基于 Wayback的索引策略研究

孙志茹[1,2]　吴振新[1]　曲云鹏[3]
[1](中国科学院国家科学图书馆　北京 100190)
[2](中国科学院研究生院　北京 100049)
[3](中国国家图书馆　北京 100081)

【摘要】通过分析目前使用 Wayback作为访问工具的 Web Archive项目,总结出资源索引工作采用的几个典型索引策略,并对其适用范围及优缺点进行初步分析和探讨,以期为界内同行提供参考。
【关键词】Web Archive　Wayback　索引策略
【分类号】G202

Analysis of Index Strategies in Web Archive

Sun Zhiru[1,2]　Wu Zhenxin[1]　Qu Yunpeng[3]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](Graduate University of the Chinese Academy Sciences, Beijing 100049, China)
[3](National Library of China, Beijing 100081, China)

【Abstract】This article summarizes several typical index strategies through analyzing Web Archive projects with Wayback as access tool, also gives preliminary analysis for the scope of application, merits and faults of each strategy. Thus hopes to give companies of this area some reference.
【Keywords】Web Archive　Wayback　Index strategy

1 引　言

随着网络的蓬勃发展,Web资源正日益成为文化遗产的重要组成部分,Web资源的采集和保存活动,即 Web Archive已成为各国保存领域的研究重点。经 Web Archive所存档的 Web资源不但继承了 Web资源本身的动态、增长、海量等特性,还强化了这些特性并呈现出一些独有的特征,这些特性也完整地反映到 Web资源的索引数据上,如内容动态增长、内容累积性、内容海量以及对硬件高性能的需求。为了支持 Web Archive资源的高效访问,Web Archive的索引系统面临着动态性、准确性、可伸缩性、高性能等多方面的挑战。制定合适的索引策略,高效地管理索引数据,从而提高索引性能,成为 Web Archive访问系统的一个非常关键的问题。

Wayback[1]是目前 Web Archive领域中广为使用的存档资源访问系统,它集存储、索引、检索、再现等功能于一体。初始版本为 Wayback Machine,由 Alexa公司受 Internet Archive(IA)委托于 2001年开发完成并投入使用。Wayback Machine由 Perl语言实现,缺乏可维护性和可扩展性,代码也并非开源。其后的 Java版本主要致力于解决这三个问题,从而促进了 Wayback的广泛应用及逐步完善。Wayback的典型应用案例是 Internet Archive(IA),据统计,从 1996年至今 IA已经保存了 850亿的网页,每天收到大约 1000万次点击,每秒钟要处理 100 - 200个

收稿日期:2009 - 04 - 02

点击,每天 10万次左右通过 URL查找,每天 400万次返回请求,但在笔者多次测试中其反应速度良好。如何应用 Wayback来处理海量数据,保证良好的索引、检索性能,是一个非常值得研究的题目。

2　基于 Wayback的索引策略研究

本文对目前基于 Wayback的 Web Archive项目进行了调研,对其所采用的索引策略进行初步归纳和分析。

2.1　基于本地访问的索引策略

基于本地访问的索引策略是 Wayback的一种最为简单、最为基础的索引策略。其基本特征是索引工作与整个访问系统的其他部分都放在同一个服务器上,索引的建立、存储和访问完全在本地服务器上进行。根据被索引文件的存储格式不同,Wayback本地访问策略又可进一步分为本地 Berkeley DB(BDB)资源索引和本地 CDX资源索引[2]。

葡萄牙 Web Archive计划(Portuguese Web Archive Initiative, PWA)就采用了这种策略利用 Wayback对系统中的 ARC文件作 URL索引。其爬行器初始访问的 URLs有 7 200万个,下载 5 600万个(2.8TB),而压缩成 ARC文件后有 2TB。PWA每三个月进行一次新的爬行,因此每年需要 9.12TB的磁盘空间来存储这些数据。从文件数量上看,该项目属于小型 Web Archive项目。为了应对不断增长的数据,该项目利用 Hadoop对这些文件作聚类处理。Wayback的索引工作则直接作用于 Hadoop处理后的文件[3]。

这种策略的优点是系统配置简单易行,在一定的数据规模内效率也比较高。但其缺点是随着集合中数据量增大会相应增加服务器的消耗,一旦超出服务器本身的负载能力,系统将无法运行,也就是说,随着资源的增长使用这种索引策略会相应提高对服务器硬件的要求,在目前的服务器性能状况下,没有哪一种服务器能够应对海量的 Web Archive数据。因此目前这种策略只适用于小规模的 Web Archive集合。

2.2　与访问系统分离的索引策略

除了本地索引策略,常用的策略还包括"与访问系统分离的索引策略"。它将索引服务从整个访问系统中分离出来,使索引与访问成为相对独立的两个部分,并部署在不同的服务器上,索引的建立、维护、存储都在远程服务器上进行,访问系统可以通过一定的访问方式(如 HTTP查询)来调用远程服务器上的索引资源。

IA就此策略提供了两个具体的实施方案[2],远程 BDB或 CDX资源索引,提供 URL检索;远程 Nutch-WAX[4]资源索引,提供 Web Archive的全文检索。图 1所示是 Wayback远程 BDB或 CDX资源实施方案。其中 Wayback资源索引服务(Wayback Resource Index Service)与访问服务(Wayback Service)分离,其资源索引(CDX/BDB)以及 BDB/CDX索引数据都处在这个分离的资源索引服务系统中。而这两种服务采用 XML作为中间格式进行通信。

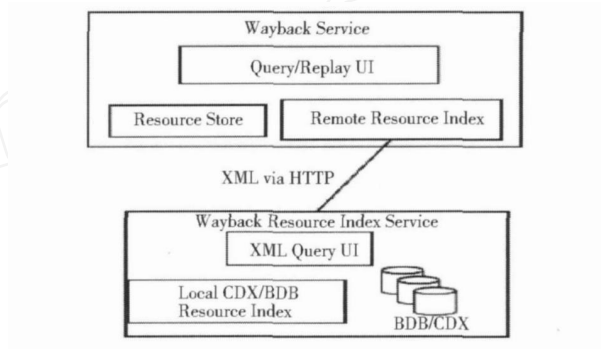

图 1　Wayback远程资源索引[2]

与访问系统分离的索引策略将索引资源的管理和存储与访问系统的其它部件分开进行,从而减少对服务器资源的需求,降低服务器工作负担,在一定程度上解决了 Web Archive索引与访问相冲突的难题。但存在的问题是需要远程通信,从而会相应提高系统对通信设施的配备要求。这种索引策略是一种解决海量数据访问的最常用方法,也是 Web Archive的基本索引策略,是其它复杂索引策略建立的基础,适用于较大规模 Web Archive集合。

2.3　基于负载均衡的索引策略

为了更好地应对 Web Archive索引与访问的冲突难题、分散服务器的负载,可以采用基于负载均衡的索引策略。这一策略的基本思路是将索引服务分成多个子服务,每个子服务都是一个完整的索引系统,同时还可以配备专门的管理设备对这些子索引服务系统进行管理,访问系统通过管理设备来分配和协调索引服务,查询索引数据,以此来降低单个设备的负担,提高索引工作效率,降低索引工作时间。

IA提供了一种按字母顺序分布资源索引的方案[2],这个方案将索引资源按照所存储 URL的字母顺

序划分成多组(如可划分为 A - M,N - Z两组),每组索引资源由一个主机来管理,另外在存档资源与这些索引文件中间配备一个配置文件(Configuration File),通过这个配置文件将索引请求提交给相应的主机,从而提高索引工作效率(见图 2)。

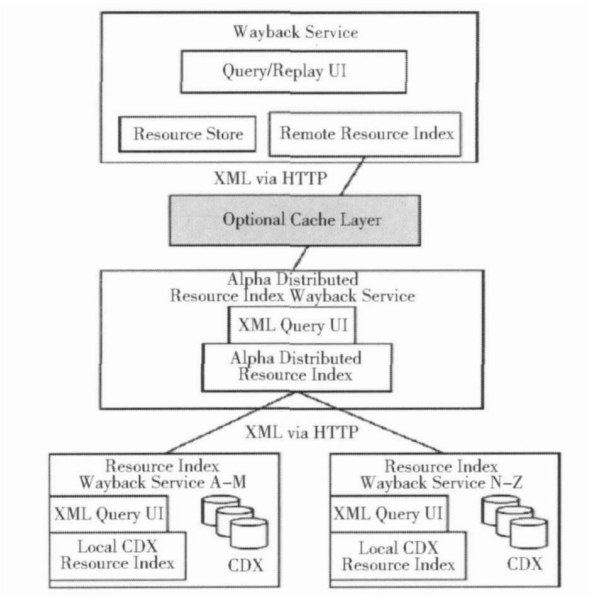

图 2 字母分布式资源索引[2]

圣地亚哥超级计算机中心(San Diego Supercomputer Center, SDSC)在 LOC/SDSC(a Library of Congress/San Diego Supercomputer Center)项目中[5]则采用配置多个 Wayback来分别对系统中的部分资源进行索引的方式来应对这一问题。在此项目中共配备了 18个"Wayback Instance",并相应地生成 18个索引数据库,而单个 Wayback instance维持的索引率约为1 000个文件/天。另外配备一个"主 Wayback"(Master Wayback)作为前端机(Front - end)访问 LOC内的资源(见图 3)。

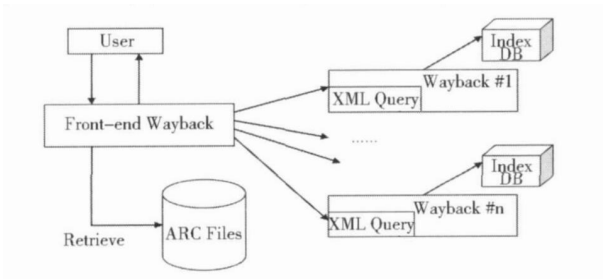

图 3 设备分布式资源索引[6]

基于负载均衡的索引策略的基本特征是索引资源的分散性,通过分散索引资源、部署多个索引设备大大提高了索引工作的效率,减少资源索引的构建时间和索引数据的查询时间,为用户对 Web Archive的快速访问提供了可能,具有很高的索引工作效率。缺点是需要相对复杂的系统部署工作。另外,当后端的资源数量继续增大时,也会面临执行上的问题,因为相对于索引数据存储在一个单一的并行数据库(Parallel Database)的方案,这种方案请求遍历所有的后端索引存储器的效率会比较低。这种方案也同样适用于大规模 Web Archive集合。

2.4 基于分布资源的索引策略

基于分布资源的索引策略原理是,为位于不同地理位置的存储系统中的 Archive资源分别配备索引工具,即为分布式资源部署多个索引设备。

这个策略与"基于负载均衡的索引策略"相类似,不同之处在于前者的存档资源被存放在一个存储设备上,而后者的存档资源被分布存放,因此需要一个资源登记系统记录这些资源的存放位置。采用这种索引策略需要为每部分存档资源各配备一个索引设备,相应生成的索引数据库也与这些资源一起存放。同样,需要有一个前端设备负责这些索引设备的管理和与用户的通信。

其典型应用案例是圣地亚哥超级计算机中心(San Diego Supercomputer Center)在 LOC/SDSC(a Library of Congress/San Diego Supercomputer Center)项目[5]中采用的分布式 Wayback方案(见图 4)。该方案就是针对索引资源的分布性所提出的。为系统配备多个 Wayback分别管理分布的 ARC文件及其相应的索引数据库,并且提供一个资源登记系统(Resource Registry)专门负责记录这些资源的位置。

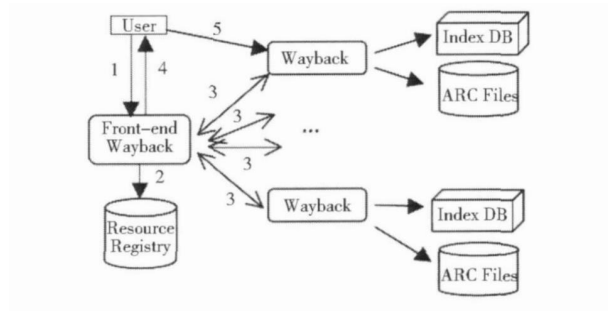

图 4 SDSC全分布式 Wayback Machines实施方案[5]

SDSC的优化方案[4]中还使用了另外一种方案来解决分布式资源的索引问题,即在 Wayback与存储资源中间配备高效的索引数据库(如 Oracle),通过这个

数据库将分布资源产生的所有索引数据都集合在一起进行专门管理,从而进一步提高对索引数据的操作效率 (见图 5)。

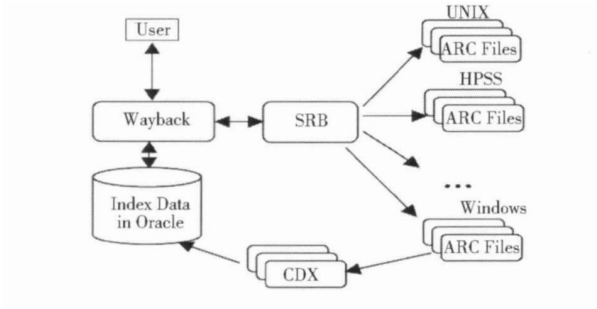

图 5　使用 Oracle数据库和 SRB的索引策略[5]

在以上 SDSC的项目中针对分布资源索引采用了两种具体的解决方案:部署多个索引设备和配备高效的索引数据库。不管是哪种解决方案,都是针对分布式资源。前一种方案的索引效率较高,但同时存在与"基于负载均衡的索引策略"相同的问题,即相对复杂的系统部署工作和面临海量数据时较低的索引效率。另外,当处理分布资源时,将一个大的集合分成多个部分的方法给数据和系统的维护与管理增加了相应的负担。因此,一些项目对该策略做了进一步优化,也就是后一种解决方案 ——配备高效的索引数据库 (如 Oracle),把分布资源产生的所有索引数据都集合在一起进行专门管理,这样不但提高对索引数据的操作效率,更能适用于海量的分布式资源。但这种方案的最大缺点就是配备高效的索引数据库会相应提高系统成本。

2.5　基于异构资源的索引策略

使用 Wayback系统还可以处理存储在异构存储系统中的资源 (如分别存储在 UNIX系统中、HPSS系统或 Windows系统中)。这也就是基于异构资源的索引策略。此时,需要配备专门的网格中间件来整合索引工具与存档资源。网格中间件是一种在网格中应用的中间件,是指一系列协议和服务软件,其功能是屏蔽网格资源层中资源的分布、异构特性,向网格应用层提供透明、一致的使用接口[7]。

图 5是 SDSC提出的一个基于网格中间件的可扩展索引方案[5]。在这个方案中,系统使用网格中间件 SRB连接异构存储系统中的存档文件和 Wayback。SRB负责管理所有存档资源的注册信息和与 Wayback的交互。在这个方案中,SRB为分布在不同地理位置、不同文件系统的存档数据提供了一个虚拟的全球 UNIX Box。

基于异构资源的索引策略是针对异构资源的索引工作而提出来的,其主要特征是在系统中使用了网格中间件。最大的优点是将异构资源的索引整合到一起,但是配备网格中间件也会相应地增加系统配置的复杂性。从目前 Web Archive全球化协作来构成 Web Archive网络的趋势来看,这种利用网格中间件的策略是最为适用的。

3　结　语

Web Archive资源所具有的积累性、海量和动态增长等特点,引发了其索引工作不同于其它资源的问题和挑战。Wayback以其充分的可扩展性在 Web Archive实践中得以广泛应用,并在此基础上形成了多种解决策略。以上笔者分析的几个索引策略,就是研究人员在实践过程中利用 Wayback对于索引效能不断深入探索的结果,这里对这几种策略的特征、优缺点及适用集合做了一个小结 (见表 1)。

表 1　基于 Wayback的 Web Archive索引策略比较表

索引策略	特征	优点	缺点	适用集合类型
基于本地访问的索引策略	索引系统与访问系统在同一服务器上	系统配置简单易行	对服务器硬件要求较高	适用于小规模的 Web Archive集合
基于远程访问的索引策略	索引系统与访问系统不在同一服务器上需远程通信	服务器工作负担小	需要远程通信	大规模 Web Archive集合
基于负载均衡的索引策略	索引资源分散	索引工作效率较高	系统部署工作相对复杂;面临海量数据执行效率较低。	大规模 Web Archive集合
基于分布资源的索引策略 (部署多个索引设备)	索引资源的分布性	索引工作效率较高	系统部署工作相对复杂;面临海量数据执行效率较低。	分布式资源
基于分布资源的索引策略 (配备专门、高效的索引数据库)	高效数据库对索引的管理	索引工作效率高	需配备专门数据库	海量分布式资源
基于异构资源的索引策略	网格中间件的使用	整合异构资源的索引	需配备网格中间件	异构资源

随着 Web Archive索引系统所面临的挑战日益复杂,人们对索引的研究进行得也愈加深入,Wayback的出现及其进一步完善为 Web Archive索引工作提供了强有力的支持工具。目前,以 Wayback为基础的 Web Archive索引策略的实施也由简单到复杂逐渐演变,索引效能亦随之不断提升。从以上的分析不难看出,虽然已有多种策略努力应对海量数据的索引和高效能访问的需求,但随着 Web Archive资源不断增加,内容不断丰富,提高索引工作效率仍是 Web Archive索引工作的研究主题。

参考文献:

[1] About the Wayback Machine[EB/OL]. [2009-02-26]. http://www.Archive.org/web/web.php.

[2] Tofel B. 'Wayback' for Accessing Web Archives[C]. In: *Proceedings of the 7th International Web Archiving Workshop*. Vancouver, Canada, 2007.

[3] Gomes D, Nogueira A, Miranda J, et al. Introducing the Portuguese Web Archive initiative[C]. In: *Proceedings of the 8th International Web Archiving Workshop*. Aaarhus, Denmark, 2008.

[4] NutchWAX[EB/OL]. [2008-05-19]. http://Archive-access.sourceforge.net/projects/nutchwax/.

[5] Minor D, Zhu B, Moore R, Cowart C. Archiving, Indexing and Accessing Web Materials: Solutions for Large Amounts of Data[C]. In: *Proceedings of the 7th International Web Archiving Workshop*. Vancouver, Canada, 2007.

[6] Data Center for Library of Congress Digital Holdings: A Pilot Project[R]. Library of Congress, CACI, San Diego Supercomputer Center and UCSD Libraries, 2007.

[7] 应宏. 网格技术及其应用[J]. *计算机工程与设计*, 2004, 25(10): 1685-1688, 1691.

(作者 E-mail: wuzx@mail.las.ac.cn)