

基于云计算架构的学术搜索引擎研制与应用*

□ 万齐鸣 / 北京万方数据股份有限公司 北京 100038

摘要: 文章通过传统搜索引擎与学术搜索引擎对比,分析传统搜索引擎在云计算技术应用方面的经验。结合863课题研究,剖析学术搜索引擎的研制机理、云计算技术应用特点、学术搜索专业特色,提出基于云计算架构学术搜索引擎的中国学术搜索网云服务应用模式。

关键词: 云计算, 学术搜索, 搜索引擎, 云服务, RMSCloud

DOI: 10.3772/j.issn.1673—2286.2013.06.005

1 引言

2011年,北京万方数据股份有限公司(以下简称万方数据)承担了国家高技术研究发展计划(863计划)“云计算关键技术与系统(一期)”项目“以科技文献服务为主的搜索引擎研制”课题(编号:2011AA01A206),研制中国科技文献学术搜索引擎,并免费为社会公众用户提供学术搜索服务。

万方数据原有的RMS非结构化数据库系统具备全文检索功能,可以针对科技文献提供学术搜索。然而随着科技文献元数据仓储资源量的快速增加、用户检索和访问量的不断增大,万方数据RMS全文检索系统单服务器执行的检索技术和业务逻辑,已经不能满足海量数据大用户量、大并发量快速检索、多维知识挖掘分析和即时响应的需求。

如何实现大数据环境下的学术搜索与知识挖掘分析,就成为863课题研究和万方数据技术发展亟待解决的问题。

2 传统搜索引擎与学术搜索引擎对比分析

针对863课题研究任务要求,以及万方数据知识服务平台业务需求,万方数据首先针对传统搜索引擎应用关键技术、学术搜索技术特点进行了调研分析。

传统搜索引擎,如Google、Baidu等,一般以网页搜索为目标,从互联网成千上万的网站上通过网

络蜘蛛抓取网页,针对网页文字内容进行索引,并提供搜索终端为用户提供搜索服务,用户通过输入关键词,针对索引进行检索,将检索结果呈现给用户。

传统搜索引擎的技术架构如图1所示。

传统搜索引擎直接对于网页文本进行索引和检索,其面对的网页数量动辄几十亿甚至上百亿条,同时面向的用户检索请求瞬时可能达

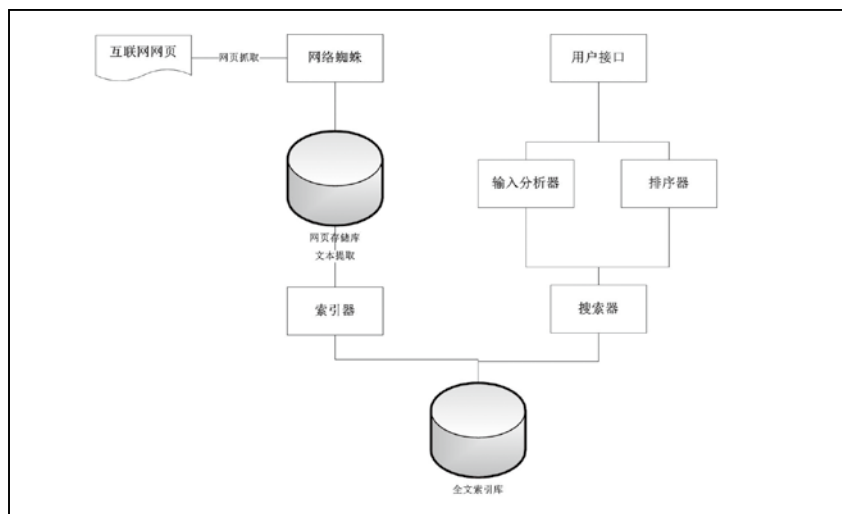


图1 传统搜索引擎的技术架构

* 本文系国家高科技发展计划(863计划)“云计算一期”重大专项课题“以科技文献为主的搜索引擎研制”子课题(编号:2011AA01A206)成果之一。

到几万个。传统搜索引擎虽然并不刻意追求搜索的专业化和精准检索,也不对搜索数据进行结构化处理和呈现,但为满足海量数据检索、大用户量、大并发量快速检索响应的需求,传统搜索引擎大量应用包括分布式、并行计算、负载均衡与调度、分布式文件系统等云计算技术。

学术搜索引擎

则主要针对学术文献,包括期刊论文、学位论文、会议论文、专利、标准、成果、政策法规、电子图书等科技文献元数据信息及全文等进行搜索,这些元数据或全文信息来源于多种途径,可能是从互联网通过网络蜘蛛采集,可能是通过OAI协议等进行收割,也可能是通过其他数据格式进行转换和整合,这些数据不同于传统搜索引擎的网页文本数据,为满足学术搜索查全率、查准率以及多维知识挖掘分析的需求,需要按照元数据标准进行数据拆分提取、深度加工标引、规范性

检查、数据质量检查,将其转换为规范的结构化元数据及非结构化全文,再对其进行索引,提供学术搜索服务,以满足学术搜索的专业化精准检索和多维度知识挖掘分析需求。

学术搜索引擎技术架构如图2所示。

学术搜索引擎面对的是规范化的科技文献元数据和全文信息,数据量比传统搜索引擎要小得多,但由于需要考虑其专业检索、海量数据的多维度知识挖掘分析需求,其对于数据质量、数据加工标引深度

则提出了较高要求,同时其对于瞬时并发性检索和快速响应要求并不比传统搜索引擎要低,甚至在某一时刻,由于其对于大的检索结果集按照不同维度聚类统计分析需求,其对并发性的检索聚类请求响应的性能要求还要超出传统搜索引擎,对于分布式、并行计算等云计算技术的应用则提出了更高的要求。

传统搜索引擎与学术搜索引擎在并行计算为主要特征的云计算应用方面的根本差别如图3所示。

举一个例子,传统搜索引擎在

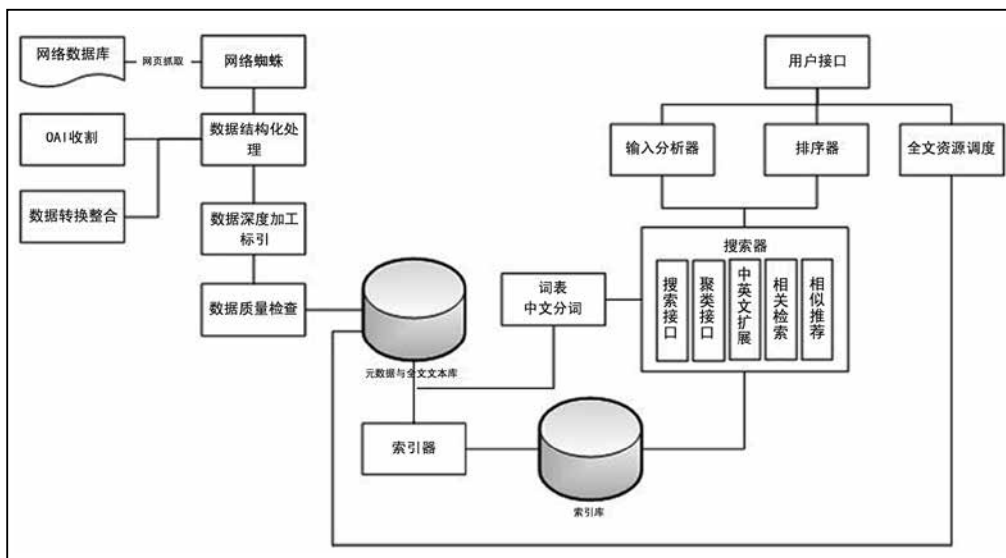


图2 学术搜索引擎的技术架构

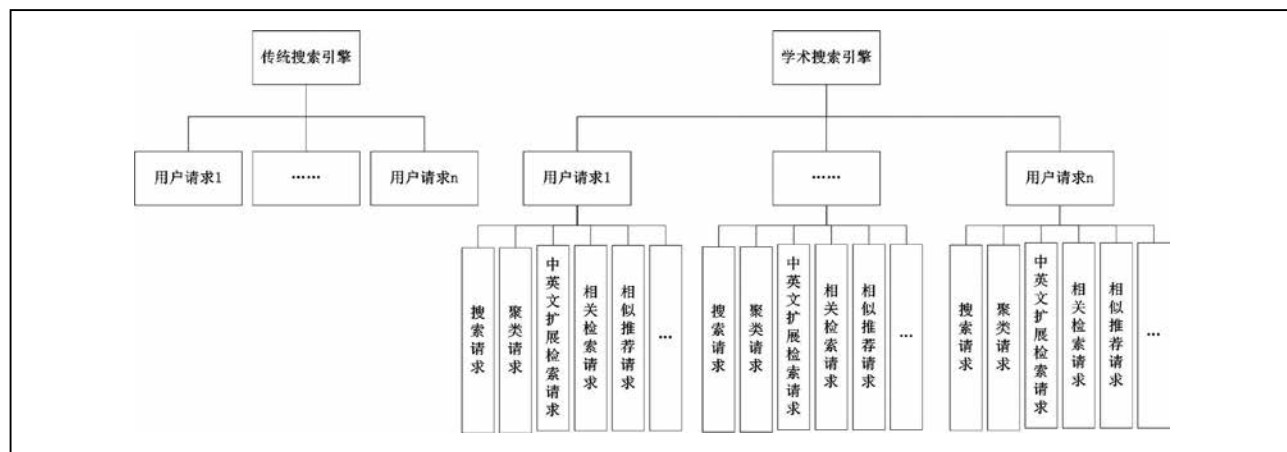


图3 传统搜索引擎与学术搜索引擎云计算应用的根本差别

进行关键词搜索时,可能是从上百亿条数据的索引中一次检索,并将检索结果直接快速返回给用户;而学术搜索引擎则完全不同,虽然其面对的数据集可能只有几亿条或者十几亿条,但用户在使用关键词进行一次搜索时,学术搜索引擎可能在后台实际需要瞬时发送几十甚至上百条检索或聚类请求,一方面为用户直接检索和返回检索结果,另一方面也为用户提供按照年代、学科、语种、基金、期刊、机构等不同维度的聚类检索分析,同时还可能为用户提供主题的本体扩展检索、中英文混合检索、相关检索、相似性检索、相关推荐等检索和结果推送。

因此,学术搜索引擎虽然以专业化检索、高查全率和查准率为目标,然而,其对于搜索性能的提升也有着极为迫切的需求,需要大量应用云计算相关技术,实现其持续的性能优化,满足学术搜索大并发量的检索、聚类统计和知识挖掘分析需求。

3 云计算技术在搜索引擎中的应用与发展

云计算是分布式处理、并行处理和网格计算的发展,其最基本的概念,是通过网络将庞大的计算处理程序自动分拆成无数个较小的子程序,再交由多台服务器所组成的庞大系统经搜寻、计算分析之后将处理结果回传给用户。稍早之前的大规模分布式计算技术即为“云计算”的概念起源。透过这项技术,网络服务提供者能够在数秒之内,达成处理数以千万计甚至亿计的信息,达到和“超级计算机”一样强大效能的网络服务。

云计算系统的建设目标是将原来运行在PC上或单个服务器上的独立的、个人化的运算转移到一个数量庞大的服务器“云”中,由这个云计算系统来负责处理用户的请求,并输出结果,它是一个以数据运算、处理和服务为核心的系统。云计算系统涉及的核心关键技术主要包括高可靠性的系统集群技术、并行计算技术、分布式文件系统、通用的终端服务支持等。

传统搜索引擎面对上百亿条网页数据,同时可能几十万人的并发检索请求,要在秒级甚至毫秒级对于用户检索请求进行响应,依靠单服务器的计算能力已经不能解决问题,需要利用集群和并行计算技术进行扩展。云计算技术的发展,给传统搜索引擎性能优化和提升提供了很好的解决思路。

Google搜索引擎面向世界各地用户,其搜索服务器达到了几百万台甚至上千万台,支撑其运行的基础是集群系统,即由多台同构或异构的计算机连接起来协同完成特定的任务,在这样的工作环境下就构成了计算的分布性。如何对于集群系统各计算机或其提供的服务进行冗余配置,进行即时监控,进行计算或服务资源调度,从而保证集群服务的可靠性和准确性,是构建云计算集群首先要考虑的问题。

Google搜索引擎同时应用了并行计算技术,其提出的Map/Reduce软件架构,就是一个大规模数据集并行计算的典型代表。它的基本思想是用多个处理器来协同求解同一问题,即将被求解的问题分解成若干个部分,各部分均由一个独立的处理机来并行计算。并行计算系统既可以是专门设计的、含有多个处理器的超级计算机,也可以

是以某种方式互连的若干台独立计算机构成的集群,通过并行计算集群完成数据的处理,再将处理的结果返回给用户。

传统搜索引擎面对海量数据,如何对于数据和索引进行存储,如何保证数据一致性,还需要用到另一项核心技术,就是分布式文件系统。分布式文件系统的设计应满足透明性、并发控制、可伸缩性、容错以及安全需求等,客户端对于文件的读写不应该影响其他客户端对同一个文件的读写。分布式文件系统需要做出复杂的交互,尽量保证文件服务在客户端或者服务端出现问题时能正常使用。分布式文件系统能提供备份恢复机制以保证分布式处理的可靠性。Google提出的GFS (Google File System) 是一个分布式文件系统的典型代表,Google GFS隐藏下层负载均衡、冗余复制等细节,对上层程序只提供一个统一的文件系统API接口,中心是一个Master节点,可以根据文件索引,找寻文件块。

可以说,传统搜索引擎是云计算技术研究和应用的急先锋,如Google Map/Reduce并行计算架构、Google GFS分布式文件系统等就是Google提出来的,并在Google搜索引擎中得到了成功应用。

学术搜索引擎和传统搜索引擎虽然在专业性、多维知识挖掘分析方面有其应用层面的差别,然而,其面对的应用场景又具有一定相似性,都需要满足大数据量、大并发量的检索需求,都需要应用云计算技术。Google的成功,Google的技术研究思路,为基于云计算架构的学术搜索引擎研制提供了很好的参考。

4 基于云计算架构的学术搜索引擎研制与构建

云计算技术的快速发展,硬件设备性能提升和价格下降,使得基于大规模集群、分布式并行计算的文献检索成为可能。万方数据通过对于传统搜索引擎的研究分析,结合学术搜索引擎的特点,决定采用云计算技术,对于现有RMS全文检索系统进行改造,研发基于云计算架构的学术搜索引擎。

1) 基于云计算架构的学术搜索引擎设计

万方数据在进行学术搜索引擎总体设计时,充分借鉴了Google在云计算应用方面的思路,在宏观层次,通过集群技术和负载均衡调度技术,在微观层次通过并行计算技术,基于云计算架构,实现学术搜索引擎的研制与构建。

基于云计算架构的学术搜索引擎总体技术架构如图4所示。

基于云计算架构的学术搜索引擎主要包括中国学术搜索网前端、负载均衡系统、静态缓存系统、搜索应用服务集群、知识挖掘分析应用服务集群、分布式动态缓存系统、元数据库服务集群、RMS Cloud核心搜索服务集群、运维管理体系等。

①多级负载均衡和任务调度实现云服务集成共享

基于云计算架构的学术搜索引擎通过多级负载均衡和任务调度,可以实现搜索应用服务、知识挖掘分析应用服务的集群部署与基于策略的分布式任务调度扩展,同时也支持多个RMS Cloud搜索引擎的分布式部署,根据资源占用、服务状态的任务调度,从而实现多个分布不同地域的云搜索服务集群的集

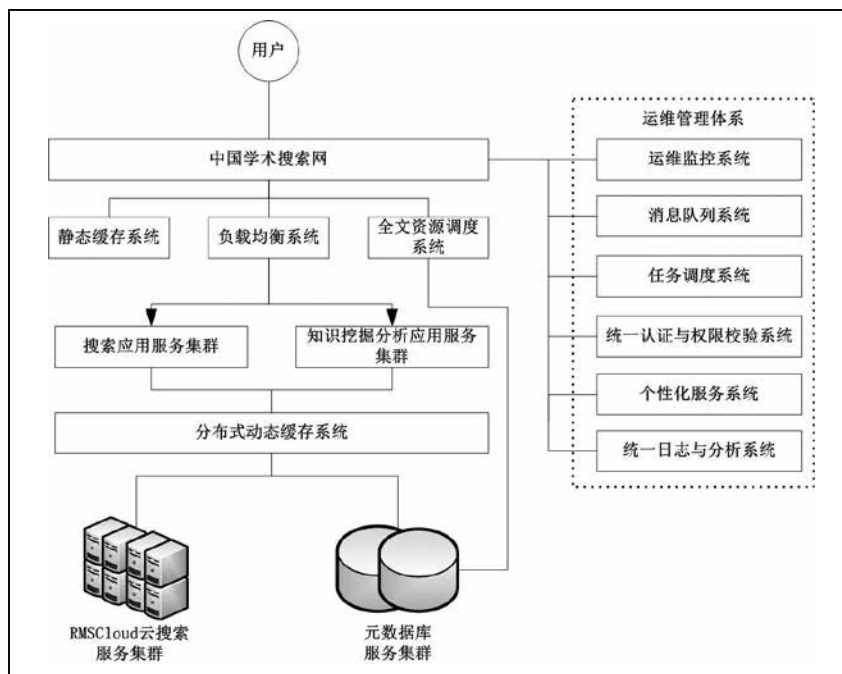


图4 基于云计算架构的学术搜索引擎总体技术架构

成与服务共享,形成一个规模更大的云服务集群。

②静态缓存与动态缓存技术提高检索性能

基于云计算架构的学术搜索引擎同时大规模应用缓存技术,如通过静态缓存,实现静态网页、图片的缓存和快速推送,通过分布式动态缓存,实现检索、知识挖掘分析结果的缓存和快速推送,从而提升学术搜索引擎的检索服务性能。

③高扩展性与大用户量、大并发量检索访问支持

基于云计算架构学术搜索引擎在Web服务器、应用服务器、元数据库服务器、RMS Cloud搜索引擎等各个环节,均支持采用集群服务配置方式,一方面通过多级负载均衡和任务调度,实现服务集群的分布式部署和扩展,另一方面,也可通过云搜索引擎的节点扩充、自动发现和自动同步机制,实现搜索服务集群的扩展。整个系统具备高度的开放性和可扩展性,能满足大用

户量、大并发量的检索、知识挖掘分析等业务应用需求,可以根据业务拓展和服务负载情况进行按需扩展。

2) RMS Cloud搜索引擎研制与构建

基于云计算架构的学术搜索引擎的核心是其RMS Cloud搜索引擎,其提供RMS全文索引检索技术及云计算架构支持。

RMS Cloud搜索引擎基于RMS中文智能分词技术,实现对于中文科技文献文本信息的分词,采用B*树索引算法,进行索引构建和全文检索,其索引技术支持最多7次实现索引的快速定位,为海量科技文献信息的学术搜索提供了全文索引和检索技术支持。

RMS Cloud搜索引擎同时采用云计算架构和并行计算技术,通过索引分片,减少单索引数据量,提高索引检索速度;通过索引副本,实现全文索引在集群多节点之间的分布,实现多节点并行计算;通过

无主从集群节点通信,实现节点数据同步,为集群节点可靠并行计算与云搜索服务提供保障。

RMSCloud云搜索引擎核心搜索服务技术框架如图5所示。

RMSCloud云搜索引擎分层结构及模块组成主要包括API接口、传输协议支撑、Java Netty框架、监控、RMS中文智能分词、第三方插件支持、云集群通信、脚本解析引擎、RMS全文索引、RMS全文检索、索引映射配置、数据源、分布式RMS索引目录支持、文件系统持久化网关等模块。

RMSCloud云搜索引擎在研制和构建过程中,应用了大量的云计算技术,实现集群与并行计算支持,满足大数据量科技文献学术搜索与知识挖掘分析需求。

①集群与分布式并行计算

RMSCloud云搜索引擎支持分布式并行计算技术,主要依赖于以下途径实现:

集群技术:集群中有多个节点,其中有一个为主节点,这个主节点可以通过选举产生,主从节点是对于集群内部来说的。对于集群外部来说,就是去中心化,从外部来看集群,在逻辑上是个整体,与任何一个节点的通信和与整个集群通信是等价的。集群节点故障不影响整个集群的对外服务,从而保证集群的可靠性。

索引分片:可以把一个完整的全文索引分成多个分片,这样的好处是可以把一个大的全文索引拆分成多个,分布到不同的节点上,在检索时,依托多个节点的计算能力进行并行计算和分布式检索。

索引副本:可以针对索引及分片设置多个副本,副本的作用一是提高系统的容错性,当某个节点某

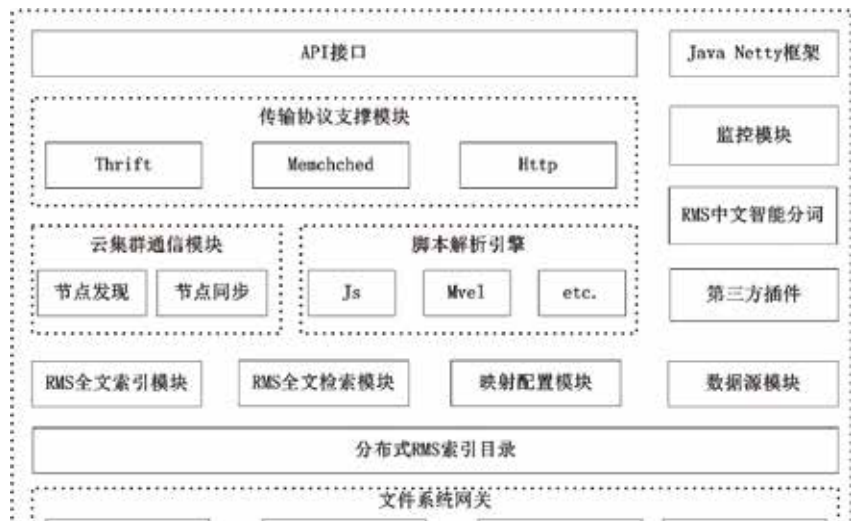


图5 RMSCloud云搜索引擎核心搜索服务技术框架

个分片损坏或丢失时可以从副本中自动恢复;二是提高检索效率,可以自动对搜索请求进行负载均衡,调度到其他节点副本上进行分布式检索。

②自动化维护与管理

RMSCloud云搜索引擎基于分布式计算模式,支持节点自动发现、节点扩展、数据自动重新分布、索引自动持久化存储等能力,可以实现对于云计算集群的自动维护管理功能。主要表现在:

自动节点发现:类似一个p2p的系统,它先通过广播寻找存在的节点,再通过多播协议来进行节点之间的通信,同时也支持点对点的交互。

数据重新分布:在有节点加入或退出时会根据机器的负载对索引分片进行重新分配,挂掉的节点重新启动时也会自动进行数据恢复。

数据源自动索引更新:可支持从数据源中自动读取数据并同步索引到RMSCloud云搜索服务集群中。

索引持久化存储网关:RMSCloud默认是先把索引存放到内存中,当分配内存满了时再持

久化到硬盘等存储网关。当这个集群关闭再重新启动时就会从存储网关中读取索引数据。RMSCloud支持多种类型的持久化存储网关,有本地文件系统、共享文件系统、Hadoop的HDFS和Amazon的s3云存储服务。

3) 学术搜索专有特性与应用体现

万方数据在基于云计算架构学术搜索引擎研制过程中,除借鉴传统搜索引擎在云计算应用方面的思路,同时也考虑了学术搜索引擎的专有特性,主要表现在:

①词表与中文智能分词技术的大量应用

万方数据在基于云计算架构的学术搜索引擎研制过程中,大量应用了词表和中文智能分词技术,主要包括:

万方科技文献主题词库:用于文献分词与索引构建,检索语句的分词与扩展检索、相关检索词提示等;

汉语叙词表:用于对于检索关键词基于词间关系,包括上位词、下位词、相关词、代用词等主题词本体扩展与相关检索、相似词推荐;

中英文主题词对照表：用于中英文词的对照翻译与中英文混合检索扩展；

学者库：用于对于作者进行规范、同名识别、检索时自动学者识别和展现学者空间；

多层次机构库：用于对于机构标引进行规范，机构合并、曾用名等进行检索分析扩展，检索时自动机构识别和展现机构空间。

②深度数据加工标引与多维度的聚类 and 知识挖掘分析支持

万方数据在基于云计算架构的学术搜索引擎研制过程中，对于中外文科技文献仓储知识库建设，制定了元数据加工标引、质量检查等一系列标准规范，提升数据加工标引的质量；同时，对于科技文献元数据，严格按照学科、主题、人物、机构、基金等“知识获取五要素”进行深度标引，为围绕五要素的检索、导航、多维度聚类和知识挖掘分析提供了基础。

③相关度计算排序与相似结果推荐

基于云计算架构的学术搜索引擎可以根据用户检索关键词进行自动识别，判断用户检索人物、机构、期刊、主题等检索意图，同时可提供按照检索词的相关度排序和相似结果推荐。基于云计算架构的学术搜索引擎通过对于数据库、字段及索引定义权重分值，以支持多字段过滤与相关度排序及相似结果计算。

4) 基于云计算架构的学术搜索引擎性能测试结果

基于云计算架构的学术搜索引擎核心搜索服务，即RMSCloud云搜索引擎在4台服务器（IBM x3650 M4: Xeon E5-2650 2GHz * 2, 32G Memory, 2.4T HD）集群配置

环境下，加载中文期刊库5800万条数据，并发用户300人，性能指标测试结果如下：平均检索响应时间为0.095秒，平均每秒钟处理事务量为6.956，每秒钟最大处理事务量为11.656。性能测试结果如图6所示。

5 围绕中国学术搜索网的万方云服务应用

基于云计算架构的学术搜索引擎的研制成功，为万方基于中外文科技文献元数据仓储知识库，为用户提供学术搜索服务和各种知识挖掘分析服务提供了坚实的基础。

万方拟以基于云计算架构的学术搜索引擎研究成果为基础，推出中国学术搜索网（<http://www.sciinfo.cn/>），为社会公众用户提供免费学术搜索、资源导航、全文资源调度服务，同时还将为有科研项目评审、审核、成果评价需求的类似各种基金管理部门、科技部、各地科技厅局等科技管理部门提供学者学术评价、机构科研能力评估、学科主题权威专家与机构推荐等科技管理类增值服务，为高校、科研院所、企业等有科研活动需求的用户提供科技查新、立项选

题分析、论文引证分析、学科热点分析、专利分析等科研活动类增值服务。

围绕中国学术搜索网的万方云服务应用模式如图7所示。

在中国学术搜索网上，各类增值服务采用云服务应用集成模式，以Web Service服务方式执行，可以是万方提供的自有增值云服务，也可集成第三方提供的增值云服务，各类增值服务可以独立部署，也可以分布式集群部署，从而满足大数据量、大用户量、大并发量情况下面向互联网提供增值应用服务的复杂业务应用需求。

6 结语

采用集群、并行计算、负载均衡调度、分布式文件系统等云计算技术，基于云计算架构学术搜索引擎研制取得了成功。在检索性能方面，从原有的秒级响应提升至毫秒级响应；在检索功能方面，则突破了原有检索系统在聚类、排序方面的局限，支持大数据的多维度聚类、知识挖掘分析和相关度排序；同时，系统高度的开放性和可扩展性，按照业务拓展和服务负载情况



图6 基于云计算架构的学术搜索引擎性能测试结果

的按需扩展,能满足大用户量、大并发量的检索、知识挖掘分析等业务应用需求,为万方围绕中外科技文献元数据仓储知识库提供学术搜索和知识挖掘分析服务提供了坚实基础。

随着基于云计算架构的学术搜索引擎研制成功以及中国学术搜索网对外正式提供学术搜索和知识挖掘分析服务,将为用户科技管理和科技创新活动提供必备工具。学术搜索引擎,可为各类科研工作者查找科技文献提供极大便利;科技管理类增值服务,将促进科技管理过程透明化,推动科技管理科学决策;科研活动类增值服务,将极大提升科研活动效率,推动科研自主创新。

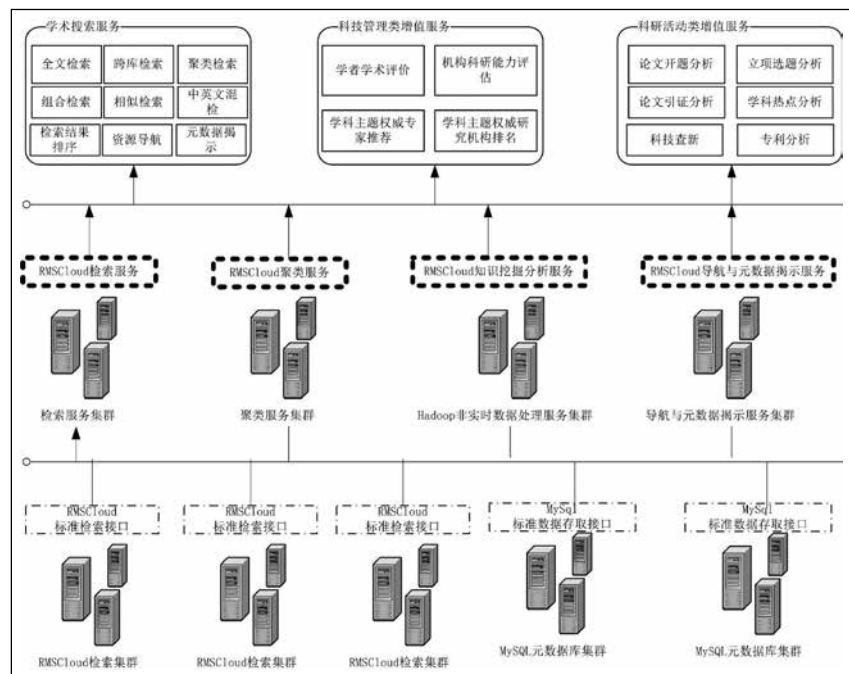


图7 围绕中国学术搜索网的万方云服务应用模式

参考文献

- [1] 林乐然,陈德龙.基于云计算的分布式企业搜索引擎研究[J].电脑知识与技术,2009(5):9429-9434.
- [2] 郑炜,梁战平.面向服务的搜索引擎技术进展研究[J].中国信息界,2012(2):64-67.
- [3] 云健,王春霞.搜索引擎技术综述[J].河西学院学报,2008(24):80-82.
- [4] 陶跃华,陈丽萍,王泳.聚类分析在搜索引擎中的应用[J].计算机工程与应用,2003(21):80-182.

作者简介

万齐鸣 (1971-), 男, 北京万方软件有限公司总工程师, RMS Cloud云搜索引擎产品经理, 863计划“以科技文献服务为主的搜索引擎研制”课题主要研究成员。研究方向: 搜索引擎、云计算与云服务、知识挖掘分析、内容管理、数字出版。E-mail: wanqm@wanfangdata.com.cn

Development and Application of Academic Search Engines Based on Cloud Computing Architectures

Wan Qiming / Beijing Wanfang Data Co., Ltd., Beijing, 100038

Abstract: By contrasting Traditional Search Engines and Academic Search Engines, the paper analyzes experiences of Traditional Search Engines in application of cloud computing technology. Combined with 863 Project, the author dissects the development mechanism of Academic Search Engines and the characteristics of its application in cloud computing technology, as well as the professional features of Academic Search, putting forward the application model of WANFANG Cloud Services around China's Academic Search.

Keywords: Cloud computing, Academic search, Search engines, Cloud services, RMSCloud

(收稿日期: 2013-05-09)