基于云计算模式的图像检索研究

韩法旺

（南京森林警察学院 信息技术系，江苏 南京 210046）

摘　要：以Web2.0技术为代表的现代技术快速发展及应用，加速了网络信息容量的膨胀。如何从如此庞大的信息源筛选出用户所需的信息，尤其是图像信息，则必须对这些信息进行高效地检索。图像检索传统算法上的改进难以解决海量数据存储、计算及传递等一系列问题，云计算作为一种新兴的计算模型，对解决图像检索发展遇到的瓶颈有着极其重要的推动作用。

关键词：图像检索；海量数据；云计算；模式

中图分类号：G350　　**文献标识码：**A　　**文章编号：**1007-7634(2011)10-1534-05

Reseach on Image Retrieval Based on Cloud Computing Model

HAB Fa-wang

(Department of Information Technology,Nanjing Forest Police College,Nanjing 210046*,China)*

Abstract: The rapid development and application of modern technology,such as Web2.0,accelerate the expansion of the network information capacity. To select the necessary information from a huge source, especially the image information ,we must retrieve the information efficiently.The traditional methods of image retrieval have diffcult to solve solve the problem of mass data storage, calculation,transition and others. Cloud computing as a new computational model has an extremely important role to solve the bottleneck in the development of image retrieval.

Keywords: image retrieval; mass data;cloud computing;model

图像检索就是根据对图像内容的描述，在目标图像集合中找到具有制定特征或包含指定内容的图像[1]。从20世纪70年代开始，有关图像检索的研究就已开始，当时主要是基于文本的图像检索技术（TBIR），本描述的方式描述图像的特征90年代以后，出现了对图像的内容语义，如图像的颜色、纹理、布局等进行分析和检索的图像检索技术，即基于内容的图像检索（CBIR）技术[2]。

随着图像检索技术的迅速发展，传统算法上的改进难以解决图像存储、数据计算以及数据传递等一系列问题。云计算作为一种在网格计算基础上发展起来的新兴计算模型，因其强大的计算能力和海量的存储空间，就是否能推动图像检索技术的发展和提高图像检索的性能已经引起了广泛的关注，本文就云计算模式下图像检索服务的系统结构和研究内容作了探讨。

1　图像检索技术与系统介绍

1.1　基于文本注释的图像检索

基于文本注释的图像检索是指在关系数据库中

收稿日期：2011-03-24

作者简介：韩法旺(1972-)，男，江苏南京人，讲师，主要从事计算机网络安全、信息检索、信息资源管理研究.

加入描述图像内容的字段,并在图像的存储路径和这些字段之间建立联系,然后利用数据库的查询功能实现图像检索[3]。然而不同的检索系统采用各自的注释结构与关键字段,缺乏描述图像的统一方案,资源共享程度低。为了在互联网方式下解决图像资源检索的问题,1995年3月在都柏林召开的第一届元数据研讨会上产生了一个简单的元数据集—都柏林核心元数据集[4],该元数据集规定了可使用的数据内容和数据格式,为在互联网下对图像进行描述提供了可能。

1.2 基于内容的图像检索

基于内容的图像检索是在系统可以自动识别视觉内容的基础上实现的,可用于特征索引的视觉内容包括包括图像纹理、颜色、空间结构等。对于图像特征的提取算法,国内外学者已作了大量研究。针对图像直方图特征,章毓晋[5]和刘忠伟[6]提出了累加直方图和局部累加直方图的方法用于图像检索;Zachary[7]在Lab颜色空间上建立了直方图特征用于图像检索。针对图像空间特征方面,张磊[8]提出了对量化后的色彩计算质心的方法来描述颜色空间特征;Tao等人[9]则利用计算几何中的Delauney三角剖分[10]的方法来描述像素的空间关系。针对图像纹理特征,Tamura等人[11]提出了纹理特征的表达,Rickner[12]等人在著名的QBIC图像检索系统中使用了Tamura纹理用于图像检索。

1.3 现有图像检索系统

互联网的图像搜索引擎如Google、yahoo、百度等采用的是基于文本注释的检索方式。基于文本注释的图像检索方式与用户认知图像的方式一致,因而查准率较高,但由于图像注释采用人工完成,工作量大,主观程度高,容易造成图像注释的歧义。基于内容的图像检索系统主要有QBIC、Virage、PhotoBook、Piction等,此类系统自动提取图像特征建立特征索引,效率高,通用性好,但低级图像特征对图像的描述往往与用户对图像的描述存在较大的差异,检索往往得不到满意的结果,并且图像资源受到数据库的限制难以及时更新满足用户的需求。

综合采用以上两种检索方法可以达到较好的检索效果,但面临图像存储、数据计算以及网络传递等一系列问题,传统算法上难以改进。云计算理念的提出,将在软件环境、应用平台、数据共享方面改善图像检索发展如今所面临的困境,下面就云计算模式下图像检索的系统结构及研究内容加以探讨。

2 云计算对图像检索发展的推动

2.1 云计算的概念

目前云计算系统没有统一的定义,云计算供应商根据自己企业业务推出相关的云计算战略。Hewitt[14]认为云计算系统主要是将信息永久地存储在云中的服务器上,在使用信息时只是在客户端进行缓存。客户端可以是桌面级、笔记本、手持设备等。Wang Li-zhe等人[14]从云计算系统应该具有的功能角度给出了科学云计算系统的定义,指出计算云系统不仅能够像用户提供硬件服务Haas,软件服务Saas,数据资源服务Daas,而且还能够向用户提供能够配置的平台服务Paas。因此用户可以按需向计算平台提交自己的硬件配置、软件安装、数据访问需求。综合来看,云计算就是通过将计算任务分布在由许多计算机组成的资源群上,使各种应用系统能按需取得计算力、存储空间及各种软件服务。

2.2 云计算模式下的图像检索

云计算模式下,将互联网中众多图像资源数据库组合成云资源群,形成一个利用率高、计算速度快的信息检索服务系统。实现云计算模式下的图像检索主要从三个方面着手。

(1) 对于系统端来说,需要建立基于云计算的海量数据存储模型和计算模型,以便将大量的存储任务和计算任务分布在云计算网络下的服务器或客户端下,达到扩大资源共享范围和提高运算速度的目标。

(2) 对于用户来说,检索系统需要提供标准的客户端检索应用程序,用户利用该程序可以方便快捷的从云资源库中检索到所需的图像。同时若用户愿意,应用程序还需提供让用户将自身图像资源上传到云资源库的功能,以供其他用户检索使用。

(3) 云计算模式下需建立统一的检索标准和资源管理机制,若缺乏这一标准和机制,云资源库难以响应不同客户端的检索请求,不同服务器或数据库的资源由于数据存储和传递协议的不同无法共享。

2.3 云计算模式下图像检索系统结构

通过对云计算模式下如何实现图像检索进行分析,可以将云计算模式下图像检索系统结构分为三

个层次，如图1所示。

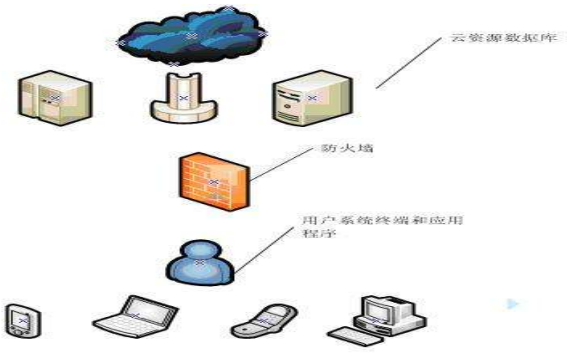

图1 云计算模式下图像检索系统结构

(1)云资源层。这一层主要由大型机、服务器及分布在云网络下的各个客户端机器组成，所有的图像信息资源存储、计算及传递都在此层上进行。

(2)防火墙层。此层用来保证图像检索系统数据信息的安全及用户信息的保密性，防止黑客入侵。

(3)用户层。用户层包括系统设备终端和应用程序。终端设备包括PC机、笔记本电脑、PDA、数字电视、手机等，应用程序是系统提供的标准检索程序，用来提交用户的检索需求。

3 云计算模式下图像检索的研究内容

3.1 海量信息的存储

云计算模式下的图像检索与传统模式下的图像检索最大的不同即是存储空间不受单一服务器容量的限制，用户可在整个云资源库中检索所需的图像。云计算模式下的图像检索系统需要同时满足大量用户的需求，并行地为大量用户提供检索服务。因此，云计算模式下的数据存储技术必须满足海量数据的高传输率和高吞吐率。

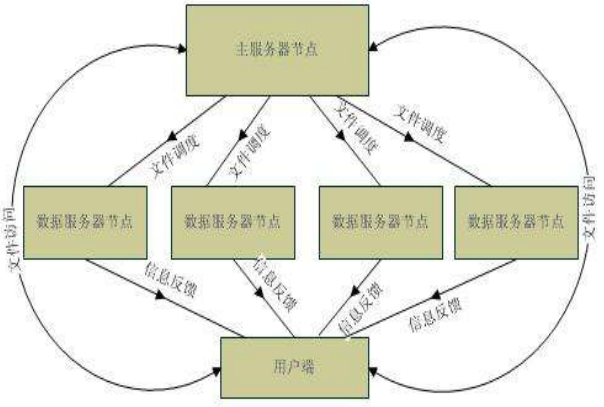

图2 HDFS系统架构图

目前，云计算的数据存储技术主要有谷歌的GFS(Google File System)[15]和Hadoop开发团队开发的HDFS(Hadoop Distributed File System)[16]。以HDFS为例，HDFS作为分布式文件系统采用三层架构，由主服务器、数据服务器和用户端组成[17]，如图2所示。用户通过主服务器向数据服务器发送数据请求，而数据服务器将取得的数据信息直接返回给用户端。HDFS作为一个分布式文件系统，适用于云计算模式下的图像检索系统主要从以下几方面考虑。

(1) 云计算模式下的图像信息资源分布在网络的各个服务器中，当一个或几个服务器出现故障不能读取或存储数据时，图像检索系统需维持正常工作状态。HDFS文件系统有着高容错性的特点，并且用来设计部署在低廉的硬件上，单个服务器的损坏或数据的丢失不会影响到检索系统正常工作。

(2) 图像信息可以看作一个二维矩阵，相对传统的文本信息，程序处理的数据计算量和传递量显著增加，HDFS文件系统提供高传输率来访问应用程序的数据，适合那些有着超大数据集(large data set)的应用程序[18]。

(3) 云计算模式下的图像检索系统需要及时更新应用程序，以满足用户的更多的检索需求。HDFS作为一个开源的文件系统，有着良好的软件拓展性和兼容性。

3.2 海量信息的并行计算

云计算模式下的图像检索系统，不同物理地址下的用户可能同时在使用检索服务，而云计算环境下的数字图像资源范围广、量大，故系统海量数据的并行计算技术是一个非常重要的研究问题。

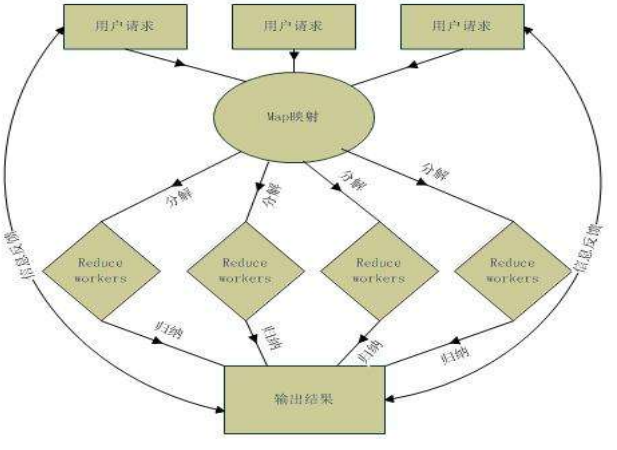

图3 Map-Reduce工作流程图

目前云计算大部分采用Map-reduce模式来实

现海量数据的并行计算。Map-Reduce是一种编程模型,适用于大规模数据集的并行运算。它将一个任务分解成多份子任务,这些子任务在空闲的处理节点之间被调度和快速处理之后,最终通过特定的规则进行合并生成最终的结果,其处理模型有点类似于传统编程模型中的分解和归纳方法[19]。Map-Reduce的工作流程图如图3所示。

云计算下的图像检索系统采用Map-reduce程序结构,检索大体需要以下步骤。

(1) 用户向客户端应用程序提交图像检索请求;

(2) 检索系统响应并将用户请求上传至云计算网络进行任务分解,如不同用户使用不同节点,或者对同一用户的检索请求进行分解;

(3) 系统将分解后的任务发送至各个Reduce-workers节点同时进行工作;

(4) 系统将各个节点完成的任务进行归纳并产生检索到的图像反馈给用户。

Map-Reduce模型具有很强的容错性[20],当workers节点出现错误时,系统会自动将该节点屏蔽并将任务转移到其他节点完成,不会影响到检索任务的进行,因此Map-Reduce模型具备在云计算模式下完成图像检索系统中海量数据的并行计算功能。

3.3 海量信息的管理

在云计算模式下,图像资源不受单一数据库空间的限制,容量显著增大,系统对海量信息数据进行处理、分析以便用户在规模巨大的资源中高效的检索用户所需的图像。云计算模式下的数据管理技术中最著名的是Google提出的Bigtable数据管理技术[22]。

Bigtable采用列存储的方式,提高数据读取效率如图4所示[21]。

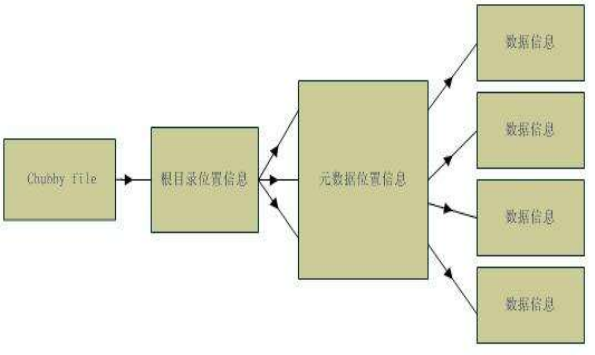

图4 Bigtable体系结构图

其中第一级的Chubby file中包含了数据根目录的位置,根目录包含所有元数据目录的位置信息,每个元数据目录又包含具体数据存放的服务器位置信息。在云计算模式下的图像检索系统中,用户检索图像时首先从Chubby file中获取图像根目录的位置信息,并从根目录中读取图像元数据的位置信息,接着从元数据位置目录中确定图像存储的服务器位置,并根据此位置信息到服务器读取图像。

Bigtable数据管理技术至今已运行了5年,基本上能够满足海量数据管理的管理需求,处理海量数据,实现高速存储与查找。目前,Bigtable的应用包括 Google Analytics、Google finance、Writely、Google earth等60多个项目[22]。

4 云计算模式下图像检索系统特点及存在的问题

4.1 云计算模式下图像检索系统的优点

(1)无限制的数据存储。云计算是基于互联网的超级计算模式,由于数据是存储在云里,不再受到单一数据库存储容量的限制。

(2)计算性能的提高。图像检索相对于文本检索其计算量显著增大,在云计算环境中,各个分布式的网络计算中心可以同时进行数据计算处理,可以显著提高计算速度和准确性。

(3)经济性的提高。传统图像检索系统数据库对数据存储容量和计算性能要求很高,需要花费很多的金钱去购买硬件设备。在云计算环境中,各个计算中心分别承担了计算和储存任务,降低了成本。

(4)数据可靠性的提高。个人电脑的意外损坏或者数据的丢失不会影响检索功能的实现,因为所有图像数据和软件服务全部储存在云中,用户只需连上互联网即可获得所需的服务。

(5)信息的及时更新。传统图像检索数据库,图像存储发生变化时难以做到自动更新或者获取最新的位置信息。而在云计算环境中,图像数据发生改变或存储位置发生变化时,系统可以及时跟踪到最新的信息并反馈给用户,避免数据的丢失。

4.2 云计算模式下图像检索系统存在的问题

虽然云计算理念对解决传统图像检索系统存在的问题有着显著的作用,但仍存在一些问题需要我们去解决。

（1）检索需求标准。检索需求是用户表达个人信息需求的一种方式。不同的人对同一图像的理解各异，带来的问题是：图像检索者可能关注图像不同层次的信息，即便对同一层次也可能关注不同的信息类别[23]。因此，各个图像检索系统之间的检索标准如何制定，不同的标准该如何协调都需要进行进一步的研究。

（2）系统之间的互操作。不同图像检索系统之间的互操作是必须要考虑的一个问题。云计算模式下图像检索系统最大的优点在于信息资源的丰富，当一个系统需要使用另一个系统的资源时，要能够提供跨云的操作方法，使得检索系统之间能够交互。

（3）系统安全问题。包括用户数据安全性和保密性、数据存储安全、用户权限安全、访问控制管理等。

5 结 语

本文借助云计算理念和模型，就解决当今网络环境下图像检索系统的海量数据信息存储、计算及管理等问题作了详细阐述，通过分析证明云计算模式下的图像检索系统能够比较好的解决图像检索发展中所遇到的问题。随着图像检索领域对云计算技术的关注，图像检索的发展将进入一个崭新的阶段。

参考文献

1 高 文，刘 峰，黄铁军，等.数字图书馆—原理与技术实现[M].北京:清华大学出版社，2000:86-86.

2 石 军，常义林.图像检索技术综述[J].西安电子科技大学学报（自然科学版），2003,(4):486-491.

3 Chang S K, Yan C W , Dimitroff D C,et al. An Intelligent Image Database System[J].IEEE Trans on Software Eng , 1988 , 14 (5) :412-421.

4 Rui Y, Huang J S. Image Retrieval:Current Techniques,Promising Directions,and Open Isues[J] .Visual Communication and Image Representation, 1999, 19 (1) : 39-62.

5 章毓晋.基于内容的视觉信息检索[M].北京:科学出版社，2003:234-236.

6 刘忠伟，章毓晋.利用局部累加直方图进行彩色图象检索[J].中国图象图形学报，1998,3(7):532一537.

7 Zachary J M. An Information Theoretic Approach to Content Based Image Retrieval[D].USA:Louisiana State University, 2000.

8 张 磊.基于内容的图像检索中人机协同问题的研究[D].北京:清华大学，2001.

9 Tao Y, Grosky WI.Spatial Color Indexing Using Rotation, Translation and Scale Invariant Anglograms[J].Multimedia Tools and Applications, 2001,15(3):247-268.

10 周培德.计算几何一算法分析与设计[M].北京:清华大学出版社，2000:169-171.

11 Rao A B,Sribari RK,Zhang Z F.Spatial Color Histogram for Content-Based Retrieval[A].Proceeding of 11th IEEE International Conference on tools with artificial Intelligence[C]. Chicago:Illinois，USA，1999.

12 Tamura H,Mori S,Yamawaki T.Texture features corresponding to visual perception[J].IEEE Transactions on Systems, Mans and Cybernetics,1978,8(6):460-473.

13 HEWITT C.ORGs for scalable,robust privacy-friendly client cloud computing[J].IEEE Internet Computing,2008,12(5): 96-99.

14 WANG Li-zhe,TAO Jie,KUNZE M. Scientific cloud computing:early definition and experience[C].Proc of the 10th IEEE International Conference on High Performance Computing and Communications,2008.

15 GHEMAWATS,GOBIOFFH,LEUNGPT.The Google file system[C].Proceedings of the 19th ACM Symposium on Operating Systems Principles.New York:ACM Press,2003:29-43.

16 Apache Hadoop.Hadoop[EB/OL].http://hadoop.apache.org/, 2010-12-21.

17 Yahoo.Yahoo!Hadoop tutorial[EB/OL].http://public.yahoo.com/gogate/hadoop-tutorial/start-tutorial.html,2010-12-21.

18 Dean J, Ghemawat S. Distributed programming with Mapreduce[C].In:Oram A, Wilson G,eds.Beautiful Code.Sebastopol: O'Reilly Media,Inc,2007.

19 Zaharia M,Konwinski A,Joseph A D.Improving Map-Reduce performance in heterogeneous enviroments[C].Proceedings of the 8th USENIX Symposium on Operating System Design and Implementation.New York:ACM Press,2008.

20 Buyya R,Yeo C S,Venugopal S,et al.Cloud computing and emerging IT platforms:Vision,hype,and reality for delivering computing as the 5th utility[J].Future Generation Computer Systems,2008,25(6):599-616.

21 Chang F,Dean J,Ghemawats,et al.Big table:A distributed storage system for structured data[J].ACM Transactios on Computers Systems,2008,26(2):1-26.

22 吴吉义，傅建庆，张明西，等.云数据管理研究综述[J].电信科学，2010,(5):34-41.

23 曹 梅，朱学芳.图像检索需求描述的研究进展[J].现代图书情报技术，2009,(12):31-36.

（实习编辑：杨 洋）