Web Archive检索系统架构分析*

吴振新[1]　向菁[1 2]
[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）

[摘要] 本文以现有 Web archive 项目为案例，初步分析这些项目中所采用的检索系统架构以及它们如何应对在海量数据中快速发现信息、呈现信息的挑战，以期从系统架构的角度来探析 Web Archive 检索系统的性能和效率，为相关研究机构、人员提供参考。
[关键词] Web archive　检索系统　系统架构
[分类号] G250

Analysis of Retrieval System Architecture in Web Archive

Wu Zhenxin[1]　Xiang Jing[1 2]
[1] (National Science Library, Chinese Academy of Science　Beijing 100190,China)[1]
[2] (Graduate University of Chinese Academy of Sciences　Beijing 100049,China)[2]
[Abstract] Based on existing Web archive projects,this paper finishes a preliminary analysis of retrieval system srchitecture that are applied by these projects and how they cope with the challenge which is to search infomation in massive data collection. From the view of system architecture , it discusses archive retrieval system performance and efficiency and wish to provide some references to the relevant institutes and researchers.
[Keywords] Web Archive　Retrieval System　System Architecture

1. 引言

网络信息资源作为人类文化遗产中非常重要的组成部分，受到越来越多的重视，随着全球 Web Archive（以下简称 WA）的开展，大量的网络信息被采集、保存并提供使用。但 WA 资源本身固有的内容累积性、动态变化和增长、海量等特点，使得 WA 资源的检索与访问服务面临着动态性、准确性、可扩展性、高效等多方面的挑战。

对于 WA 这样的大规模复杂信息系统来说，系统结构的设计比起算法和数据结构显得更为重要，系统架构从根本上决定了系统整体的可用性(Availability)、可管理(Manageability)、可信赖(Reliability)、可扩充(Scalability)和安全性(Security)，并对整个系统的性能以及持续地调优起着非常关键的作用。在此种背景下，研究人员充分认识到系统架构的重要性，并通过对系统架构的深入的、系统化的研究来提高系统的性能和效率。

经过多年的研究和实践，很多国家已经建成了诸如澳大利亚的Pandora[1]、加拿大的GCWA[2]、荷兰的e-Depot[3]、葡萄牙的Tumba[4]、美国的LOCKSS[5]、中国的Web Infomall[6]、北欧的NWA toolset[7]等投入实际服务的Web保存系统，还出现了一些专门的访问工具,如Internet Archive的Wayback[8],IIPC的WERA[9]、Xinq[10]等。本文以现有Web archive项目为案例，初步分析这些项目中所采用的检索架构以及它们如何应对在海量数据中快速发现信息、呈现信息的挑战，以期从系统

* 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。

架构的角度来探析Web Archive的系统性能和效率，为界内同行提供参考。

2. WA 检索系统的基本架构

如下图所示，WA检索系统基本架构同一般的Web检索系统一样，由几个基本组件构成：

- 索引点：从“网页数据库”中得到数据进行预处理，依据所使用的信息检索模型对文档进行形式化表示，按照索引策略建立和存储索引，同时负责索引数据的管理和更新。
- 查询点：依据所使用的信息检索模型对用户的查询进行分析，并通过查询索引数据（库）来查找匹配数据文档，同时计算各个文档的相关度，将相关度大于阈值的所有文档按照相关度递减的顺序排列，并返回给用户。
- 分发点：按照查询点所获得的数据文档对象指针从存储设备中获得数据并按一定格式分发给用户。

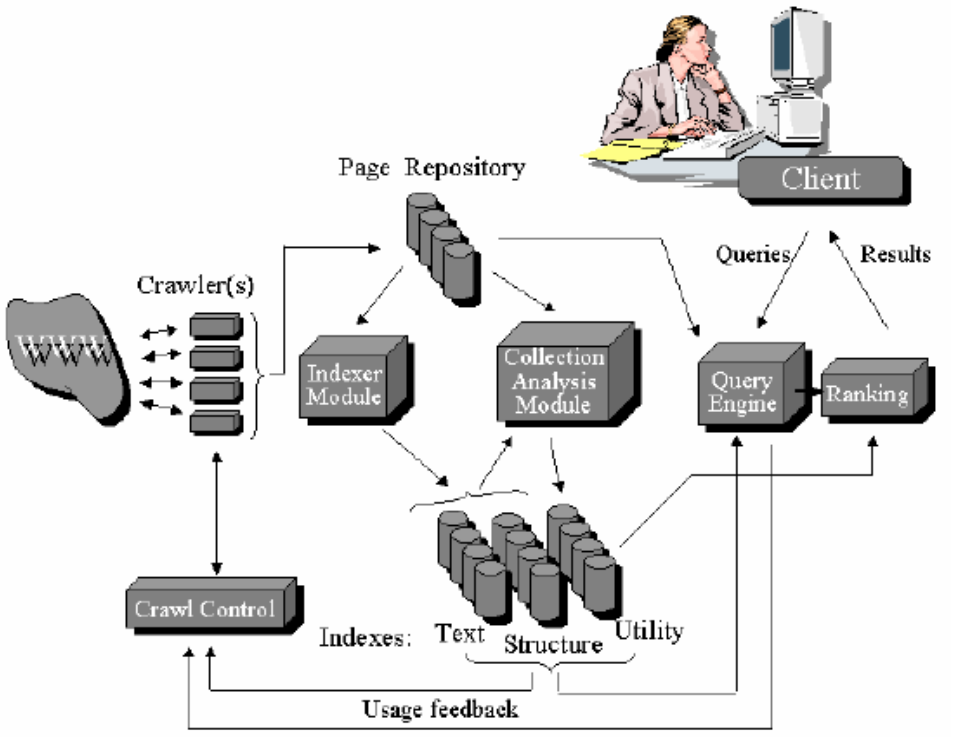

图 1　WA检索系统基本架构[11]

这些组件相互协作组成了一个有机整体。如何在现有的网络环境和硬件条件下，对检索系统的各部分进行合理配置和部署，有效应对数据规模的不断扩大，成为检索系统“效率”必须首先考虑的问题。

3. 现有 WA 检索策略分析

3.1 基于 Broker/Client 的分布式检索架构

基于 Broker/Client 分布式检索机制是利用代理服务器（broker）将查询请求分割为若干个子请求后，再发送给多个查询服务器，并对查询服务器返回的结果进行整合认证后通过 client 客户端呈现给用户使用，从而缩短服务器对检索请求查询响应的时间。

Tumba[12]是由里斯本大学的XLDB研发小组开发，于 2002 年开始使用的一个葡萄牙WA搜索引擎。它的检索点和索引点能够根据需要进行线性扩展，并提供容错功能，而且检索响应时间迅速。

Sidra是Tumba系统中执行索引、检索、排名的子系统，整个检索架构主要包括查询服务器（QueryServer）、代理服务器（Brokers）和客户端（Client）三个部分，通过分布式架构来改善索引的规模化和可扩展性。它的这种灵活的、分布式的检索机制，可以根据不同搜索引擎的要求进行配置部署。

图2显示了Sidra架构的一种配置方案。查询服务器负责匹配每个索引片段与查询式，可同时支持关键词、地点和主题的检索。Sidra的索引可被多个查询服务器按字母顺序分割予以存储（A-H、I-Q、R-Z），可在不同检索维度并行进行检索。代理服务器接收客户端查询请求，运用分布式索引上共享的编目信息和查询服务器位置将查询子请求分发给选择的查询服务器，每个子查询请求与相关文档信息匹配后，再由代理服务器整合从查询服务器返回的结果，从而减少了查询的数量，提高查询效率。其轻量级的客户端通过程序界面与用户互动，客户端响应代理请求，并将数据转化处理呈现给用户使用。

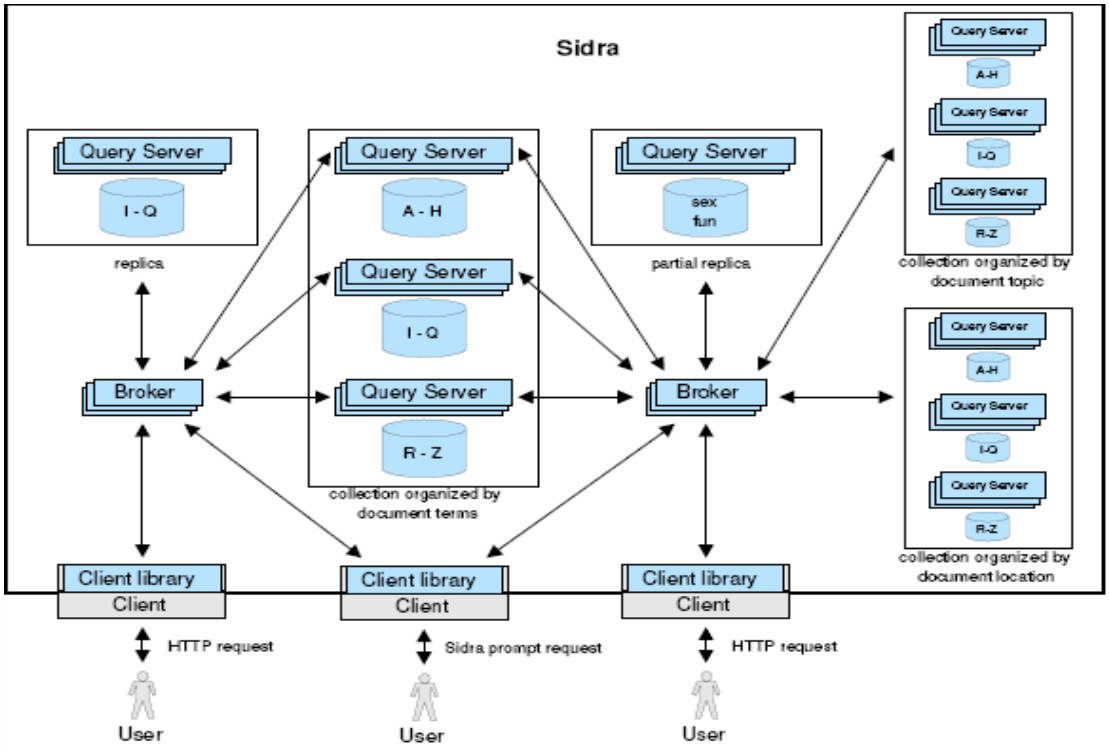

图 2　Sidra架构一种配置方法[13]

通过分析Tumba日志文件可以发现[14]：Tumba搜索引擎能够处理不同规模数据，响应速度快，显示出Sidra架构对大规模数据的良好适应性。据统计：156.8 GB数量文档的检索反应时间为毫秒级，313.6 GB大小文档索引建立时间为 56.38 小时，初步解决了大规模数据检索的效率问题。

该检索机制通过配置多个查询服务器，分散查询压力，缩短反应时间，从而提高检索系统的性能。同时可根据需求对查询服务器进行线性扩展，使得采用该机制的检索系统保持了一定的可扩展性。

3.2　可多维度扩展的分布式检索架构

为了控制不断增长网络信息，传统搜索引擎采取倒排序文件方法来维护、检索索引文档，但此种方法效率不高。为了应对这个难题，早于 Google 的 FAST 搜索引擎构建了一种灵活的、可根据数据体量大小对检索结点和发送结点进行线性扩展的架构，从而实现了可多维度扩展的分布检索机制。

FAST 分布式的架构主要有检索结点（search node）和发送结点（dispatch node） 组成，采用多层发送服务器架构，该架构呈树状结构，便于搜索引擎的扩展，采用允许数据库进行动态变化的异构架构处理网上数据的动态更新，从而提高了大规模数据的索引和检索的效率。检索结点有结点文档数量、查询率（流量容量）两个参数；发送结点有发送功能，通过发送到检索结点进行查询。每个检索结点可单独处理其他检索结点的查询。

多层发送系统有一个超级发送结点，和其它发送点组成树状结构，保证线性扩展能力。检索结点分隔检索查询的数据库，发送结点并行处理所有检索结点的查询，合并检索结点的结果，构建结果集，合并的效果受检索结点接受到查询数目的限制。为提高查询率，发送结点对所有检索结点排成一列，采用循环算法在不同纵列间进行循环（|D|是数据线性扩展增加的大小，Sij 是检索结点 j 保留 i 的划分），并实现不同纵列间可以负载均衡，执行扩展线性能力（见图 3）。

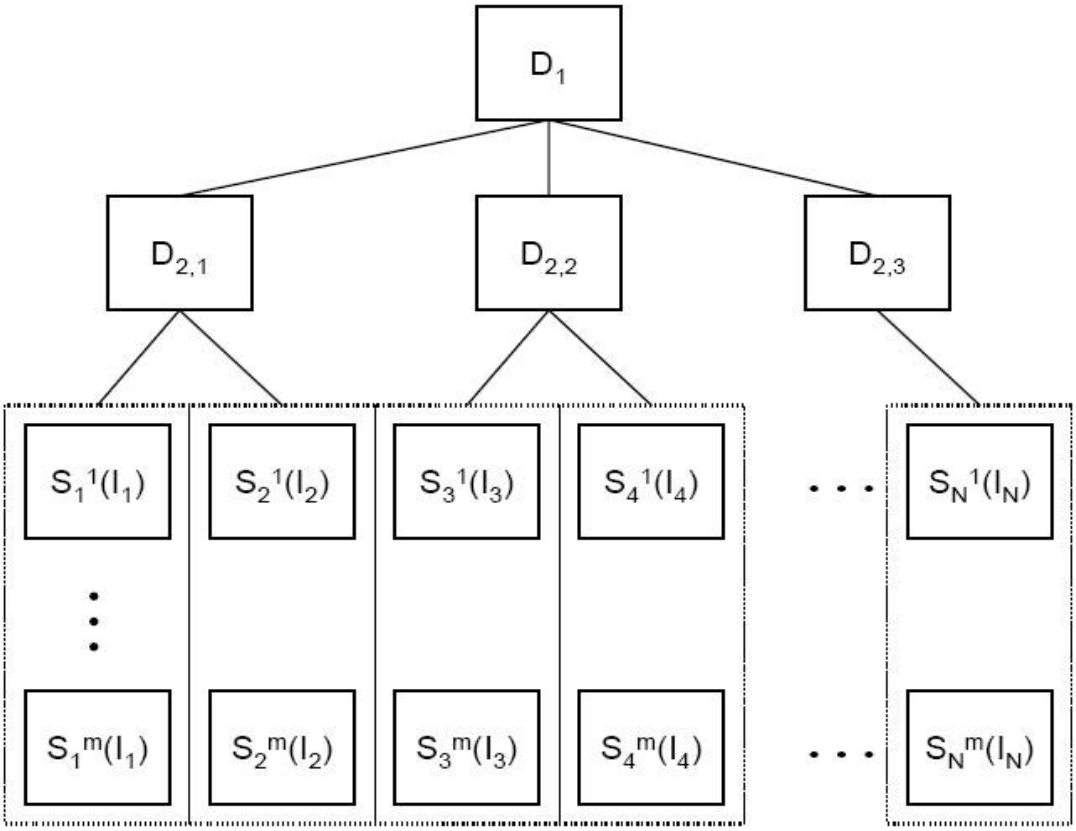

图 3　多层发送系统架构图[15]

FAST 为北欧的 NWA 项目提供了一个可扩展、分布式的高性能的检索工具，在文本检索速度、执行的复杂度方面均有不错的表现，NWA 顶层采用一个适用于任何国家的超级发送结点，同时在挪威、瑞典、丹麦、冰岛、芬兰设置分布式发送子结点，每个子结点下再继续细分为 n 个索引结点（见图 4）。

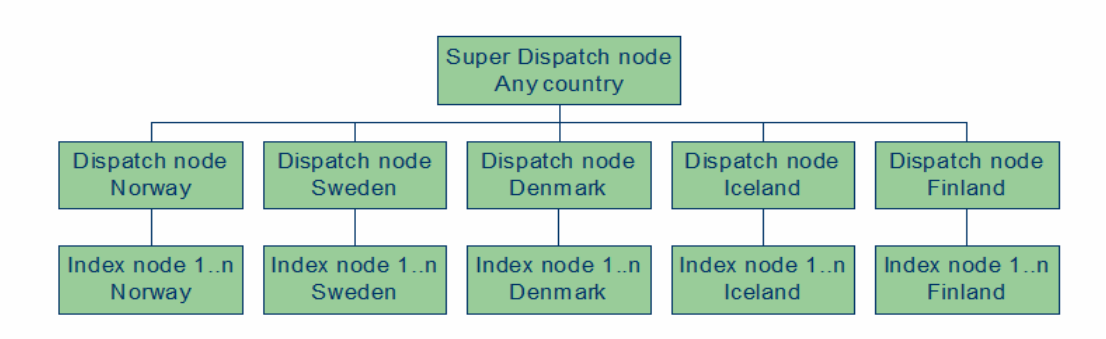

图 4　NWA项目搜索引擎的架构[16]

为同步检索北欧保存的资源，国家发送结点需要作为前端的分布式结点，当用户进入访问界面提出检索请求，访问模块通过前端的国家发送结点向北欧所有的发布结点和索引结点发出请求，得到结果后再将结果返回给访问模块，访问模块再将此结果传递到浏览器供用户浏览。

这种多维度扩展的分布式检索机制可根据需求分别对查询结点（服务器）和索引结点（服务器）进行线性扩展，从而保障整体系统的可扩展性。同时可对信息进行动态更新，但受各结点的容量限制，在内容规模扩展到一定程度，就不能再继续进行信息更新，需要重新划分检索结点并进行索引重建。

3.3　负载均衡的检索架构

负载均衡是大负载网络服务中常用的解决方案，主要思路是按对称方式搭建多台服务器，每台服务器都具备等价的地位，都可以单独对外提供服务而无须其他服务器的辅助。然后通过某种负载分担技术，将外部发送来的请求均匀分配到对称结构中的某一台服务器上，而接收到请求的服务器都独立回应客户机的请求。由于采用多台服务器同时提供网络服务，并将网络请求分配给这些服务器分担，这样就可以提供处理大量并发服务的能力。

为减少 WA 搜索引擎检索反应时间，基于 Web 服务的分布式检索机制将查询请求同时提交给网络上的多个主机，由位于这些主机上的检索程序分别独立检索并将检索结果返回到检索代理程序，经过整理后显示给用户。

加拿大政府网络保存项目[17]（Government of Canada Web Archive: GCWA）为了解决大规模数据的检索访问速度问题，采用了负载均衡的检索机制，通过分散索引资源、部署多个检索设备来提高检索的效率，实现负载均衡，从而有效地实现大规模数据的检索和访问。它为数字对象建立全文索引，并按索引存储的大小平均分配给多个检索服务器，通过分布式的检索服务器实现负载均衡。该架构包括 3 个负载均衡的四核刀片访问服务器，每个刀片访问服务器上建立了 4 个Nutchwax分布式检索服务器，每个检索服务器指向一个索引片断（部署在一台服务器）。全部资源的索引分为数个“碎片（片段）”，每个片段代表约 2500 个ARC文件，每个ARC文件约 100MB大小。当检索页面接收到查询请求时，分布式的访问系统将请求传送给Nutchwax检索控制工具，多个检索控制工具在多台检索服务器中进行检索查询（图 5），从而提供大规模数据检索访问的速度，并体现出良好的性能。使用JavaTM 装载测试框架—Grinder测试搜索引擎的性能和浏览性能的结果显示：每秒能进行 2.5 次检索，每秒可浏览 8 个网页。

该项目的搜索引擎自 2007 年秋季发布以来，加拿大国家图书馆已经提供有关政府网站 1000 万数字对象（超过 4TB）的访问。

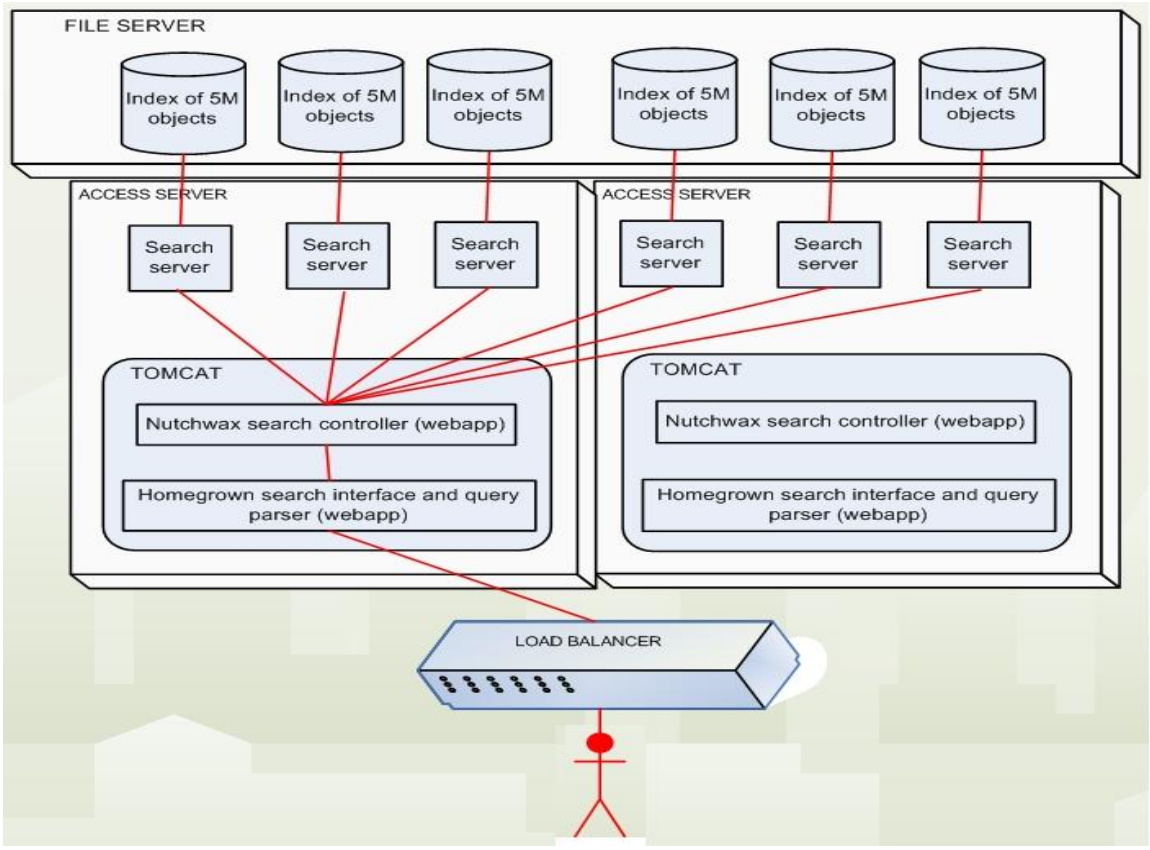

图 5　加拿大政府网络保存项目索引及查询机制[18]

3.4　有效利用缓存的分布式检索架构

“中国Web信息博物馆”(简称Web Infomall)[19]是在国家 973 和 985 项目支持下，北京大学网络实验室开发建设的一个大规模的Web仓储系统，长期保存中国范围内的网页，并在此基础上研究对历史网页进行整理归档、分析挖掘与展现回放的方法和技术。目前已经维护有 30 亿以中文为主的网页，并以平均每月四千五百万网页的速度扩大规模。

Web Infomall整合了天网搜索引擎（简称天网）作为它的资源采集器和检索访问系统，通过天网获得包含时间信息的查询，因此天网不但是一个典型的Web搜索引擎，也成为WA中具有代表性的搜索引擎[20]。

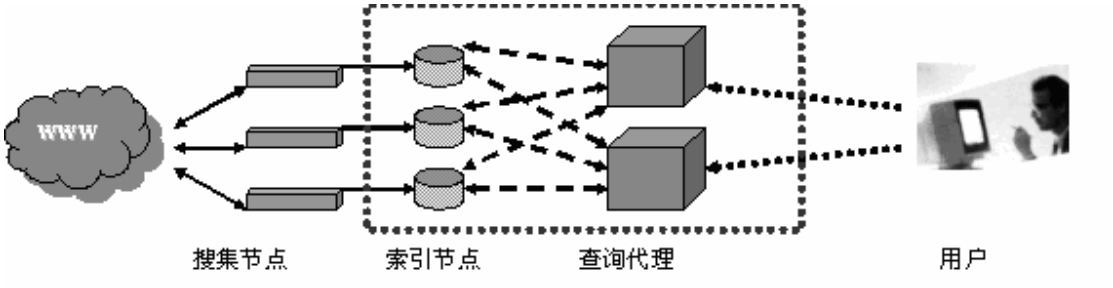

图 6 天网分布式结构模型[21]

在天网检索系统中，包括检索结点、查询代理结点和文档服务结点。文档服务节点和查询代理结点使用同一服务器;索引节点和查询代理之间的工作模式属于典型并行算法—主从方式（Master/worker)；查询代理节点(Query Broker)通过多播向所有索引节点发送查询命令，由索引节点并行完成查询并返回结果，

查询服务器上的 Retrieval Agent 负责结果数据的合并，完成访问文档服务，并格式化结果页面返回给用户，用户的后续查询（翻页），将会在缓存命中，不必再次启动查询，这将大大降低查询系统的负载，从而提高查询系统的性能。

从数据在索引节点的分布来看[12]，属于无共享数据分布模式（shared-nothing data distribution），每个节点维护一个互不相交的数据子集。天网利用多台独立计算机实现的分布式算法本质上是分割索引数据，优点在于更好的系统可扩展性。现在运行的检索服务系统共使用20台PC(PIII733/1GB)，其中一台为查询服务器，其余19台为索引服务器。

同时研究人员还提出了一个进一步提高性能的方案：将原来的二级结构变成三级结构—在中间加一层缓存服务器(QueryCache)，它也可以是由多台机器组成的集群。QueryCache 负责接收前段查询代理节点的查询命令，首先在本机的结果缓存中查找，对于没有命中的查询才向索引数据节点发送查询命令。查询代理节点按一定策略将用户查询（比如对查询词做 Hash）映射到不同的 QueryCache，这样可以有效地分布负载。QueryCache 搜集用户的查询词，对用户经常查询的单词作管理，利用自身的大容量磁盘缓存它们的查询结果。三层结构是对原来分布式系统结构的自然扩展，可以满足大量负载下的性能要求。

4. 结语

海量网络资源的存在对 WA 系统的检索性能提出越来越高的要求，从上面的分析我们可以得到一些启示：

- 检索系统是由多个部分相互协作组成的有机整体，所谓有机是指其不同的组合和不同的协作会产生不同的效果，所以一个好的系统架构会影响整个系统的成败。
- 检索系统本身复杂的结构和目前复杂的运行环境决定了整个系统必须以分布式实现。
- 资源的动态变化要求系统必须可以动态变化，可动态添加和删除各结点服务器，即需要保持动态扩展能力。
- 以提供同样功能的机器组成集群的方式分担压力，提高性能。这种方式可以允许我们采用低价的 PC 机，有效地降低系统成本。
- 以上的几个实例都从不同角度利用了负载平衡的原理，如在查询结点、索引结点和分发结点，以及它们之间的协作，都有不同程度的应用，负载平衡成为 WA 检索系统架构中很重要的特征。
- 通过保证查询结点、索引结点、分发结点的线性扩展能力，保障检索系统整体性能的（多维度）可扩展性。但索引结点的线性扩展是一种伪扩展，由于各结点服务器受容量限制的，在内容扩张到一定规模，需要对索引重新划分并重建索引才能实现索引结点的线性扩展，所以一定要根据结点容量和内容的扩展速度及可能的规模，对索引结点进行很好的规划。

目前 WA 检索系统在大规模数据索引、访问、检索质量控制、数据挖掘、智能检索方面正在不断予以试验。如何利用新技术来提高 WA 搜索引擎的存储性能、系统性能、检索性能，提高检索结果的准确率和全面性实现高效检索，并进行相应的数据挖掘用于学术研究、追踪动态等 WA 长远发展问题，都是未来 WA 检索技术发展的重要关注点。

参考文献

[1] Pandora[EB/OL].[2008-07-27]. http://pandora.nla.gov.au/.
[2] Government of Canada Web Archive[OB/OL].[2008-08-10]
http://www.collectionscanada.gc.ca/whats-new/013-315-e.html.
[3] e-Depot[EB/OL].[2008-04-7]. http://www.kb.nl/dnp/e-depot/e-depot-en.html.
[4] Tumba[EB/OL].[2008-07-27].http://www.tumba.pt.
[5] LOCKSS[EB/OL].[2008-06-2]. http://www.lockss.org/.
[6] Web Infomall[EB/OL].[2008-09-24]. http://www.infomall.cn/.
[7] NWA toolset[EB/OL].[2008-05-6]. http://nwatoolset.sourceforge.net/.
[8] Wayback [EB/OL].[2007-11-3]. http://archive-access.sourceforge.net/projects/wayback/.
[9] WERA[EB/OL].[2007-11-3]. http://archive-access.sourceforge.net/projects/wera/.
[10] Xing[EB/OL].[2007-11-3]. http://www.nla.gov.au/xinq/.
[11]赵江华.“天网”高性能分布式检索系统的设计与实现[D].北京大学，2002.
[12]同 4
[13] SIDRA:a Flexible Web Search System[R/OL].[2008-07-17].
http://www.di.fc.ul.pt/tech-reports/04-17.pdf.
[14] Daniel Gomes.Web Modelling for Web Warehouse Design[R/OL].[2008-08-17]
[15] Search Engines and Web Dynamics[J/OL]. [2008-07-27]
http://www.idi.ntnu.no/~algkon/generelt/se-dynamicweb1.pdf.
[16] Sverre Bang. The Nordic Web Archive[OB/OL].[2008-08-05]
http://www.deflink.dk/upload/doc_filer/doc_alle/1023_SBA.ppt.
[17] 同 2.
[18] Gillian Cantello, JOHN STEGENGA . Government Web content in Canada
A national library web archive perspective [J/OL].[2008-08-10]
[19] 同 6
[20] Li Xiaoming,Zhu Jiaji.Some Characteristics of Web Data and their Reflection on our Society: an Empirical Approach[R/OL].[2008-09-07]
http://www.law.gmu.edu/nctl/stpp/us_japan_pubs/internet_IICIID.pdf
[21] 同 11

(作者:E-mail:wuzx@mail.las.ac.cn)