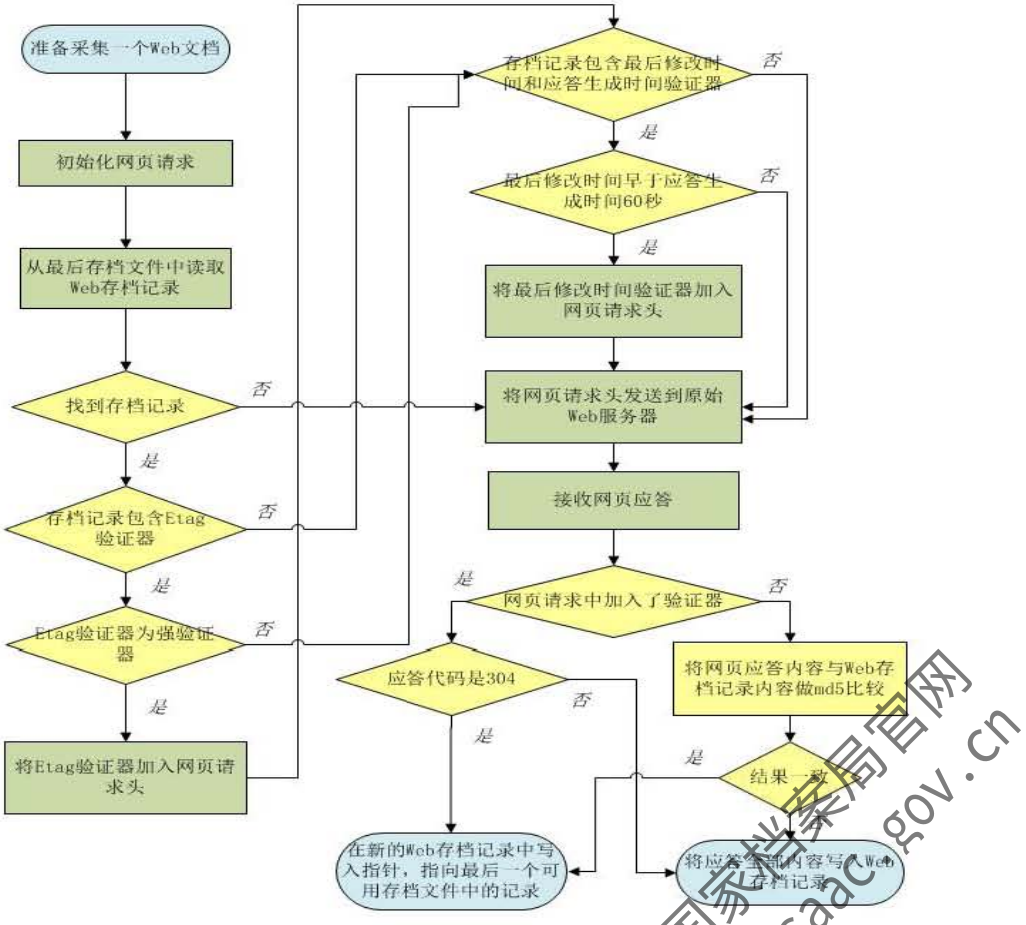

图18 验证器验证网页发生变化示意图

（三）解决了容错采集的问题

即当原始网页中有一错误时(如在URL中使用中文符号，文档头标明的文件编码与文件实际使用的编码不符等)，系统依然可以将在线网络资源文件按原样采集下来。

由于目前网页实现技术越来越复杂和多样化，对网页（Javascript、flash等）的解析能力的就决定了网页爬虫程序的优劣。我们开发的网络信息资源归档平台采用的网页爬虫程序除了具有一般传统爬虫程序的能力外，还增加了很多网页容错功能，如对URL中使用中文字符的情况，对网页头部编码标记与网页实际使用的编码不符的情况的处理，使得我们的爬虫程序可以采集一些一般传统爬虫不能采集的含有错误的网页。

（四）解决了归档网络资源按归档时间点进行回放的问题

即系统可以利用存储于Web存档文件中的Web资源文件，让用户像浏览当前网站一样，浏览查看网站的旧貌，即网站以前发布的网页。

网站回放程序分为上下两部分，上部为回放控制面板。下部为网站内容回放区。用户可以先在上部的回放控制面板中选择要回放的网站和要回放的时间点。然后单击打开站点按钮。这时，该网站在选中的回放时间点的内容就会在下部的网站内容回放区显示出来。用户可以在网站内容回放区，

通过点击页面链接，来进一步查看该网站在该时间点的其他页面的内容。下图为目前存档网站回放的界面截图。

图 19 存档网站发布利用界面

存档时间导航页面列出了该网站所有的 Web 存档文件的采集时间。单击某个采集时间，系统在网站回放系统中打开这个时间采集的 Web 存档文件，使用户可以像浏览在线网站那样，浏览这个采集时间点的网站的旧貌。

在网站回放系统中，用户可以点击链接，查看更深层次的网页。也可以拖拽上方的时间轴控件，调整回放网站的时间点。

（五）实现了对归档文件建立索引并实现高效的全文搜索的功能

在 Web 存档导航页面，用户可以通过拼音索引，查看系统中已发布的存档网站。也可以在网址文本框中输入想查看

的网站的网址，然后单击“回放网站”按钮。系统进入存档网站的存档时间导航页。下图为检索结果界面和 Web 存档导航页面的用户界面截图：

图 20　Web 存档导航页面

除了以网站回放模式查看 Web 文档存档外，用户也可以直接在 Web 存档导航页面右上角的全文搜索框中输入自己感兴趣的关键词，然后单击搜索按钮。系统会将单个站点或不同站点之间，将所有包含此关键词的 Web 文档，返回给用户。

图 21　搜索结果界面

九、未来平台建设努力方向及展望

（一）网络信息资源归档平台相关功能的完善和扩展

本课题研究在国内相关领域中属前沿研究，在对网上信息资源的归档研究过程中解决了很多技术难题。但在网络信息采集工作方面，仍然存在一些技术难题未能完美解决，这些技术上的局限性是整个网络特性、信息资源特点决定的，是在线进行网络信息归档普遍存在的难题。主要有以下几点：

1.归档平台通用性问题

由于不同网站配置不同，平台建设针对性较强，缺少普适性，不同站点、不同栏目需要人工参与的部分比较多，这一点的改善需要在下一步的平台完善阶段进行，主要在策略设计、功能模块的智能化程度上继续深化研究。

2.特殊网站或栏目采集完整度问题

对于一些网络资源，需要输入检索词，才能获得资源列表。对于这中情况，系统无法采集。这是因为，爬虫程序无法穷举所有可能的搜索条件，这样爬虫就无法采集到全部网络资源。

对于在 flash 文件中使用绝对 URL 的网络资源，这类网络资源多是一些网站采用 flash 制作的导航菜单。如果导航菜单在制作工程中，菜单使用绝对 URL 指向目标网页，在归

档文件回放系统中，浏览器会使用 flash 菜单中的绝对URL去打开当前的在线网络资源，而不去打开网络资源归档系统中的归档资源。

对于一些网络资源列表，使用服务器端Session来实现翻页机制，这样用户在点击“下一页”、“上一页”链接时，浏览器中的URL始终不变。由于爬虫只能获得浏览器端的状态信息，而无法获得服务器端保存的状态信息，因此爬虫无法实现翻页采集，即爬虫只能采集网络资源列表第一页的内容。

一些链接的URL通过Javascript经过复杂地变换后生成。由于目前爬虫还不能模拟javascript运行，因此爬虫无法准确地解析出这些URL，从而这些URL所对应的网络资源就不会被爬虫程序所采集。

大量使用AJAX技术的网站，由于目前的爬虫还没有模拟javascript运行的能力，所以对大量使用AJAX技术的网站无法采集。

需要用户登录验证码，全自动采集有困难。很多通过政府网站发布的信息都属于隐蔽网资源,如动态网页、实时数据、网络数据库及只对注册用户开放的网页等，据2001年Bergman 的一份报告显示，有超过8.5% 的隐蔽网存在于政府网站之中。隐蔽网资源是目前的全面自动保存工具无法搜集的,而人工选择性保存尽管可以搜集到这类信息且更能保

证信息质量，但与全面自动保存相比，它会耗费大量的人力财力及时间，且搜集到的信息数量有限，无法及时跟上政府信息更新的步伐。

3. 多网站采集和更新时的带宽问题

网络资源归档是在互联网上实时进行的，采集和更新的网站较多时，需要占用很多带宽，采集过程中会对整个网络的访问速度有一定的影响，采集平台需要部署在较高端的服务器上，否则会影响采集速度。也会影响维护与查询利用的速度，因此在下一阶段的平台和业务扩展上，需要在硬件上的增加投入。

4. 知识产权和隐私权的保护问题

大部分由政府网站发布的信息，如政府文件、政策、法律法规等都具有公共性，公众一般可以免费获取，在对它们加以保存时往往没有知识产权的限制。但是也有一些通过网络发布的政府信息是受到知识产权保护的，如由政府与研究机构或企业合作或委托第三方生成的技术报告、统计信息等。在开展保存项目时，如何合理、清晰地界定受知识产权保护的信息范围，对于有知识产权限制的信息，首先需征得知识产权所有人同意再保存，还是保存以后再告知他们或采取其他方式，目前还没有一个广泛而统一的意见。

各国著作权法基本都不支持在未获得版权所有者许可的情况下对资料的免费存取。按照我国《著作权法》第二章

第一节第十条的相应规定，著作权人拥有发表权、发行权、广播权、网络传播权。2006年颁布的《信息网络传播权保护条例》第七条规定：“图书馆、档案馆、纪念馆、博物馆、美术馆等可以不经著作权人许可，通过信息网络向本馆馆舍内服务对象提供本馆收藏的合法出版的数字作品和依法为陈列或者保存版本的需要以数字化形态复制的作品，不向其支付报酬，但不得直接或间接获得经济利益。当事人另有约定除外”。

可以看出，通过网络向公众提供作品的权利属于个人，从这个角度来说，网络信息资源保存机构在没有授权的情况下，没有权利将收集保存的网络信息向用户提供使用。

综上，但这些技术局限性并非不可避免的，不是每个网站必然会出现的问题，并且可以通过建立相关标准、规范来辅助解决。因此，在下一步的研究中我们将在此研究基础上，继续深化和扩展。如：通过技术手段，尽可能解决目前网络信息采集平台不能处理的问题。并且在归档网络信息资源时对如何协调通过建立标准、规范，解决技术上存在的难题，以达到经济、高效与优质的统一方面，将进一步研究。

（二）网络归档资源管理相关制度规范的深化研究

管理与规范是开展网络资源归档与利用等工作所必须遵循或参照的一系列准则，目前国内外可参照的相关标准规范很少，还需在实践中探讨和总结，如《网络信息归档与利

用操作规程》、《网站建站技术规范》、《网络信息归档信息管理规范》等。为此项工作提供科学和理性的指导。同时可以解决建设和管理过程中出现技术难题和理论难题，从而便统筹安排资源，节约成本。

（三）网络归档资源管理与知识管理理念的深度融合

网络信息资源整体无序化，内容组织程度也不高，信息资源之间交叉关联程度较低，用户需要在不同的网络环境之间穿梭漫游，需要在不同的信息空间来回切换，需要掌握不同检索软件的使用方法，用户在信息空间中的“迷航”会使他们感到厌倦而丧失获取信息的信心，从某种意义上说，信息资源数量越大，给用户造成的负担也就越重。

所以，我们必须重新审视、研究网络环境下信息资源的新情况，并对其进行积极有效的整合，只有这样才能真正实现信息资源的有序化，实现信息资源共享效用的最优化，否则，必然会使用户陷入不得门径而入的困惑境地，影响信息资源的有效利用。

基于此我们构想建设基于知识管理的网络归档资源库，实现档案信息资源的增值，提高档案馆在不断变化的数字环境下的应变能力，也是档案界有效应对知识经济时代未来挑战的必然选择。

1.知识管理理念

知识是具有不同形态的，根据1996年经济合作与发展组织(OECD)的《以知识为基础的经济》报告，知识包括四大类：(1)“知道是什么(know-what)”的事实知识；(2)知道为什么(know-why)”的原理知识；(3)“知道怎样做(know-how)”的技能知识；(4)“知道是谁(know-who)”的人际知识。OECD报告将前两类知识归为显性知识,将后两类知识归为隐性知识。显性知识是经过编码的、有序地承载于某种可见载体之上的知识，包括传统的书面化的文件或电子化后的档案。这类知识是客观的、理性的知识，可以在不同个人之间快速而简单地传递，便于组织成员之间的沟通与分享。隐性知识是未经正式化的知识，包括个人的思维（心智）模式、主观信仰和观点，是基于个人经验与直觉的知识，不易用语言来沟通与表达。它是知识创新中最为基础的东西，具有很强的抽象性、主观性和个性特征，这类知识很难被交流和传递。

知识与信息是两种不同性质的概念。知识属于人类认识成果范畴，信息是事物属性的反映。从形式上看，知识具有最小的构成单位“知识单元”，而信息没有这种普遍意义上的“知识单元”；从基本作用上看，知识价值主要体现在能够在一定程度上正确地指导实践，信息的基本作用是作为认识的媒介，可以使主体对客体有所了解或从信息中获得某种感受；信息是正确决策的依据和知识生成、进步的原料，而

知识是有效地认识、选择、整理、加工、传播、开发的利用信息，使信息充分发挥作用的前提。

可将“知识管理”理念归纳为以下几点：一是对记录有知识的载体的管理（如，文献实体的排架和保护管理等）；二是对知识信息或有关知识的信息的管理（如，档案文献编研、将分散的知识整合成知识地图等）；三是指知识实践活动（如，知识的识别、学习和创新活动等）；四是指对知识实践活动的管理（如，对科研活动和知识产权的管理等）。在不同使用场合，“知识管理”可以表示其中一种或几种含义。简而言之，对知识或与之直接密切相关的知识实践的有关方面进行的管理，是古今中外知识管理不可缺少的基本特征。知识管理简明扼要地可概括为：以知识为核心的管理。

2.将知识管理理念融入网络归档资源管理可行性

知识管理涉及的技术工具主要有：内部网(Internet)和外联网(Extranet)、数据库管理系统(DataBase Management System， DBMS)、推拉技术(Push&Pull Technologies)、存储结构技术(Storage Architectures)、元数据技术(Metadata)、群件技术(Groupware)、数据仓库(Data Warehouse)、数据挖掘技术(DataMining)、信息查询与检索引擎技术(information Search and Retrieval Engines)、工作流技术(Work Flow)、联机分析处理技术(Online Analytical Processing， OLAP)、中间件技术(Middleware)、

多维度分析技术(Multidimensional Analysis)、文档管理技术(Document Management)、过程建模与思想地图(Process Modeling & Mind Mapping)、评估和报告(Measurement & Reporting)等。

知识管理技术工具分类如下：

分类标准	类别	包含的信息技术
按知识类别的不同，所需要的技术进行划分	管理显性知识的技术	数据库管理、文档管理系统、检索信息技术、互联网、检索引擎、内部网、工作流技术、使用和应用知识的技术、决策支持系统、业务支持系统、数据挖掘、数据仓库。
	管理隐性知识的技术	电子邮件、电视会议、协作的电子工作平台、电子式论坛，群体工作系统、技能知识分布系统。
按知识来源的不同，所需要的技术进行划分	知识是从信息系统中转换得来	数据仓库、数据挖掘、数据库、数据发现、地理信息系统、数据可视化技术、在线分析处理引擎等技术
	知识是独立生成的	文件管理技术、搜索技术、合作技术和知识库技术
从知识的生命期来看	知识的生成	Internet 、KDD（knowledse discovery in databases，中文译名“知识发现”)。知识综合的工具有 IdeaFisher（能够将相关的词句组合起来，帮助人们将分散的创新观点整合起来）和 Inspiration（能够帮助用户形成一种概念图，从而提高使用者对知识进行合成的能力)。知识创造的工具有 IdeaGenerator 和 Mindlink（这两种工具均通过引导人们突破思维定势来提高创新能力）等
	知识编码化	知识库、知识仓库、知识地图、组织词表/词典等。

	知识转移	模拟器、关系地图；降低时间距离的工具，如电子邮件；缩短空间距离的工具，如网络会议；缩短社会距离知识的工具，如学习地图等。
按知识管理的支撑技术进行划分	商业职能	由数据、文本挖掘技术、联机分析处理技术(OLAP)以及数据仓库技术组成。
	协作技术	包括实时协作技术和异步协作技术。
	知识传播	包括基于计算机的培训技术(CBT)、分布式学习技术(e-learning)以及实施电子课堂、电子研讨会和讨论会技术。
	知识发现	包括检索技术、内容分类技术以及数据导航(知识地图)和文档管理技术等。
	专家定位	包括专家网络、可视通信、密切度度量以及其他帮助迅速定位目标人群的技术。

表 4　知识管理技术工具

根据知识管理理念及现有计算机技术分析，当今计算机软件的智能化，完全可以采用技术手段对网络归档的信息资源进行综合分析、合理决策，把对信息的管理提升到对知识的挖掘、对知识的管理和对知识的利用的高度。现在大多数档案馆只把计算机当作存储设备来用，没有充分发挥其现代化、智能化的功能。数字化的对象只是停留在馆藏资源上，而与管理及业务工作相关知识的数字化未给予足够的重视。

3. 网络归档资源管理与知识管理理念进行融合的意义

运用知识管理理念是对网络归档信息资源管理的升华。知识管理是将可获得的海量信息转化和升华为知识，并且使得知识被人们所应用的过程。知识管理又是对归档信息管理

的进一步发展。第一代信息化管理的是数据；第二代信息化管理的是信息；而知识管理将信息化推进到第三阶段，第三代信息化管理的对象是知识。从技术角度来说，知识管理是对知识进行程序化管理，以便于知识的进一步明晰、提取和重复应用。而知识管理的基本流程应当包括对知识的采集、识别、获取、开发、整合、存储和使用等等。此外，根据知识管理的特性，在知识管理中必须要把握并且切实满足对知识管理的积累、交流以及共享的三种需求的基本原则。

通过知识管理模式管理网络归档的信息资源可以在原来的信息管理的分散-集中，无序-有序的基础上，使信息能够达到有机的结构化。经过一系列的筛选、鉴别、归类、综合，使得知识资源更加结构化、综合化、系统化，使得人们更加方便地、快捷地应用知识资源。因此，从这个意义上说，知识管理可以促进信息资源得到最大程度的开发利用。具体体现在以下三点：

（1）增强信息资源体系的关联性

信息资源是一个紧密联系的有机整体，而现有数据资源系统内的数据对象大都是孤立存在的，无法体现信息资源体系的内在联系。而经过整合后的信息资源，包括不同的学科、不同载体类型、不同数据库之间的知识、信息，其覆盖面广泛，能够促进各种类型间的信息资源的相互渗透、相互作用，保持信息资源体系的整体性和关联性，发挥信息资源的整体

功能，为用户提供系统的服务，更好地满足用户需求。

（2）消除用户使用过程中的差异性和复杂性

由于不同的数字资源系统有着不同的编码结构和表达方式，数据格式的不同导致描述和组织标准的差异，导致检索途径和方法的不同，另外，不同的数据库使用不同的检索软件，使得数据库检索界面也风格各异，迥然不同，用户在使用时会感觉到明显的差异性和复杂性。而整合后的信息资源系统，用户可以通过一个统一的用户界面，在多个网络数据库平台上实现检索操作，用户感觉如同在一种信息资源系统中操作，不需要了解、掌握各数据库系统的结构及使用方法等。

（3）实现最大程度的资源共享

网络信息资源共享不仅是互联网的重要特征，也是网络环境下用户对网络信息服务的要求。互联网的出现使网络信息资源的分布、利用没有了时间、空间、地域的限制。 但是用户在使用这些信息资源时仍然会出现一些问题。比如使用数据库时，用户只能一个一个的检索，若在这个数据库中找不到所需要的信息，则退出来，再进入另一个数据库进行检索。这种未经整合的数据库信息资源不能进行统一的检索，给用户查找信息带来很大的麻烦。如何实现各种类型的异质信息资源的统一检索，最大范围、最大深度的共享信息资源，便是信息服务机构进行资源整合的重要目的之一。

4.基于知识管理理念下归档平台建设任务

（1）建立大规模不同类型的网络资源归档库

建立信息资源库是开发利用的前提，初步设想是在采集辽宁档案信息网信息资源的基础上，将采集范围扩展至全省档案信息网，进而建立全省档案信息资源库。在技术与理论成熟的环境下，继续将采集范围扩展到其他领域，建立其他类型的信息资源库，如：教育信息资源、企业信息资源、文献信息资源、政务信息资源库等。

（2）针对归档资源库进行信息资源整合，建立信息服务的基础和平台。

整合对象

信息资源系统是一个相互关联的有机知识整体，它包括不同学科、不同载体类型的各种信息资源。所以，整合后的信息资源可能来自不同的学科，不同的信息资源体系，其知识覆盖面更为广泛，能够为用户提供系统的学科知识，同时，信息间具有统一性和有机关联性，可以为用户提供一体式的信息服务，满足用户全方位、多渠道地获取信息的要求。

信息资源具体的整合对象可分为：不同载体、不同类型的信息资源之间的整合

平台功能

该平台能够提供基于归档资源内容的多个个性化专题，并具有文献自动分类功能。提供基于文献内容的自动摘要和

关键词生成。提供基于内容的文献自动消重功能。提供基于文章内容的相关文章列表功能。提供基于内容相似度聚类的信息自动推荐功能。例如：在辽宁档案信息资源库内分析现行文件资源，将我省各市的档案信息网上发布的现行文件聚集，形成我省现行文件资源归档库与共享库。并能够方便的信息检索，结果集排序、动态分组等功能。

网络环境下，信息服务机构在信息社会中所起的核心作用就是服务，服务人员只有在信息资源整合的基础上才能为用户提供准确、有效的信息，同样信息检索也是基于信息在一定程度上的开发和整合。可以说，有效地开发电子信息资源是信息服务机构在网络环境下进行信息服务的基础和平台。信息服务机构要运用科学的手段和方法，充分整合各种不同类型间的信息资源，从而全面地满足社会的信息需求。

（3）建立智能分析模块，能够智能导航、推荐浏览

该模块能够提供按分类的当前热点文章自动推荐，每篇文章自动提供相关文章列表；能够按分类、地域、时间等特征的统计图表。提供Web Service调用，便于集成系统；按信息源类型进行频道组织。垃圾链接的过滤，有效链接和内容清洗；附件内容的下载、抽取，索引。采集日志存储、统计、分析等。

（4）建立多维导航、概念检索引擎

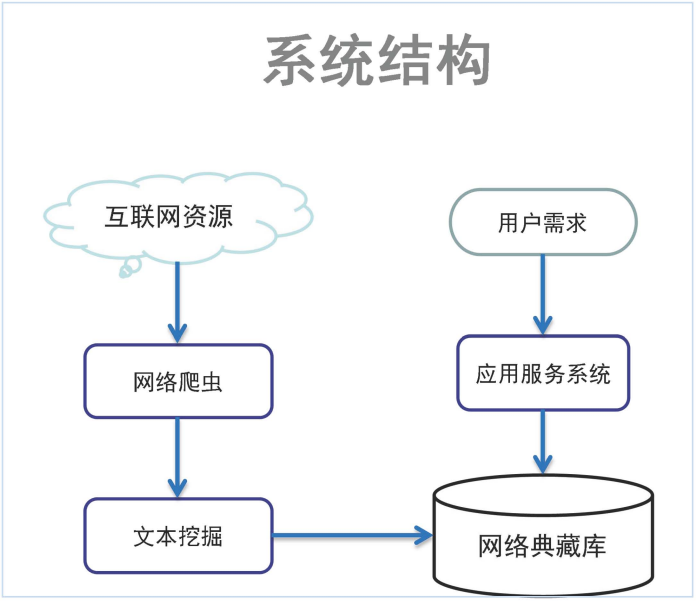

图 22 检索流程示意图

对归档网络信息资源进行深度知识挖掘，能够显示信息之间的内部联系；能都实现全文检索数据库，中文自然语言处理和分析，能够通过叙词表（受控语言）精确检索信息、并能够对相关信息或知识点进行标引等功能。

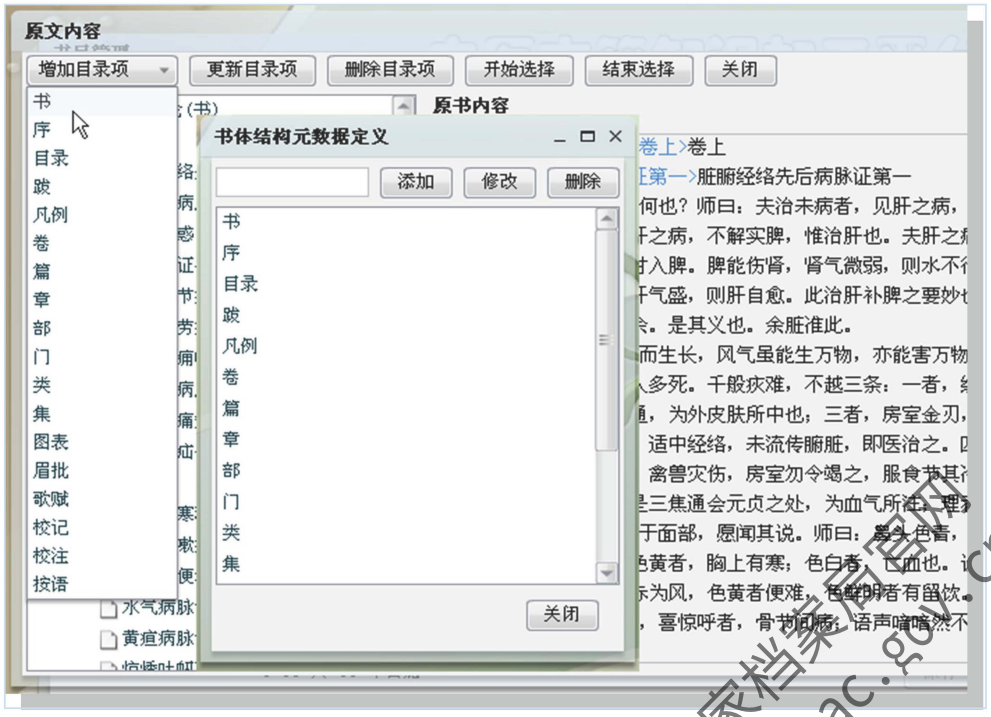

图 23 叙词表精确检索信息

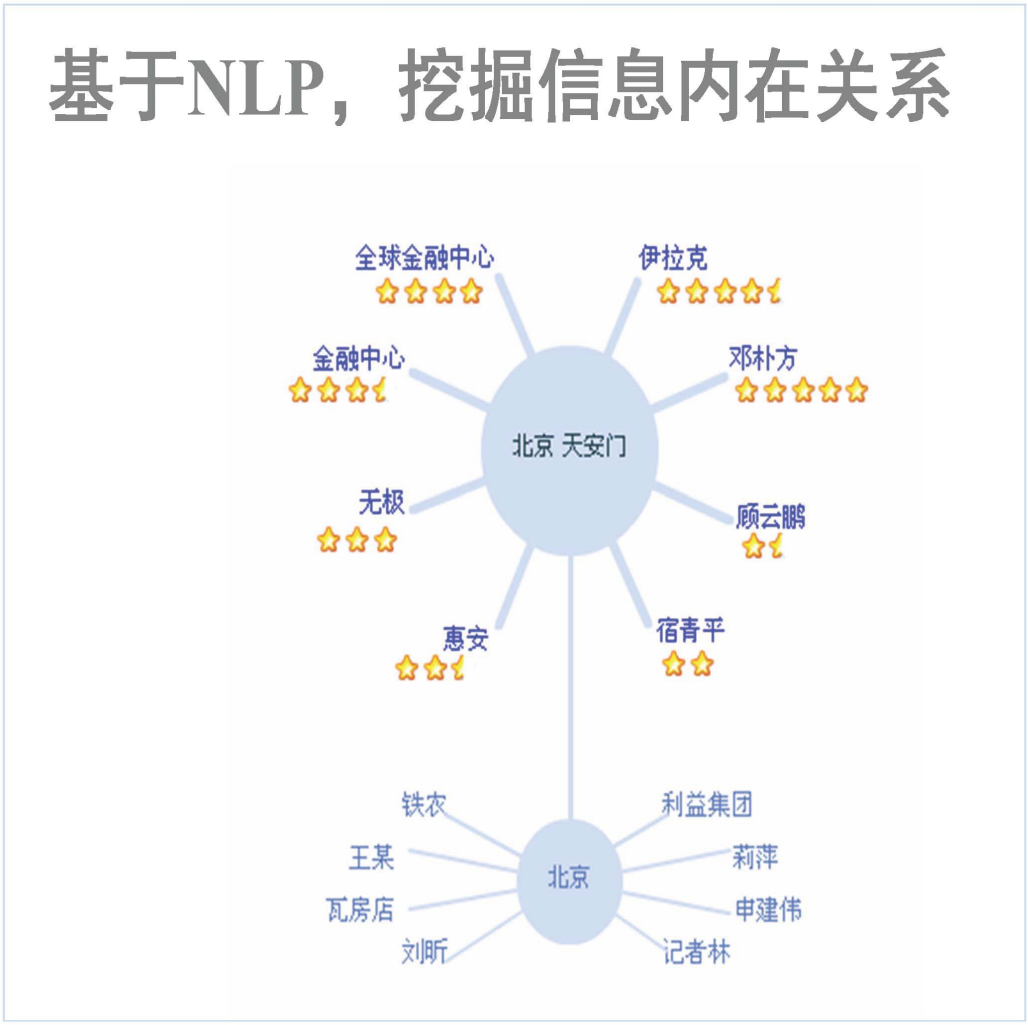

图 24 信息之间内部关系挖掘

参考文献

国内参考

1.web Archive发展历程与发展趋势研究 李华1 吴振新2 郭家义 3向菁2,4《现代图书情报技术》

2.网络信息资源长期保存研究《图书馆理论与实》2007(2)唐琼

3.国互联网信息中心:《第28次中国互联网络发展状况统计报告》.2011年7月19日http://www.cn

4.网络信息存档:档案部门的责任及其策略。周毅 《档案学研究》2010 年第1 期

5.网络信息归档保存的理论与实践探索 周毅

6.国际主要Web_Archive项目介绍与评析 向 菁等《国家图书馆学刊》 2010年第1期(总第71期)

7.政府网络信息资源长期保存研究 唐琼图书馆理论与实践 2007 (2)

8.丰富数字档案馆馆藏的新视角 周文佳 浙江档案 2008 年第9 期

9.守护E时代的记忆 北京图书馆出版社 赵俊玲 2007年2月第一版

10.互联网档案十年发展评述 2008年 罗勇《档案与建设》月刊

11.赵俊玲:《国外关于网络信息资源保存的研究》,《中国图书馆学报》2004年第3 期

12.罗勇:《亟待开展的互联网档案学研究》,《图书情报工作》2006 年第11 期

13.网站的归档李翠云穆林 2006.05 中国档案

国外参考

1. FC 616, 13 Caching inHTTP

http://www.w3.org/Protocols/rfc2616/rfc2616-sec13.html

2.ISO 28500-2009 信息和文件.WARC文件格式

3. http://www.ifs.tuwien.ac.at/anaola/

4. http://pandora.nla.gov.au/index.html

5. http://www.netarchive.dk/

6. http://www.sino.uni-heidelberg.de/dachs/

7. http://deposit.ddb.de/

8. http://www.archipol.nl/

9. http://www.kb.nl/kb/ict/dea/index-en.html

10. http://webarchiv.nkp.cz/

11. http://www.loc.gov/minerva/

12 .http://govinfo.library.unt.edu/

13. http://www.archive.org/

14. http://www.digitalpreservation.gov/

15. http://www.nb.no/paradigma/eng_index.html

16. http://www.kb.nl/coop/nedlib/

17. http://warp.ndl.go.jp/

18. http://www.kb.se/kw3/ENG/

19. http://www.webarchive.org.uk/

20. http://www.pro.gov.uk/webarchive/

21. http://www.infomall.cn/

22. http://webarchive.nlc.gov.cn

23. http://netpreserve.org/about/index.php