互联网网站存档增量采集研究

杨云鹏
（国家图书馆，北京 100081）

摘要：互联网网站存档随着互联网的普及，每年的存储量都在快速增长，导致服务器的存储空间、运行负载和网络带宽已无法满足采集量的增长速度。因此，采集系统过滤掉采集周期内重复的文档实现增量采集将是解决这些问题的关键。本文首先讨论增量采集的采集策略和工具，然后根据采集策略选取合适的工具进行实际采集验证增量采集效果。通过对采集系统添加附加工具的形式实现互联网网站存档增量采集，并对采集的结果进行分析讨论，实现减轻服务器的运行负载、减少网络带宽的占用、降低互联网网站存档存储空间和提高采集资源展示质量的目标。

关键词：互联网网站存档；增量采集；采集策略；网络抓取
中图分类号：TP393.092　　DOI：10.3772/j.issn.1673-2286.2020.12.003
引文格式：杨云鹏. 互联网网站存档增量采集研究[J]. 数字图书馆论坛，2020 (12)：17-21.

第46次《中国互联网络发展状况统计报告》显示，截至2020年6月，网络新闻的使用率达77.1%，网络视频和娱乐的使用率达94.5%，此外40.4%的网民通过微博等社交媒体获取信息[1]。2020年上半年，面对突如其来的新冠肺炎疫情，网络新闻行业深入开展疫情相关报道，通过多种形式助力抗疫斗争[2]。互联网上存储着类似新冠肺炎这种重大事件的全部信息，然而网络信息的寿命一般在90~100天，因此互联网信息存档尤为迫切。

互联网网站存档由于网站部分内容的改变，而必须反复对其进行采集。采集系统如何处理未改变的网站内容是互联网网站存档的一个新的发展方向。采集时如何只保留更改的内容，剔除重复的内容，需要根据不同采集网站的采集需求而采用不同的方式。对于那些需频繁采集且大范围深入采集的网站，重复数据量会非常庞大。如从中国国家图书馆年鉴的互联网网站存档栏目发现2009—2018年随着每年频繁采集相同网站，存储量迅速增加达到了93.73TB[3]，导致国家图书馆服务器存储空间严重不足，由于采集系统没有增量采集功能，导致大量重复资源占据存储空间，造成存储空间的严重浪费。因此，互联网网站存档增量采集是一个亟需解决的问题。

互联网网站存档过程中，使用增量采集主要基于以下4个方面考虑：①减少采集系统服务器的运行负载；②减少单位网络带宽的占用；③降低互联网网站存档存储空间；④提高采集资源展示质量。

1 互联网网站存档增量采集研究

通常认为互联网网站存档增量采集就是保存网站上更改的文档和新文档，它要求采集程序监测在线资源何时发生变化及发生变化的内容。从理论上讲，这种方法是完全没有问题的，它甚至允许软件根据监测到的变化调整采集流程。但是，在实际操作中，这种方法并不可行，因为这种方法会让软件过于敏感，网站细微的变化都会进行采集，导致占用大量内存和采集许多无用数据反而占用存储空间。虽然可以通过更多的人工干预来解决，但是花费大量的人工并不可行，而且即使按照人工干预检测方法也没有一个解决方案可以完全保证不会错过任何有效内容的更改，做到准确的增量采集。因此，除了普通的采集以外，互联网网站存档仍然面临最初的问题，即采用何种采集策略来实现准确增量采集。

1.1 互联网网站存档增量采集策略分析

互联网增量采集策略分两个方面，一方面是根据网站标头、URL和文档内容的唯一Hash值为判断依据进行增量采集；另一方面是分析网站的结构类型，找出变化可能性很小或者不变化的类型，过滤掉这部分进行增量采集。

1.1.1 根据网站标头进行增量采集分析

网站标头包含许多元数据，因此在下载文档正文之前，利用它来检查网站变化是一个很好的方法[4-5]。特别是datestamp（last modified）字段，它包含文档最后一次更改的时间，还有由服务器产生的一个资源标签Etag，可以用于资源的比较，判断资源是否已经被修改[6]。

通过采集测试证明，使用网站标头检查网页是否更改是准确可靠的。但是，此方法并不适合所有网站，对于没有网站标头或者网站标头不准确的网站采用此方法会造成重复下载。尽管如此，使用网站标头检查网站变化是一种非常必要的手段。

1.1.2 根据网站URL进行增量采集分析

国际互联网保存联盟（IIPC）的成员机构是采用Heritrix开发的采集系统，每次采集的时候采集程序将获取所有URL，包含那些重复URL。基于URL的互联网网站存档增量采集，将把每次采集的URL自动存到数据库中，当进行重复采集的时候，会访问数据库，如果其中存在相同URL将放弃采集，从而实现增量采集的要求。

对没有确切链接和内容相同URL不同的内容无法使用URL方法进行增量采集，因此基于URL的增量采集可以作为一种手段，但不能作为唯一的手段。

1.1.3 根据网站文档内容的唯一Hash值进行增量采集分析

Hash函数是一种算法，将任意长度的数据映射到固定长度的Hash值，该值是数据的唯一标识[6]。Hash值具有唯一性，只要有微小的变化就会得到不同的Hash值，只有完全相同的内容才会得到相同的Hash值。

基于以上理论，采集系统使用SHA-1算法自动计算出网页文档的Hash值。增量采集时，通过对比Hash值以检查网页是否重复。通过Hash值进行的增量采集很可靠，但是Hash值过于敏感，微小的变化都将重新采集，导致占用大量内存和采集许多无用数据反而占用硬盘存储空间。

1.1.4 根据网站结构探讨增量采集

网站按照存储结构由网页文档（html）、其他文档（dns、JavaScript和css等）、图片（jpeg、gif和png等）、视频（mp4、mpeg和jav等）和其他（xml、octet-stream和x-shockwave-flash等）组成。网站并非所有类型的存储结构都是经常改变的，某些类型的存储结构比其他类型更能保持不变[7-8]。这并不是说它们永远不会改变，而只是说它们改变的可能性很小。

如图1所示，在大型采集周期中，网页文档URLs在采集中比例最高为42%。但是，对比其数据量时，网页文档只占11%。同时，除纯文本和图片格式以外的其他文件占数据量的22%，而在文档总数中只占18%。

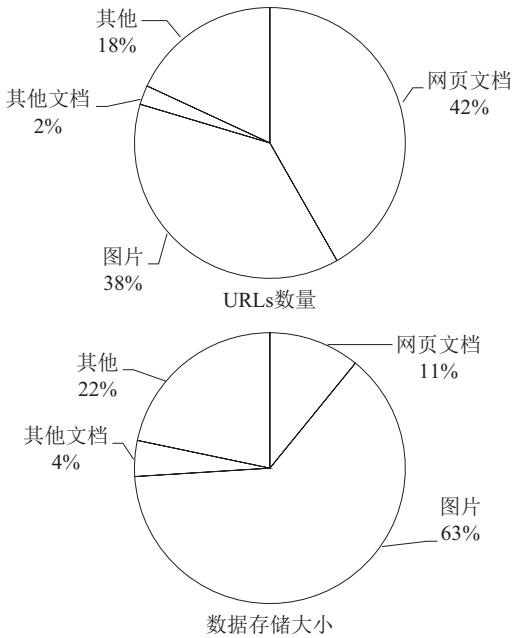

图1 网站快照收集的不同文档类型的数量比较

显然，过滤掉未更改的非文本文件，可以大幅减少存储在硬盘上的冗余数据。实际上，通过增量采集的方法采集，会发现随着采集深度的增加，非文本数据的重复率从50%提高到了90%。

通过分析网站的存储结构类型，只针对很少改变的文件类型进行增量采集可以大幅降低系统的内存占用率和硬盘存储压力，是互联网网站存档增量采集的前提和必要手段。

1.2 互联网网站存档增量采集工具分析

互联网网站存档增量采集是互联网网站存档的研究热点，存在多种增量采集的工具。英国大英图书馆使用UKWA-Heritrix和OutbackCDX两个工具进行增量采集[9]。美国国会图书馆通过在采集系统添加FetchHistoryProcessor和PersistStoreProcessor模块的方法进行增量采集[10]。埃及亚历山大图书馆利用warcrefs工具进行增量采集[11]。冰岛国家图书馆和日本国家图书馆利用DeDuplicator工具进行增量采集。

通过增量采集策略分析发现单一方法并不能满足增量采集的要求，需要根据实际情况用多种方法同时进行才能取长补短满足增量采集的要求。

大英图书馆和美国国会图书馆都是利用Hash值采集策略将采集数据的Hash值自动存储到数据库中，每次采集的时候比对数据库中的数据进行增量采集。如上文分析的单独用Hash值进行查重过于敏感，同时这种增量采集的方法会严重影响爬虫的性能。这种增量采集的主要问题是维护其已发现的采集数据库。通过添加另一个数据存储库或扩充现有数据库来解决此问题都会影响爬虫的性能。

埃及亚历山大图书馆的增量采集方法是对采集之后的文件进行分析，利用Hash值原理将重复的数据剔除，然后重新生成一个只有新数据和修改数据的文件。这种方法是采集后进行增量采集，并不会影响爬虫性能，也避免Hash值过于敏感造成频繁采集的情况，但是这种方法需要先采集，占用大量存储空间，而且如果遇到大型网站会耗费大量时间来删除重复数据，同时占用服务器大量内存，因此这种方法更适用于小型网站的增量采集。

冰岛国家图书馆和日本国家图书馆采用的DeDuplicator具有很多参数可以选择。①网页文档类型（mime-filter），选择过滤文本文档、非文本文档或者两者都选；②去重参数（matching-method），查找匹配项时需要确认先查找URL还是Hash值，如果将其设置为BOTH，则任何一种设置都将起作用；③网站标头时间戳参数（timestamp），即上次收集文件的时间。DeDuplicator可以根据不同的网站选择不同的参数组合来进行增量采集。由于增量采集会根据网站的更新频率确定两次采集之间的时间间隔，而DeDuplicator是在两次采集之间创建索引，因此，其对采集程序性能的影响最小。

综上所述，通过分析各个增量采集工具，DeDuplicator优势明显，更适合互联网网站存档增量采集。

2 增量采集工具DeDuplicator采集分析

在2011年IIPC大会[12]中，日本国家图书馆预估重复数据删除将使日本每月的网络归档减少80%，每季度的归档减少45%，证明DeDuplicator工具在日文增量采集中效果明显。本次将测试DeDuplicator工具在中文网页增量采集的去重率。

2.1 增量采集系统配置方法

在抓取或处理网站时，采集系统使用了一系列采集处理器。采集系统可以根据采集需求由管理员决定采用哪些处理器，但处理器的顺序是固定的，不可随意更改。因此，采集系统很容易引入其他处理器。

DeDuplicator处理器设计为在下载HTTP文档的FetchHTTP处理器之后运行。它具有两种不同的匹配模式，分别是URL和Hash值。增量采集时如果选取URL模式，则将DeDuplicator处理器matchingMethod的value填写为URL即可；反之，则填写为Hash。

2.2 增量采集测试

为了测试DeDuplicator工具对中文网站的增量采集效果，对国家图书馆二级网站和专题网站进行了多次采集，将采集方式分为两种类型。

（1）整站增量采集。采集间隔为90天，进行了2次增量采集。每次采集的最大深度是3层。采集覆盖大约1.5万个网址，每次的采集时间大约为3天。

（2）专题增量采集。采集间隔为一周，每次采集持续24小时，覆盖1.2万个文档。

两种类型增量采集都完成了深入的采集，满足了对网站的全面捕获。停止采集时，只有极少数非常深的站点仍在处理中，采集完成率达到99.5%。

2.3 增量采集数据分析

整站增量采集时，通过分析网站存储结构类型，采用非文本文档进行索引，增量采集索引文件与采集时间间隔2个月并采用了一样的配置文件。两次采集网站下载的数据总量在未进行增量采集前，数据量预估大约为800MB（未压缩数据量），但是进行增量采集后，磁盘上实际存储的数据量（压缩后数据量）从691MB减少到了156MB，存储空间节省77%。

如图2所示，DeDuplicator插件处理的非文本文档中有83.6%是重复的。在采集的1.5万个网址中有7 412个是非文本文档，并且其中有6 197个是重复项，占总网址的41.3%。

```
Processor: is.hi.bok.digest.DeDuplicator
  Function:          Abort processing of duplicate records
                     - Lookup by url in use
  Total handled:     7412
  Duplicates found:  6197 83.60%
  Bytes total:       637140246 (608 MiB)
  Bytes discarded:   510643182 (487 MiB) 80.14%
  New (no hits):     1215
  Exact hits:        6197
  Equivalent hits:   0
  Timestamp predicts: (Where exact URL existed in the index)
  Change correctly:  7
  Change falsly:     29
  Non-change correct:6126
  Non-change falsly: 54
  Missing timpestamp:49
```

图2 增量采集统计信息

通过URL匹配索引中的网址，初步筛选出6 265（7+29+6 126+54+49）个重复项，进一步采用Hash值准确匹配，筛选出精确的重复数据为6 197个文档。这意味着实际上只有68个已采集文档（约0.9%）发生更改，此结论支持了增量采集策略最初的假设，即有些类型的文档很少更改。

本次测试非文本文档占608MB，其中487MB被视为重复数据，节省约80%存储空间。这个数字证明增量采集工具采用策略的正确性，非文本文档大部分文件保持不变。

第二次整站增量采集是在第一次采集的基础上进行索引，两次采集间隔90天。第二次增量采集结果证明了第一次增量采集结果的正确性。由于它是在第一次增量采集的基础上进行的第二次增量采集，因此重复数量更高达10 050项，占总网址的67%。

专题增量采集相比整站采集规模要小得多，一般仅由几十个网页组成。专题也比网站的更新频率更低，只是在特定时间进行更新。本次测试针对热点专题进行采集，采集频率为一周，每次只采集更新的内容，结果发现专题增量采集效果比整站增量采集更好，有98%的数据是重复的。对于不频繁更新的专题采集，增量采集的效果更为突出，大幅降低了每次采集的存储量。

通过比较两种方式增量采集结果，根据网站采集的经验，发现以下3个因素会影响到增量采集的重复率。①网站采集的时间间隔，间隔时间越短重复率越高；②网站采集的深度，越深层的采集将会遇到越多不变的数据；③网站的更新频率，网站更新频率越高重复率就越低。

专题增量采集采用相同配置和时间间隔，采用文本文档作为索引数据源进行了增量采集对比测试。文本文档的增量采集匹配出34%的文档为重复数据，占据总存储量的25%。这个结果比非文本文档的结果差很多，并且由于文本文档数量较多，会大幅降低采集效率，增加采集时间，因此综合考虑非文本文档更适合普通网站的增量采集。

2.4 增量采集整体评估

通过长时间的采集测试，完成了DeDuplicator增量采集插件对采集系统性能影响的评估。受采集规则和可用资源的限制，专题增量采集量较小，对采集系统不会产生任何影响。增量采集过程中尝试使用完整索引（过滤出文本文档和非文本文档）时，会对采集程序产生较大影响。对于大型采集任务，需要根据实际情况选择文本文档或非文本文档，但对于小型的采集任务则可以忽略不计。采集测试过程中发现，仅过滤掉非文本文档对大型采集的采集速度影响很小，实际上，增量采集开始的前几个小时，每秒文档的采集速率对比普通采集反而有所提高，但随着采集时间的增加，采集速度对比普通采集大约会降低10%。

大型采集任务出现采集速度降低的现象与采集系统处理内部存储分布方式有关，大部分采集数据存储在Berkley数据库（BDB）中。虽然BDB已将数据写入磁盘，但因内存中有数据缓存，最初几乎不需要从磁盘读取数据。但是，随着采集时间变长，BDB必须越来越多地访问磁盘上的内容，开始与增量采集插件Lucene的磁盘访问请求产生冲突，造成采集速率降低。未来通过优化采集系统的内存结构，将会大幅提高增量采集的采集速度。

3 结论

通过测试证明，互联网网站存档增量采集插件DeDuplicator对普通的中文网站同样效果明显。DeDuplicator工具根据网站结构灵活地选择增量采集策略，有效地减少了采集过程中服务器的负载和网络带宽的占用。增量采集工具的使用不但降低了采集带宽负担，而且不用再人为地限制采集速度，以免影响其他服务，也不用因长期采集导致数据大量占据服务器内存从而需要频繁重启采集系统。

增量采集工具DeDuplicator的最大优势是降低存储空间占用。通过采集测试发现整站增量采集能节省约80%的存储空间，专题增量采集能节省超过90%的存储空间。增量采集工具是根据网站结构进行增量采集的，与网站的大小关系不大，只要网站结构不变，增量采集去重率变化不大。采用增量采集工具进行增量采集可以节省大量存储空间，尤其是对于更新不频繁和深度采集的大型域名类网站，效果更加明显。

实际上，所有的增量采集都还有一个优点，在回放系统进行搜索时，均不会显示多个相同网址，或多个一样的重复资源。因此，采集资源在进行集中展示和服务的时候，不需要再进行人工删除和筛选，大大提高了资源的可利用价值，也节省了时间和人力资源。

随着互联网技术的快速发展，越来越多的资源和数据都存储在网络中，其中包含大量的重复数据，因此增量采集将是互联网数据采集的一个重要研究方向。现今国内外许多互联网保存组织也都在积极地研发增量采集技术，相信未来必定会有更先进的技术，既能节约存储空间，又能提高采集效率。

参考文献

[1] 中国互联网信息中心. 第46次《中国互联网络发展状况统计报告》[EB/OL]. [2020-09-29]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/202009/t20200929_71257.htm.

[2] 周文泓，许强宁，高振华，等. 新冠肺炎疫情的网络信息存档实践——兼论其对重大社会事件网络信息存档的策略性启示 [J]. 浙江档案，2020（7）：31-33.

[3] 中国国家图书馆. 国家图书馆年鉴 [M]. 北京：国家图书馆出版社，2019.

[4] 赵思远. HTTP协议头及错误码详解 [J]. 计算机与网络，2016（11）：40-43.

[5] 任栋，刘连忠. HTTP头信息在Web网站开发中的应用研究 [J]. 计算机应用，2002（3）：65-67.

[6] 张皓，周学广. 基于Heritrix的增量式网络爬虫研究 [J]. 软件导刊，2013，12（11）：135-137.

[7] 孟庆浩. 互联网数据增量采集系统的设计与实现 [D]. 北京：北京邮电大学，2015.

[8] 孟庆浩，王晶，沈奇威. 基于Heritrix的增量式爬虫设计与实现 [J]. 电信技术，2014（9）：97-101.

[9] 高婷，白如江. 基于OutbackCDX的增量式Web信息采集研究 [J]. 山东理工大学学报（社会科学版），2020，36（4）：99-105.

[10] UK WEB ARCHIVE [EB/OL]. [2020-12-12]. https://www.webarchive.org.uk/en/ukwa/index.

[11] ELDAKAR Y，NAGI M. Deduplicating Bibliotheca Alexandrina's Web Archive [C] //iPRES 2015，North Carolina，2015.

[12] IIPC General Assemblypresentation [EB/OL]. [2020-11-01]. http://www.netpreserve.org/ general-assembly/2011/Overview.

作者简介

杨云鹏，男，1986年生，硕士，工程师，研究方向：数字图书馆资源整合和互联网采集，E-mail：syzyyp@126.com。

Research on Incremental Collection of Internet Archive

YANG YunPeng
(National Library of China, Beijing 100081, China)

Abstract: Internet archive with the popularity of the internet, the amount of storage is growing rapidly every year. The storage space, operating load, and network bandwidth of the server can no longer meet the growth rate of the collection volume. Therefore, the key to solve these problems is to filter out the repeated documents in the collection cycle and realize incremental collection. This paper first discusses the acquisition strategy and tools of incremental acquisition, and then selects the appropriate tool according to the acquisition strategy for actual acquisition to verify the effect of incremental acquisition. Through the addition of additional tools to the collection system, the Internet archives incremental collection is realized, and the collected results are analyzed and discussed. The goals of reducing the operating load of the server, reducing the occupation of network bandwidth, reducing internet archive storage space and improving the display quality of collected resources are achieved.

Keywords: Internet Archive; Incremental Acquisition; Acquisition Strategy; Web Scraping

（收稿日期：2020-11-14）