数字图书馆与开放源码软件2009

基于开源软件搭建网络信息资源采集与保存平台

吴振新 曲云鹏
李成文 向 菁
中国科学院国家科学图书馆&
国家图书馆

2009－3－12

主要内容

1 背景介绍
2 搭建网络信息资源采集与保存平台
3 采集实验及结果分析
4 WA领域应用开源软件的一些思考

1.背景介绍

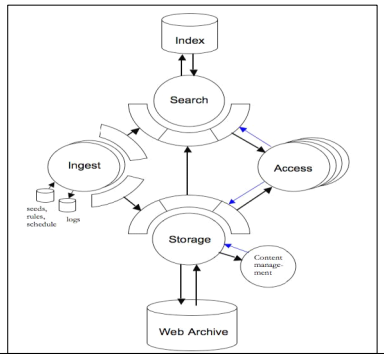
- 网络信息资源的重要性
  - 随着现代信息技术的迅速发展，特别是网络环境的形成，信息的生产、存储和传递的方式发生了革命性的变化，网络成为交流和共享信息资源的重要平台，网络上存在大量对文化遗产、学术研究具有重要价值的信息，因此网络信息资源保存（**Web Archive**，**WA**）成为非常重要和急迫的要求 。
- 图书馆改革其传统的资源建设和存储模式
  - 图书馆作为人类信息资源的主要保存者，也随之改革其传统的资源建设和存储模式，对开放存取的网络信息资源的开发和利用，已经成为图书馆资源建设和保存中的有机组成部分，因此构建一个有效的网络信息资源采集、保存、服务平台成为很多图书馆系统建设所关注的问题。

1.背景介绍

- 国家社会科学基金项目“网络信息资源保存的理论和方法研究”
  - 合作者
    - 中国科学院国家科学图书馆信息系统部
    - 国家图书馆数字资源部
  - 实验
    - 利用开源软件搭建网络信息资源采集和保存平台
    - 完成基于选择策略的专题采集实验

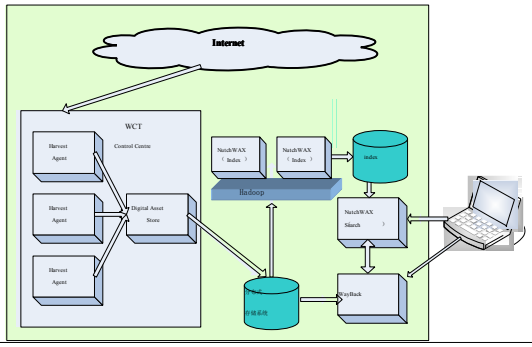
2.搭建网络信息资源采集与保存平台

- 遵循IIPC的Web archive系统技术体系框架



2.搭建网络信息资源采集与保存平台

- 网络信息采集与保存平台系统框架图



2.搭建网络信息资源采集与保存平台

❖相关软件的选择

系统环境	
操作系统	Linux RedHat AS 5
开发语言	Java: JDK1.6
开发平台	Tomcat 5.5
分布式软件平台	Hadoop 0.9.2
关系数据库	MySQL 5.0.51
软件及版本号	
采集工具	WCT 1.3
全文索引工具	NutchWAX 0.10.0
访问工具	Wayback1.4.0

搭建网络信息资源采集与保存平台

❖相关软件的介绍

- 采集工具Web Curator Tool（WCT）
  - 是一款在heritrix基础上开发的选择性网络采集过程管理工具，具备了对Web资源采集进行较好的流程控制、过程管理的功能，拥有完善的采集许可制度、规范流程管理和采集质量审核机制，采集质量高，并能与Nutchwax、wayback进行整合提供索引和访问，非常适合非技术人员（如图书馆员）操作使用。

2.搭建网络信息资源采集与保存平台

❖相关软件的介绍

- NutchWAX
  - 目前Web Archive中应用最为广泛的全文本索引工具，它在Nutch的基础上添加了一些附加软件（add-ons），从而可以为存档的Web数据提供索引和搜索的功能。其在索引中添加额外字段和对非text/html内容的索引功能，索引质量和效率较之Nutch都有所提高，因此NutchWAX是目前最为有效的全文索引工具。
- Wayback
  - 目前Web Archive中应用最为广泛的检索和浏览工具。

3.采集实验及结果分析

❖ 采集实验的技术说明

- 采集时间
  - 2008年7月15日到2008年9月15连续8周
- 采集配置说明
  - WCT采取增量采集方式
  - 采集宽度：针对采集目录中的种子站点，每个站点最多采集5，000个URL；
  - 采集深度：每个站点的 URL采集深度限定为5层，20次跳转；
  - 采集对象大小：超过 100MB的文档将不被收割，并提供关于这些文档的报告；
  - 采集对象格式：最小单元 item，即各种类型文档，如 PPT、PDF等。
  - 每个对象限定1分钟下载时间，当采集器 5分钟内不运行，采集器监测设备将采集器重启；
  - 遵守Robots.txt规则

3.采集实验及结果分析

❖ 采集实验的技术说明

- 采集种子目录的收集
  - 实验利用搜索引擎Google和百度通过限定国家域和网站类型（.com,.org,.edu等），利用给定的检索词来分别获得各国和各类型网站的相关主题URL，经机器自动筛选和人工筛选，确定了2000余个URL作为种子站点。

3.采集实验及结果分析

❖ 实验结果

- 任务在运行开始，保持了比较稳定的采集速度，在130KB/s上下。随着任务不断进行，速率有所下降，最终稳定在20KB/s上下。截止到实验结束，任务持续稳定运行了51天零8小时，共采集数据159GB，URL为4888768个，平均速度53KB/s,URL0.45个/s。
- 从采集目标的数量和状况来看，队列中URL总数为4888768个，其中有效数为3256628个，占66.6%。
- 从已采集资源的类型来看，Text/html类文档共3372220个，占下载总量的６９％，体量约80ＧＢ，占总体量的５０.３％；image/jpeg 类文档共1365922 个，占下载总量的２７．９％，体量约４１ＧＢ，占总体量的２５.８％；video/x-flv类文档共4930 个，占下载总量的０.１％，体量约３２ＧＢ，占总体量的２０.１％。

3.采集实验及结果分析

- 实验分析
  - 爬行深度分析受限
    - 由于存在特定文档格式或者传输机制（如使用Flash技术）
    - 以为已经采集了所有公共内容文档和支持网站显示和功能的文档，如stylesheets和JavaScript文档
    - 采集器在智能化方面还亟待提高

3.采集实验及结果分析

- 实验分析
  - 自动采集收割内容获取的限制因素
    - 无效URL
      - 等待一段时间（可以设置等待时间），造成进度假死
      - 采集前对URL进行有效性检查
      - 利用专用的网站检测工具进行 URL测试，可以排除80%以上的无效URL （探测URL可访问性 ）
    - 临时文件夹的管理
      - 约相当于ARC文件的三分之一
      - 给临时文件夹分配足够的空间，并进行及时的清理

采集实验及结果分析

- 实验分析
  - 自动采集收割内容获取的限制因素
    - 内存的合理分配
      - 实验系统多个地方需要设置 JVM，如linux操作系统、hadoop平台、tomcat、WCT都各自需设置JVM
      - 需合理搭配，否则极易导致系统内存溢出的错误
      - 参考各软件安装手册中所建议的配置外，主要依靠实验中多次调整参数值，观察系统运行情况来决定
    - 网络礼仪的影响
      - 削弱了采集的成果
    - 存档文件的体量
      - Wayback采用的策略是实时访问 ARC文件并解析，而且没有采用缓存策略来增加重复访问速度，因此需要控制ARC文件尺寸以减少访问速度的问题

3.采集实验及结果分析

- 实验分析
  - 访问方式单一
    - 对所保存的资源存在着访问方式单一的问题，虽然已经有些项目在进行该方面的探索，但目前还没有开源的可用软件可以采用
  - 大型网站的扩展性问题
    - 通过下载数量、24小时采集数据量来衡量WCT可扩展性，每天采集集群增加20个采集处理发现：WCT对大规模网站的并行采集的可扩展性效果不理想

4.WA领域应用开源软件的一些思考

- 关注开源软件的易用性
- **WA**需求有别与其它网络信息采集
- 更好地理解开源软件的使用成本
- **WA**开源软件普遍缺乏有力地技术支持

致谢

- 项目组成员
  - 吴振新，张智雄，王志庚，李春明，曲云鹏，李成文，徐健，向菁，周晨，林颖，郭家义，刘兰，孙志茹，王婷

兰色字体部分为项目的平台技术组成员

数字图书馆与开放源码软件2009

谢谢

Thank You !

2009－3－12