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

网页变化监测相关技术与方法研究*

□ 高建秀 吴振新 张智雄 / 中国科学院国家科学图书馆 北京 100190

摘要：有效的网页变化监测技术可以极大提高地保存的效率，因此网页变化监测研究对网络资源长期保存显得十分必要。文章总结了现有的网页变化特点研究成果，指出了目前所采用的网页监测变化频率选择策略，分析了网页变化检测的技术和方法，并指明其发展面临的困难和挑战。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：长期保存，网页变化监测，重访频率
DOI：10.3772/j.issn.1673—2286.2009.07.003

1 引言

随着信息技术的发展，Internet已经成为国家的信息基础设施，融入了社会生活的方方面面，由此，网络资源成为社会文化遗产的重要组成部分。Web资源作为全球最大的信息资源库，本身具有海量、无序等特征，并且每天在以指数级速度不断增长。如IA保存的数据已达2 PB（1 PB=1024 TB，约等于1千千兆比特），并且目前在以每月20 TB的速度增长[1]。 同时，网页的更新十分迅速，相对于印本资源，网页的寿命相当短暂。

为此，很多国家和机构积极开展了网络信息资源保存（Web archive，简称WA）工作。然而，由于Web资源自身的特性使得WA面临着一系列的困难。由于网络资源变化速度快，需要在网页消失或变化前对网页进行及时采集，但由于资源的海量，全部采集所需时间也非常可观。另外，由于WA累积性保存要求，不可能对所有的资源重复存储，而是通过变化监测找出发生变化了的资源，仅对变化的资源增量式地累积存储。如何有效采集和保存网络资源，一直是WA面临的难题。为此，WA专家致力于研究如何更好地掌握网页发生改变的特点以及规律，找出更好的监测方法，使得WA系统能够更有效地对网页进行及时、有效地存档。

2 网页变化的特点

网络中的网页数量庞大以及时刻动态变化的特性，使得对这些海量URL地址的变化监测十分困难。合理的监测频率可以大大减少不必要的重载，采用高效的网页变化监测算法，可以显著提高网页变化监测系统的性能。同时，如果为了监测网页变化，时刻重载所监测的页面并进行比较，对系统的并行性要求很高，对于大型的WA项目，这显然是不可行的。因此，寻找网页变化的特点和规律，成为网页变化监测的首要任务。

根据搜索领域的相关研究，发现网页变化确实存在一些规律。

- 在短时期内发生变化的网页数量相比于网页内容总量仍然较小

孟涛等人[2]在2003年8月15日至9月16日的一个月中对1,637,253个网页做了121次搜集检查，发现每次变化的网页数量占总体的百分比例如图1所示，整个过程中，内容曾经发生变化的网页数量仅约占样本容量的8.52%，绝大多数网页在短期内不会发生变化 。

- 网页变化的频率与域名种类、页面大小等因素关系密切

DennisFetterly、MarkManasse等人所做的实验表

* 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。

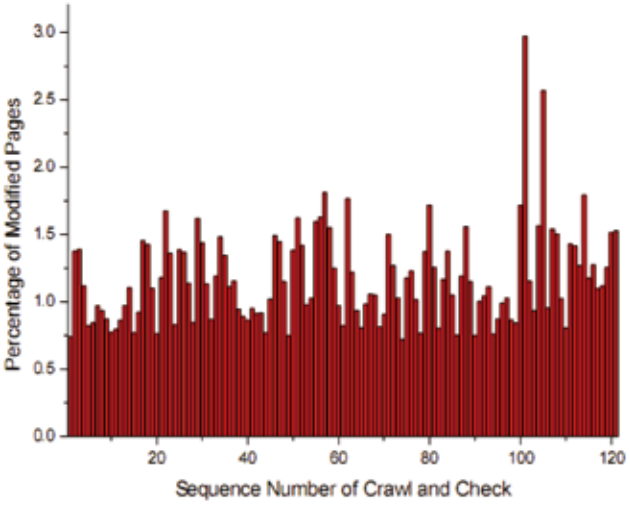

图1 全局网页中发生变化的网页数量[1]

明[3]，以com和net为顶级域名的网页要比.gov和.edu的变化更为显著。从图2中可以看出：经常更新的网页主要是.com的网页，有超过40%的网页在每天发生改变，有10%的其他域名的网页同样以这种频率变化。

该实验还表明：网页文件的大小和网页变化程度、快慢密切相关。大的网页文件比小的网页文件更容易发生变化。

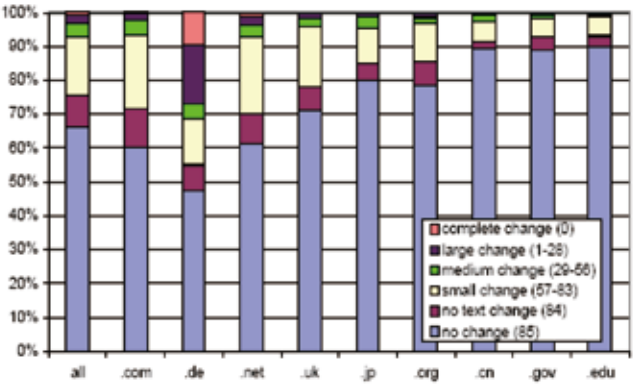

图2 不同域名下的网页变化分布[2]

- 发生变化的页面一般会在不久后继续发生变化

DennisFetterly、MarkManasse等人所做实验还得出结论[3]，一个网页过去所发生的变化可以对将来变化程度和频率做出很好的预测。即存在类似于内存被使用之后可能再次被使用的情况，网页在发生变化之后，在不久的将来有可能会再次发生变化。

由此可以发现网页变化确实具有一定的规律可循，网页变化的特点和历史记录可以为以后的变化监测提供依据。

3 网页变化监测的频率选择策略

网络上的网页资源是独立于本地数据进行更新的，所以需要定期重访网页以监测变化，并将新变化后的网页资源重载到资源数据库中。这种变化监测和页面重载，由于各种因素限制通常会按照一定的频率进行，因此选择和确定一个合适的监测频率，对于保存的完整性和有效性是非常关键的。目前常用的几种频率选择策略包括：统一策略、基于网页变化历史的策略、基于样本的策略。

3.1 统一策略

这种策略使用相同的频率，重新访问URL列表中的所有链接，而不考虑它们各自的更新频率。原理是对监测频率给定一个取值，爬行器每到需要检测网页变化的时间点，就重复爬行一次所有网页，将变化了的网页下载、保存。

这种策略简单易行，被当前的很多系统所采用。但是，通过统一策略监测网页变化，由于体量大所需周期会比较长。如果需要维护100万的网页，而每周只能重载10万张网页，那么每个网页被更新的周期就是10周。另外，各种页面变化频率差别很大，对于生命周期短暂的网页而言，其中部分网页已经改变或消失，难以及时准确地保存已经变化的网页。同时，重载所有页面并进行对比，对数据量十分庞大的系统来说，将消耗过多的系统资源。

3.2 基于网页变化历史的策略

基于网页变化历史的策略，需要搜集网页变化的历史轨迹。简单的方法是变化的总次数X/时间间隔T。起初为每一个网页设定一个生存周期，到达生存周期结束时刻就进行重访监测。当对某个网页的变化频率有一定的统计估计值后，根据估计的网页变化频率来调整这个网页的生存周期。比如，如果一年内检测到四次变化，那么就预测它的周期是3个月。这样，基本能够保证网页的变化频率越高，更新频率也越高。

当能够精确掌握数据的变化频率且这个频率相对稳定时，这种方法是比较理想的。但在实际应用中，需要相当长的时间才能为每个网页获得足以对它的变化频率作有效预测的历史变化轨迹，另外，网页的变

化频率常常是不规律的，通常很难分析出网页的精确变化频率，现有的分析方法获得的结果往往不可靠，从而导致制定了不可信的重载策略。同时，为了分析变化频率，需要跟踪每个网页的变化历史，在网页数目很大的情况下，需要存储和维护庞大的数据，会极大地耗费系统资源，因此网页历史变化轨迹维护会成为这种方法的瓶颈。

贝尔实验室的E．G．Coffmant及斯坦福大学的Cho等人对页面更新频率进行了研究[4]，利用统计方法分析页面更新历史记录，提出用泊松 (Poisson)过程来模拟独立页面的变更过程，并给出一系列方法来估算独立网页生存期和整个 Web间内的页面生存期的分布规律。

3.3 基于样本的策略

为了避开上述策略对网页变化建模的需求而带来的一系列弊端，有些项目采用了通过样本采样估计网页变化频率的方法，即基于样本的策略。

这种策略的基本出发点是：绝大多数网页以网站或其它群体形式聚集，不同的网页群体之间的平均变化频率相差极大，但同一群中变化频率接近，因此通过采集一定数量的样本页面，以样本页面的变化频率来确定所属群体的变化频率。

首先对网页进行分组，例如可以将同一网站内的网页作为一组。从每组中选出一部分作为样本进行采集监测，分析样本页面有多少发生了变化。然后根据分析结果，计算分配到每个网页分组中应该重载的网页数量，将除去样本之后剩余的待搜集网页数量分配到各个分组中。如：以10，000个站点作为样本（包含100，000个网页样本），从每个站点中下载10个网页。然后计算样本中多少个网页发生了变化，根据这个计算值，来分配剩余的900，000个下载网页的名额。

目前，美国加州大学（UCLA）的WebArchive系统中采用了这种方法。Junghoo Cho、Alexandros Ntoulas等人设计了自适应算法和贪心算法来优化样本大小的选择[5]，并通过实验分析得出结论：如果要获得最优性能，可以将基于变化频率和基于样本的策略相结合使用，即在初始时先通过基于样本的方法，当搜集到足够多的变化历史轨迹数据时，开始采用基于变化历史的策略。但是，何时进行这种策略转换仍是待研究的问题。

4 网页变化检测的方法

由于关注点不同，对于网页变化的认定也不尽相同，通常包括下面几种变化：

（1）内容/语义改变：从读者观点来看，发生的内容上的改变；

（2）表现/外观改变：HTML标签改变；

（3）结构性改变：网页与网页之间链接的改变；

（4）行为改变：网页的动态成分，如脚本程序、插件、应用程序的改变。

目前有多种方法可以实现上述几种类型网页变化的监测，常用的网页变化检测技术方法有三种：基于HTTP协议头的状态检查、基于特征字符串比较法、基于文档树结构比较法。

4.1 基于HTTP协议头的状态检查

表1　HTTP响应的头域信息

应答头	说明
Content-Length	表示内容长度。
If-Modified-Since	标识自指定日期以来，此资源是否已经被修改。
Last-Modified	指定被请求资源上次被修改的日期和时间。
Etag	确定实际被发送的资源是否为同一资源。

当对一个页面发出访问请求时，浏览器首先获得HTTP响应。在HTTP 消息的头域中可能会包含以下可以利用的信息：

最常用的检查页面变化方法是把Last-Modified 和ETag请求一起使用。当第一次抓取请求某一个URL时， Last-Modified的属性标记此文件在服务器端最后被修改的时间，ETag记录被请求变量的实体值，格式类似：

Last-Modified: Fri, 12 May 2006 18:53:33 GMT

ETag: "50b1c1d4f775c61:df3"

当第二次请求此URL时，根据 HTTP 协议的规定，浏览器会向服务器传送 If-Modified-Since 和If-None-Match报头，询问该时间点之后文件是否被修改过：

If-Modified-Since: Fri, 12 May 2006 18:53:33 GMT

If-None-Match: W/"50b1c1d4f775c61:df3"

如果服务器端的资源没有变化，则自动返回 HTTP 304 （Not Changed.）状态码。当服务器端资源发生改变时，服务器的响应代码是 HTTP 200 (OK)。

除此之外，还可以比较Content-Length的值，因为内容的有效变化几乎一定伴随文档长度的变化。

通过这种方法，可以判断网页是否发生了变化。如果值不同，就进行重新采集。否则，认为网页没有变化，不进行实际的下载操作。这种方法的优点在于快速易行，能够为系统节省大量不必要的网络资源，并节省时间，在一定程度上能够审查网页是否有更新。但是在网络信息资源保存中这种策略往往是不可信赖的，如果单纯依赖这种方法，对于时间戳或页面大小未发生改变的网页则认定网页没有发生变化，会错失很多更新。另外，网络中很多网页是动态页面，或者并不设置Last-Modified 和ETag等头域信息，这样通过上面的方法并不能监测到网页发生了变化。因此，在实际应用中常常将这种方法与其他方法相结合使用。

日本京都大学研发的Past Web Browser采用了这种策略，采集模块下载数据将URL和时间戳传递给浏览器，浏览器进行比较后决定页面是否为可供用户浏览的新版本。这种方法适用于网页数据量大、内容更新变化较少的系统中。

丹麦国家图书馆曾对丹麦域进行采集实验，通过矩阵方式比较哪种指标更能准确判断网页的更新，从而判断各种方案的有用性和可靠性。他们的实验并没能证明时间戳和Etag在指示网页变化方面的可靠性，实验表明通过比较时间戳和Etag来进行网页变化监测的方案并不是很理想。

4.2 特征字符串比较法

由于网页资源的海量和系统对时间要求的高效性，对网页内容逐字精确比较是不可能的，对于内容较长的网页，一般采取计算网页特征字符串的方法，取其MD5摘要值来作为判断网页变化的特征字符串值。MD5(Message-Digest Algorithm 5：信息-摘要算法)是根据数据生成一个特征值（即对原数据作HASH变换得到一个HASH值），可以证明数据的“唯一”性。一旦MD5值发生了改变，则网页内容一定发生了变化。

这种方法常常与上述基于HTTP协议的状态检查结合使用，首先利用HTTP协议头初步判断网页是否发生了变化，如果发生变化，则不计算特征字符串值，直接进行重载。初步判定结果未发生变化，则进一步计算MD5值，比较是否与原来系统中保存的网页特征字符串值相同。目前，北大天网搜索的增量采集系统中使用了这种方法，运行效果良好。

但是，这样定义的网页变化往往过于严格。例如网页中常有一段显示当天时间或者显示访问站点人数的计数器的脚本，这段固定格式的内容可能每天都会变，但这种变化是没有意义的，不必进行网页的更新保存。系统期望排除并不关心的那部分网页内容的变化，直接分析想要关注的内容，根据这部分内容是否发生了变化来决定是否重新保存网页。因此有些系统中提出了基于文档树结构的比较法。

4.3 树结构比较法

基于文档树结构的比较方法特别适用于检测网页特定部分内容变化的情况。在进行网页变化检测中，常常关注网页某特定部分内容，如只关心显示在浏览器的窗口中BODY标签中的网页主体内容，允许对网页中嵌入的广告、网站统计信息等变化不敏感。这种方法利用了HTML的半结构化特性，将网页视为嵌套了标签元素的文档树。比如，一个网页可以解析成如下所示的树状结构视图图3：

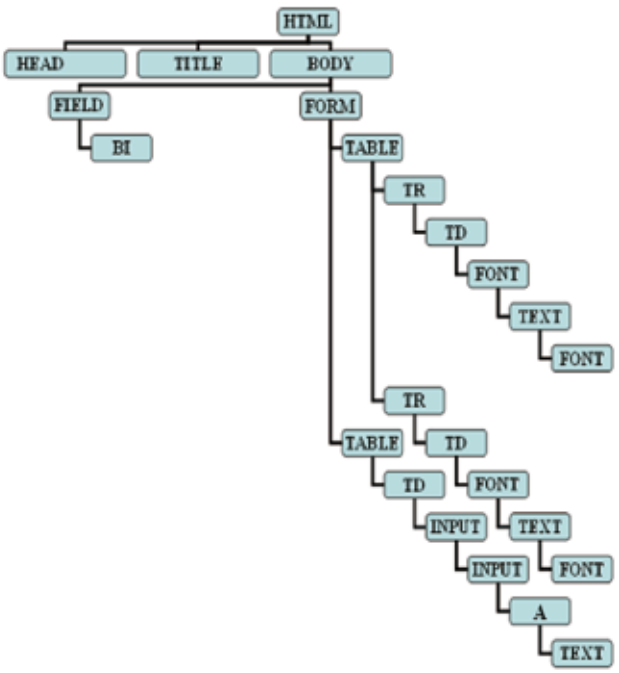

图3 网页树形结构图[3]

监测网页的变化，可以通过比较新旧网页树结构视图的不同，如结点数目、结点位置、结点取值等。进行基于树结构的变化监测时，一般需要具体包含的功能模块有：

（1）文档树构造：将HTML文件作为输入，利用语法分析器，解析出树，把打开的标签作为树的节点；

（2）语法分析器：按层次顺序列出儿子节点；

（3）绘制结构树；

（4）与旧网页对比做出变化监测。

用这种方法可以自动从网页中提取特定内容，比较所关心内容部分是否发生了变化。这样，能大大提高保存网页版本的保存效率。基于文档树结构的比较法较常用于网络上新兴的在线变化监测系统，如ChangeDetection[6]、URLy Warning[7]、WebVigiL[8]等。这类系统允许用户定制：监测哪个网页、监测什么内容（所有变化还是仅关键词）、单个的网页还是某个路径下的所有网页、如何监测、监测的频率或什么时候监测、监测开始和终止的时间、通过什么方式通知给用户、多久发送一次等等。

这类系统一般是针对小规模的网页进行检测，不适用于大规模的采集保存活动。基于文档树的比较法目前在Web Achieve领域还没有成型的应用系统出现。

5 结语

基于HTTP协议头的状态检查比较简单，能够节省时间和资源，可以应用于对网页变化要求不太严格的系统，但会错失某些网页的更新。将其与基于特征字符串比较法结合使用，能够克服该缺点，却往往过于严格。WA系统主要关注内容上的改变，不期望对对某些广告信息变化、站点访问统计信息的变化进行监测。为此，研究人员提出了基于文档树结构比较法，该方法能够提取出特定部分内容进行比较，在应用上却存在一定的困难，如每次重访对文档结构进行解析会牺牲系统的时间代价。

因此，虽然存在着多种网页变化监测技术和方法，但由于受网络带宽和系统并发等因素限制，目前在网络信息资源保存领域仍没有十分合理、有效的方法实现页面变化监测, 网页变化监测技术和方法依然是有效采集和保存的研究难点和重点。随着技术环境的变化、网络资源体量的飞速发展以及网络媒体形式的复杂化，网页变化监测技术将面临更多的问题和挑战。

参考文献

[1] Internet Archive [EB/OL].http://www.archive.org/about/faqs.php#9 [2009-04-09]
[2] 孟涛，闫宏飞，王继民：Web网页信息变化的时间局部性规律及其验证[J], 情报学报, 2005,24(4): 398-406.
[3] FETTERLY D, MANASSE M, NAJORK M, WIENER J L. A large-scale study of the evolution of web pages[J]. Softw Pract Exper, 2004, 34(2):213-237.
[4] Junghoo Cho, Alexandros Ntoulas. Effective Change Detection Using Sampling[C]//Proceedings of 28th International Conference on Very Large Databases, Hongkong, China: Morgan Kaufmann, 2002: 514-525.
[5] Alexandros Ntoulas, Junghoo Cho, Christopher Olston. What's New on the Web? The Evolution of the Web from a Search Engine Perspective[C]//Proceedings of the 13th World-Wide Web Conference, New York, USA: ACM Press, 2004: 1-12.
[6] ChangeDetection [EB/OL]. [2009-04-09].http://www.changedetection.com/.
[7] Divakar Yadav , A. K. Sharma, J. P. Gupta. Change Detection in Web Pages[C]//Proceedings of the 10th International Conference on Information Technology, 2007: 265-270.
[8] Sharma Chakravarthy, Subramanian C. Hari Hara. Automating Change Detection and Notification of Web Pages (Invited Paper)[C]//Proceedings of the 17th International Conference on Database and Expert Systems Applications,2006:465-469.

作者简介

高建秀（1985-），中国科学院国家科学图书馆2008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，100190。E-mail：gaojianxiu@mail.las.ac.cn
吴振新，中国科学院国家科学图书馆副研究馆员，数字资源长期保存方向。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
张智雄，中国科学院国家科学图书馆研究馆员，知识技术方向。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: zhangzx@mail.las.ac.cn

Research on Web Change Detection

Gao Jianxiu, Wu Zhenxin, Zhang Zhixiong / National Science Library, Beijing, 100190

Abstract: The effective web change detection technology may enhance the efficiency of web archive greatly, so the web change detection is necessary to the web archive. This paper sums up the research of change characteristics of web pages, points out the selection strategy of the web detection frequency, analyzes the technology and method of the web change detection. It aslo indicates the obstacle and challeng during its development.

Keywords: Web archive, Web change detection, Revisit frequency

（收稿日期：2009-05-15；责任编辑：贾延霞）