DLF 数字图书馆论坛
http: // www.dlf.net.cn

大数据、云计算与用户行为分析*

□ 吴恺 苏新宁 邓三鸿 / 南京大学信息管理学院 南京 210093

摘要：云计算的研究与应用方兴未艾，大数据近两年又成为国内外研究的热点问题和重要方向。文章介绍了大数据、云计算的概念和内涵，重点分析了大数据和云计算为用户行为分析带来的机遇和挑战。最后文章讨论了在大数据和云计算的背景下，用户行为分析在行为理论、信息规范、信息整合、分布式数据挖掘和数据可视化等方面的研究展望。

关键词：大数据，云计算，用户行为分析
DOI: 10.3772/j.issn.1673—2286.2013.06.004

1 引言

人类社会已经全面走进了信息时代，随着自动化机器的广泛使用，越来越多的用户行为信息被记录下来。如果能够挖掘这些用户行为数据，从中找出行为模式，进而分析用户的需求或预测用户要做的事情，对于掌握经济和社会的真实需求规律，为用户提供个性化服务，从而提高生产生活效率，减少无效浪费，无疑具有重大的现实意义。

现代社会的信息量已经远远超过人们在一二十年前对信息社会的构想。根据IDC的研究报告，人类社会的信息量每两年就会翻一番，2011年新产生和复制的数据总量达到1.8ZB（1.8万亿GB），其中75%的数据是个人产生的[1]。人们日常生产生活中使用的网络、手机或其他电子设备，每天都在不停地产生大量新的数据，超出了以往系统所能分析的能力，这种情况催生了大数据（Big Data）的问题。云计算（Cloud Computing）则为我们打开大数据的宝藏提供了钥匙，充分利用网络的存储和计算能力，将是突破数据处理的瓶颈，降低数据分析成本的可行之路。在大数据和云计算的背景下，用户行为分析的研究将翻开崭新的一页。

2 大数据与云计算

大数据指那些不能用传统的方式分析和处理的数据[2]。在2001年META Group的一份研究报告中，Doug Laney从三个角度定义了数据增长的挑战与机遇，即目前被广泛引用的3V模式：数量（Volume）、速度（Velocity）和多样（Variety）[3]。Volume是指数据的规模远远超过传统数据，例如Twitter和Facebook每天分别新产生7个和10个TB的数据，传统的数据分析技术对这样规模的数据将无能为力。Velocity指数据周转的速度，IBM认为大数据是在流动着的数据，我们必须在数据流动的过程中就开始分析和应用，而不可能等到一批数据结束后才开始分析[2]。Variety指数据的来源和类型多样，分析大数据时必须能同时处理结构化和半结构化，甚至是原始格式的数据。此外，人们还认为大数据的特征还有第四个V（Value），即蕴含巨大价值，对国民经济和社会发展有重大影响。

与传统数据的来源不同，大数据的来源不再仅仅局限于ERP、CRM等业务数据，还包括机器生成数据和社交数据[4]。机器生成数据（Machine-generated /sensor data）包括电话呼叫数据、各类服务器日志、传感器数据等，随着物联网的不断发展和传感器设备的普及，可获取的传感器数据变得越来越多。社交数据（Social data）则指在Web 2.0网络中用户参与的微博、社交网站、用户反馈等数据。

云计算概念的出现和受到关注略早于大数据，在2007年，Google在搜索引擎大会（SES San Jose 2006）首次提出“云计算”的概念，较早时候，亚马逊已经推出了弹性计算云的服务。云计算的概念描述了一种可以通过互联网进行访问的可扩展的应用程序。“云应用”

* 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。

使用大规模的数据中心以及功能强劲的服务器来运行网络应用程序与网络服务。任何用户可以通过合适的互联网接入设备以及一个标准的浏览器就能够使用云计算资源访问一个云计算应用程序。

根据美国国家标准和技术研究院的定义，云计算的服务模式有三种[5]：软件即服务（SaaS）、平台即服务（PaaS）和基础架构即服务（IaaS），这三种服务模式都已经有相应的商业实践。用户可以根据自己的要求，按需定制云计算服务，从而避免高额的硬件购置和维护的费用。

与以往研究首先起源于科研院所不同，云计算和大数据的概念首先是业界公司（Google和IBM）提出来的，可见其有强烈的社会现实需求和经济效益推动。2006年以来，Amazon、Google和IBM逐步推出了云计算服务，2011年，IBM又率先提出了大数据的概念，并且将投资1000万美元对大数据的基础平台等进行研究，将大数据作为其“智慧地球”工程的一个重要组成部分。去年奥巴马政府宣布启动“大数据研究与发展计划”，多个政府部门和机构联合开展大数据的研究，将大数据技术正式作为美国的国家科技战略[6]。

3 大数据与用户行为分析

近年来，用户行为分析研究已经逐步应用到大数据集并取得了一些初步成果。事实型数据是科技情报研究工作的基石，特别在用户行为分析领域，可用于分析的用户实际行为数据越多，期望得到的用户行为知识也就越丰富和越可靠。然而大数据的出现也为用户行为分析带来了诸多新的挑战，主要体现在以下几个方面：

(1) 数据存储问题。相对于传统数据，大数据在规模上不只是量的增长，而且是质的变化。传统的用户行为分析在数据存储上主要依赖于数据仓库。但在大数据情况下，数据仓库将会面临两个问题和一个鸿沟[7]。两个问题是数据移动代价过高和不能适应快速变化。一个鸿沟是巨型数据和数据处理能力之间的鸿沟。解决巨型数据的存储和处理问题是大数据用户行为分析的前提和基础。

(2) 信息规范问题。大数据的重要来源是机器生成数据和社交网站数据，这些数据中有许多半结构化的数据，不少数据还是原始数据。由于缺少对数据结构和涵义的说明，大量数据是定义不清、真假不分的杂乱数据，这对数据的预处理工作提出了更高的要求。

(3) 知识组织问题。基于大数据的用户行为数据挖掘是一个持续和逐步累积的过程，由于数据量巨大，数据挖掘的结果需要存在云平台的不同数据节点中。面向大数据的知识组织需要具备高扩展性、支持动态更新、便于信息整合等特性。

(4) 信息安全与公民隐私问题。一方面，大数据中存在着许多用户隐私信息，例如个人检索浏览信息或手机定位信息。如何保证这些信息得到合法合理的利用，在何种情况下可以使用这些信息，需要有更清晰的法律或互联网行为规范。另一方面，大数据的处理很难限制在封闭的环境中运行，这种情况下，信息安全也从传统计算机网络安全、保密管理等可控安全管理变为无法确知安全隐患的不可控的安全管理[8]。如何使得数据不会被盗用、泄露商业机密，处理数据开放与信息安全的矛盾，是用户行为分析能够成功应用的必要保障。

4 云计算与用户行为分析

近些年来云计算的概念正在逐渐重塑整个IT行业，也给用户行为分析的研究和实践带来了前所未有的机遇。面对当前海量的用户行为数据，仅仅依靠单机软件或数据仓库已经不能有效实施用户行为数据的分析，需要引入云计算的架构和技术。对于用户而言，软件技术的发展日新月异，用户不需要关心或维护本地的软件环境，通过云计算的平台就可以获取到自己所感兴趣的资源或知识。云计算能够为基于大数据的用户行为分析主要提供以下一些技术支持：

（1）基于云的计算服务模式。一直以来，系统成本是大规模用户行为数据分析未能得到广泛应用的主要原因。一些高端科研项目或国防军事项目可以耗费巨资组建大型计算机，而商业领域中的中小企业则望而却步。云计算提出了以租代买的计算服务模式，能够大幅度提高现有系统计算能力的有效使用率。对于商业领域而言，低敏感性的数据可以考虑在公用的云数据平台中处理以降低成本。

（2）基于云的数据存储模式。为了支持大数据的存储与访问，许多公司很早就开始了基于分布式和网络的文件系统的研究。GFS（Google File System）是Google公司为了满足其需求而开发的基于Linux的专有分布式文件系统。GFS的硬件系统是一个大规模中低端计算机集群，其中包括两类节点，一个主节点（也可

以还存在一个影子备份节点）和许多的数据节点。应用程序在访问数据时，首先访问主节点，获得数据节点的信息和授权，然后访问数据节点。某个数据节点出现问题，并不会影响整体数据的使用[9]。

（3）基于云的分布式计算架构。与GFS类似，MapReduce也是面向大规模计算机集群设计的，由主节点控制和分配各子节点的计算资源[10]。任何一个子节点都可以从集群中移除，而不会影响当前任务的执行。在MapReduce架构下，分布计算、容错、负载平衡等技术细节可以由系统自行完成，用户不需要这方面的知识就可以有效地使用分布式计算。

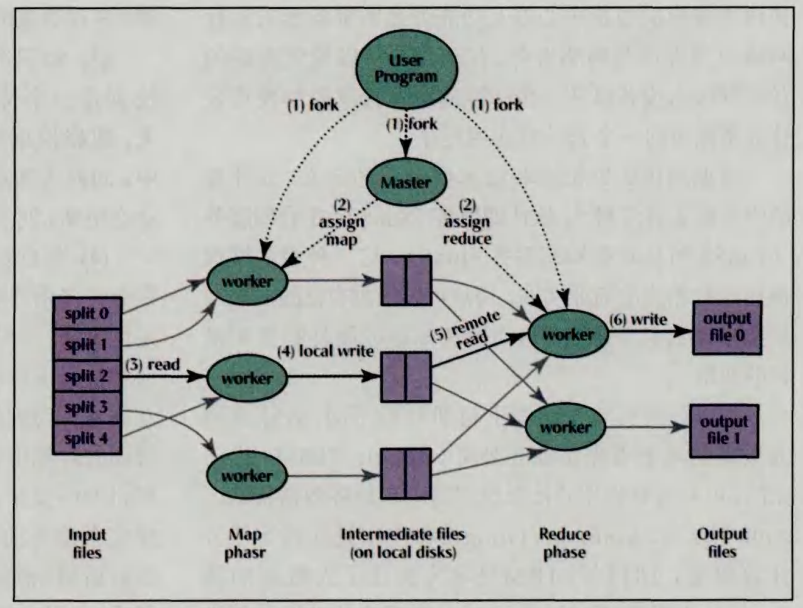

图1 MapReduce工作模式示意图

在Google提出的GFS和MapReduce思想的基础上，Apache软件基金会开发了一个开源分布式计算框架Hadoop，它是一个能够对大量数据进行分布式处理的软件框架，并且已经在实践中取得了很好的效果[11]。Hadoop整体架构包括四个模块：

（1）Hadoop Common，提供支持Hadoop系统的公用组件；

（2）Hadoop Distributed File System，用户实现高数据吞吐量的分布式文件系统；

（3）Hadoop YARN，任务计划和子节点调度程序；

（4）Hadoop MapReduce，基于YARN实现的大规模数据集的平行计算。

5 大数据、云计算背景下的用户行为分析研究展望

在大数据和云计算的背景下，以往用户行为分析中的数据采集、模式分析、知识应用等三步骤模式已经很难实现数据分析的目标。大数据的3V特征，意味着对于这样规模的数据，需要在数据不断生成的过程中边分析、边利用。可以预见，在大数据的背景下，我们需要对传统用户行为分析的一系列环节开展深入研究。

5.1 行为分析理论研究

在大数据的背景下，用户行为分析的对象不再局限于人们的经济行为、而是逐步扩展到人们的信息行为、社会行为，甚至情感行为。较早的用户行为分析中，“啤酒与尿布”的例子形象地说明了在用户经济行为中适用的关联规则分析；在信息检索时，协同过滤理论可以基于相似用户的检索评分记录为用户提供个性化推荐[12]；在分析社交网站数据时，社会网络分析模型及中心度、密度等概念有助于我们更好地刻画用户之间的社会联系，从中发现有价值的规律和知识。在信息扩散、大众情感行为分析中，更需要我们从用户心理、社会传播等多个学科寻找理论基础，建立适用于相应大数据集的行为理论分析模型。

5.2 信息组织规范研究

现有的大数据是以业务数据、日志数据的形式产生的，并未考虑如何使这些数据更适用于用户行为分析，这为有效地分析用户行为数据带来了困难。用户行为分析研究可以根据企业和个人用户的需求，提出大数据的信息组织规范，通过元数据、标记语言等形式为大数据的信息涵义提供说明，便于用户行为分析软件进行数据的提取和分析。

面向大数据的信息组织规范应当具有高扩展性、容错性、地理时间维度和便于分布式计算。在信息组织过程中需要兼顾信息的精确性和普适性两方面的需求，大数据的建模不能完全参照关系数据库的范式，可以允许有一定的数据冗余和数据不一致。考虑到大数据用户行为分析具有宏观应用、中观应用、微观应用等多个层次，信息组织的规范应当具有一定的层次性和弹性，使得用户行为数据可以在不同的层次进行分析。

5.3 信息资源整合研究

信息资源整合的目标是将分散的资源集中起来，把无序的资源变为有序，使之方便用户查找信息、方便信息服务于用户[13]。理论上数据资源整合可以表述为本地资源向全局视图的映射[14]，映射构建的主要方法主要有GAV和LAV两种，GAV虽然查询效率比较高，但映射关系的可扩展性差。对于大数据集，LAV方法的可扩展性强，可以集成大量的数据源并允许在数据集中过程随时增加和去除数据源节点。

信息资源整合可以分为两种类型，第一种是同类数据的汇合，例如将多个高校BBS上的用户发帖信息汇集起来可以研究整体大学生的关注热点。这些数据虽然是同类的，但仍存在数据格式、数据规模等方面的不同，在信息资源整合时既要考虑到数据格式的转换和统一，又要考虑到不同数据规模的代表性问题。第二种信息资源整合是异类数据之间的关联，例如在科技文献平台上，将用户检索日志和文献摘要、关键词信息整合起来。

5.4 分布式数据挖掘研究

分布式数据挖掘对于进行大数据集的用户行为分析具有两个优势，其一是由于数据安全性、数据异构性以及法律约束等多方面因素，将数据集中到一起进行分析是不可行的，需要将数据在本地完成一定的处理后才能进行后续的模式分析。其二是将数据源分成较小的模块，尽量在本地完成数据预处理和数据挖掘，最后将数据挖掘的结果合并，可以有效地降低数据传输和系统计算的开销。图2就是一个基于云计算的分布式用户日志分析框架，日志预处理、模式挖掘和模式分析可以在云端的不同节点分别完成。

对于同结构数据的分布式数据挖掘，研究者们分别提出了元学习（meta learning）和相互学习（coactive learning）。元学习的基本思想是从已经获得的知识中再进行学习，从而得到最终的数据模式[15]。相互学习是多个数据挖掘节点在进行同一学习任务时可以互相共享学习成果[16]。对于异构数据的分布式数据挖掘，Kargupta等人提出了协作数据挖掘（Collective Data Mining），对各节点的异构数据首先使用正交化基函数生成分量模型，最后再汇总生成全局模型[17]。

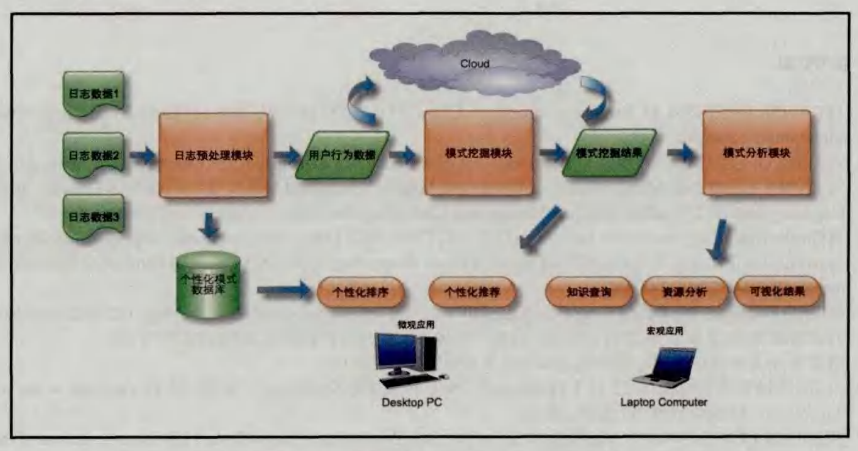

图2 基于云计算的分布式用户日志分析框架

5.5 知识呈现方式研究

在知识呈现方面，由于大数据多是表现宏观的变化趋势，因此更需要借助数据可视化的技术来表现行为分析的结果。以SmogFarm为例，SmogFarm是一个基于Facebook情感数据的大数据分析实例[18]。SmogFarm将人类的情绪细分为高兴、愤怒、悲伤、满意等12类，每种情绪有其特有的触发类型、心理症候、面部表情和行为刺激。经过数据比较，SmogFarm的情绪分析波动比Gallop的国民情绪调查领先一天。SmogFarm通过可视化的方式呈现美国的大众情感波动，从下图中可以看到，2013年4月15日波士顿爆炸案引起了美国民众普遍的悲伤情绪。借助用户可视化技术，用户可以更为直观地看到大数据所揭示的整体规律和发展趋势。

6 结语

大数据的出现意味着有更大规模和更多类型的数据可以进入用户行为分析的应用范围，这既为用户行为分析带来新发展，也对现有的理论框架和技术提出了挑战。我们在面对大数据时既可以用传统的情报学研究范式作为指导，将数据产生、信息分析和知识组织看作一个整体，用综合的视角研究基于大数据的用户行为分析问题，又应该看到在某些方面需要我们采用新的视角研究问题，而不仅仅是对原有方法的扩展和补充。在深入分析大数据的特性和应用云计算技术的基础上，进一步加强对用户行为理论、信息组织规范、信息资源整合、分布式数据挖掘和知识呈现方式的研究，可以带领我们进入大数据的宝藏，发现隐藏在大规模用户行为数据里面的规律，构建人类的知识大厦。

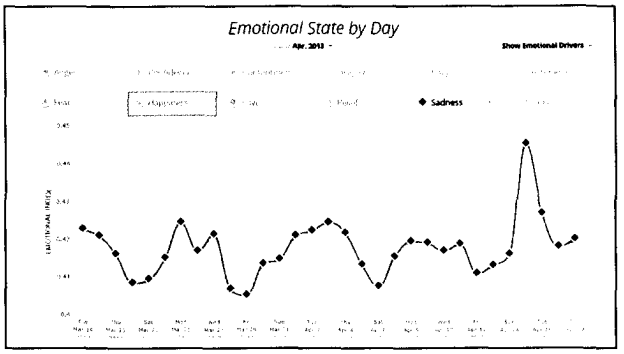

图3 大众情感分析的数据可视化

参考文献

[1] GANTZ J, REINSEL D. Extracting Value from Chaos [J/OL]. IDC IVIEW, 2011(12) [2013-03-25]. http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.
[2] ZIKOPOULOUS P, EATON C. Understanding big data: Analytics for enterprise class hadoop and streaming data [M]. McGraw-Hill Osborne Media,2011.
[3] LANEY D. 3D Data Management:Controlling Data Volume,Velocity,and Variety [J]. Application Delivery Strategies, 2001 [2013-03-18]. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.
[4] Oracle. Oracle:Big Data for the Enterprise [OL]. 2012 [2013-03-25].http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf.
[5] MELL P, GRANCE T. The NIST Definition of Cloud Computing [J/OL]. NIST Special Publication 800-145, 2011 [2013-03-25]. http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.
[6] TheWhiteHouse. Big Data is a Big Deal [OL]. [2012-12-09]. http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal.
[7] 王珊,王会举,覃雄派,等.架构大数据：挑战、现状与展望[J].计算机学报,2011(34):1741-1751.
[8] 贺德方.大数据环境下的情报学[J].数字图书馆论坛,2012(11):3-6.
[9] GHEMAWAT S, GOBIOFF H, LEUNG S-T. The Google File System [C]// SOSP '03 Proceedings of the nineteenth ACM symposium on Operating systems principles, ACM New York, NY, USA, 2003.
[10] DEAN J, GHEMAWAT S. MapReduce: Simplified Data Processing on Large Clusters [J]. Communications of the ACM 2008, 51(1): 107-113.
[11] Apache [OL]. [2013-03-25]. http://hadoop.apache.org.
[12] GOLDBERG D, NICHOLS D, OKI B M, et al. Using Collaborative Filtering to weave an Information Tapestry [J]. Communication of ACM, 1992(35): 61-70.
[13] 苏新宁,章成志,卫平.论信息资源整合[J].现代图书情报技术,2005,21(9):54-61.
[14] LENZERINIM. Data integration: a theoretical perspective [C]// PODS '02 Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2002, 233-246.
[15] PRODROMIDIS A, CHAN P, STOLFO S. Meta-learning in distributed data mining systems Issues and approaches [C]// Advances in Distributed and Parallel Knowledge Discovery, 2000.
[16] GRECU D, BECKER L. Coactive learning for distributed data mining [C]// Proceedings of KDD, 1998, 209-213.
[17] KARGUPTA H, PARK B, HERSHBERGER D. Collective data mining: A new perspective toward distributed data mining [C]// Advances in Distributed and Parallel Knowledge Discovery, 1999.
[18] SmogFarm. SmogFarm: Harvesting the Cloud [OL]. [2013-04-15]. http://smogfarm.com/.

作者简介

吴恺（1979-），博士生，研究方向：用户行为分析，网络日志挖掘。E-mail: wukai@nju.edu.cn

Big Data, Cloud Computing and User Behavior Analysis

Wu Kai, Su Xinning, Deng Sanhong / School of Information Management, Nanjing University, Nanjing, 210093

Abstract: Along with growing research and application of cloud computing, big data has become a hot topic of domestic and international research in recent years. This paper introduces the concept and connotation of big data and cloud computing and discusses the impact and challenge of big data on user behavior analysis. Moreover, under the background of big data and cloud computing, the paper illustrates the research prospects of user behavior analysis such as user behavior theory, information standard, information integration, distributed data mining and knowledge representing.

Keywords: Big data, Cloud computing, User behavior analysis

（收稿日期：2013-04-23）