基于大数据、云服务的科技情报工作思考*

□ 贺德方 / 中国科学技术信息研究所 北京 100038

摘要：本文简要回顾情报学的历史，从科学研究的4个范式引入，描述了情报工作与大数据、云服务的关联。在简要总结大数据、云计算、云服务特征及发展趋势后，归纳总结了情报工作面临的挑战与机遇，结合WorldCat、CALIS、NSTL案例，给出了情报工作在发展方向上的思考。

关键词：情报学，云服务，大数据，数据科学，知识创新
DOI：10.3772/j.issn.1673—2286.2013.06.002

前言

情报学的历史

1945年7月，美国科学研究与发展局局长Vannerar Bush在《*The Atlantic*》上发表了情报学不朽的篇章“As We May Think”（诚如所思）一文[1]，标志着情报学作为一门学科正式诞生。二战时期情报工作被看作是军队战斗力提高的因素之一，决定着国家的生存；20世纪50年代，情报工作与国家安全维系在一起；60年代，是美国与前苏联的科技情报竞争时代；70年代，情报学研究重点转向新技术；80年代，情报技术在支撑科学家理解文献知识方面受到了质疑；90年代，计算机及网络技术开始冲击着情报学；进入21世纪，情报学的冲击又多了一层——社会信息化；时下——大数据云服务时代，情报学能否实现支撑科学家获得客观世界的“insight”（洞察）？尽管众说纷纭，莫衷一是，但国际情报学的发展与时代的标签“大数据与云服务”已经不可分离。

回顾中国，1956年10月，国家建立第一个情报中心——科学情报研究所，开创了我国正规的科技情报事业。在追踪国际和政府科技信息决策服务的战略指引下，我国情报学开始了发展之路。70~80年代，我国的情报科学基于科技情报和情报交流而建立；90年代，情报科学的研究从科学技术逐步扩大到整个社会，发展了“社会——知识”的理论观点；21世纪初，诞生了“信息——知识——智慧”的理论；而大数据、云服务的背景下，中国科学技术信息研究所提出了“事实型数据库+工具方法+专家智慧”的情报研究思路与情报实践方法，并以广大的科研用户为驱动，向着提供“insight”的智慧服务而发展。

情报学与科学范式：基于大数据云服务的情报学

情报学服务于科学（Science），科学又反向作用于情报学。科学以“范式”来标志其特征，情报学与科学范式自然密不可分。伴随着社会信息化的高度发展，科学研究的主导范式与研究手段经历了4个阶段的变迁，参见图1。作为“数据密集计算”科学的第4范式，其内涵可以理解为基于大数据驱动、以“云服务”为手段的科学范式。由此我们也可以推断：情报学发展的必由之路只能是“基于大数据云计算的情报学”。

事实上，我们可以发现：情报学的研究从来没有离开过数据。二

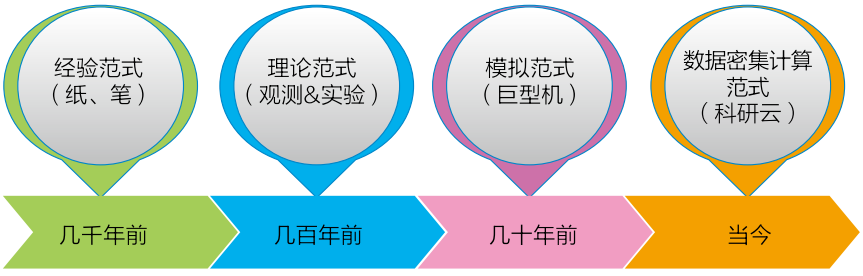

图1 科学范式与科学研究方法的迁移

* 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。

战后，美国科技信息爆炸的知识困惑问题，在Vannevar Bush的带领下走向了用机器管理大规模数据，并为科学研究所用的探索与发展之路。这已经标志着情报学从诞生之初就与大规模数据结下了不解之缘。

综合利用多来源、多形式的数据是现代情报工作的鲜明特点。如何快速地获取准确数据、高效地分析海量数据、清晰地解读系列数据是情报工作者当前面临的严峻挑战[2]。本文以“基于大数据、云服务的科技情报工作思考”为题目，希望起到抛砖引玉的作用。

以下将简要概述大数据、云计算和云服务以及大数据与云服务的关系，并结合国内外经典的3个数字图书馆云服务案例，归纳总结情报工作当下面临的机遇和挑战。

1 大数据

大数据概念的提出可以追溯到《自然》杂志2008年9月专刊中发表的文章“Big Data: Science in the Petabyte Era”，此后大数据概念被广泛应用和传播。维基百科对大数据的定义是：“大数据是由于规模、复杂性、实时而导致的使之无法在一定时间内用常规软件工具对其进行获取、存贮、搜索、分享、分析、可视化的数据集合”。互联网数据中心将大数据定义为：为更经济地从高频率的、大容量的、不同结构和类型的数据中获取价值而设计的新一代架构和技术。

PB甚至更高级别大数据的出现是近年来移动通信、互联网、传感器、物联网等技术发展和应用的结果。大数据一方面对于企业来说意味着巨大的经济效益，另一方面，大数据对IT业也意味着对海量、分散、变化、异构特性数据进行分析和管理的技术挑战。大数据带来的技术挑战涉及数据的收集、存储、检索、共享、分析以及可视化等各个方面[3]。

1.1 大数据的特点

关于大数据的特点也有很多种不同说法。百度百科给出的大数据的特点是4个“V”，即：数量巨大（Volume）、类型繁多（Variety）、价值高（Value）、处理速度快（Velocity）。Forrester分析师布赖恩•霍普金斯和鲍里斯•埃韦尔松在“首席信息官，请用大数据扩展数字视野”报告中给出的大数据的4个特点是：海量（Volume）、多样性（Variety）、高速（Velocity）和易变性（Variability）。

1.2 大数据的应用

2012年被称为大数据元年，首先是美国政府公布“大数据研发计划”，接着世界各国以及各大商业公司也对大数据给予了极大的关注。中国政府、各学协会及大型企业也给予了积极响应。但全球范围内对大数据的研究、开发、应用的引领者依然是强势的IT跨国企业。

- Google：Google所提出的GFS、BigTable、MapReduce技术奠定了云计算研究和应用的基础。GFS是Google提出的分布式文件系统，可以支持对分布在大量廉价硬件的数据进行有效可靠的访问。BigTable是Google构建在GFS之上的一种压缩高效的专属数据库系统。MapReduce是支持在大规模集群上的大数据进行并行计算的软件框架。基于这3项针对大数据存储访问和计算的关键技术，Google可以进行海量数据的搜索和分析挖掘，保证了其在搜索领域的主导地位。
- IBM：IBM在数据分析与挖掘领域的聚焦由来已久。IBM的数据分析产品研发和解决方案将其数学分析的才能应用于石油勘探、医疗健康等各领域，提升各行业的智能水平和决策的科学性。
- Microsoft：微软在数据管理、商务智能、数据挖掘的研发和解决方案是以其结构化查询语言（SQL）Server平台为基础的。目前微软已有大数据实施成功的案例，包括目前正在成都投入运作的云计算中心。该中心采用了大数据平台、虚拟化、BI商业智能分析等一系列技术。大数据的发展可以概括如图2所示：

2 云计算与云服务

2.1 云计算

云计算一词2006年由Google的CEO埃里克•施密特最初采用，从Google App Engine到Amazon EC2的出台，2008年云计算开始得到普及。云计算（参见图2）是一种无处不在、便捷、能按需获取网络上可配置计算资源的共享池（比如网络、服务器、存储、应用和服务）模式，这些计算资源能以最小的管理成本或通过与提供服务者之间的交互快速地部署和发布。这一模式可以概括为5个基本特征、3种服务模式和4种部署模式构成[4,5]，如图3所示。在云计算中，通常主要提供捆绑软件利用的就是SaaS[6]。

5个基本特征

1) 按需的自助服务On-demand self-service

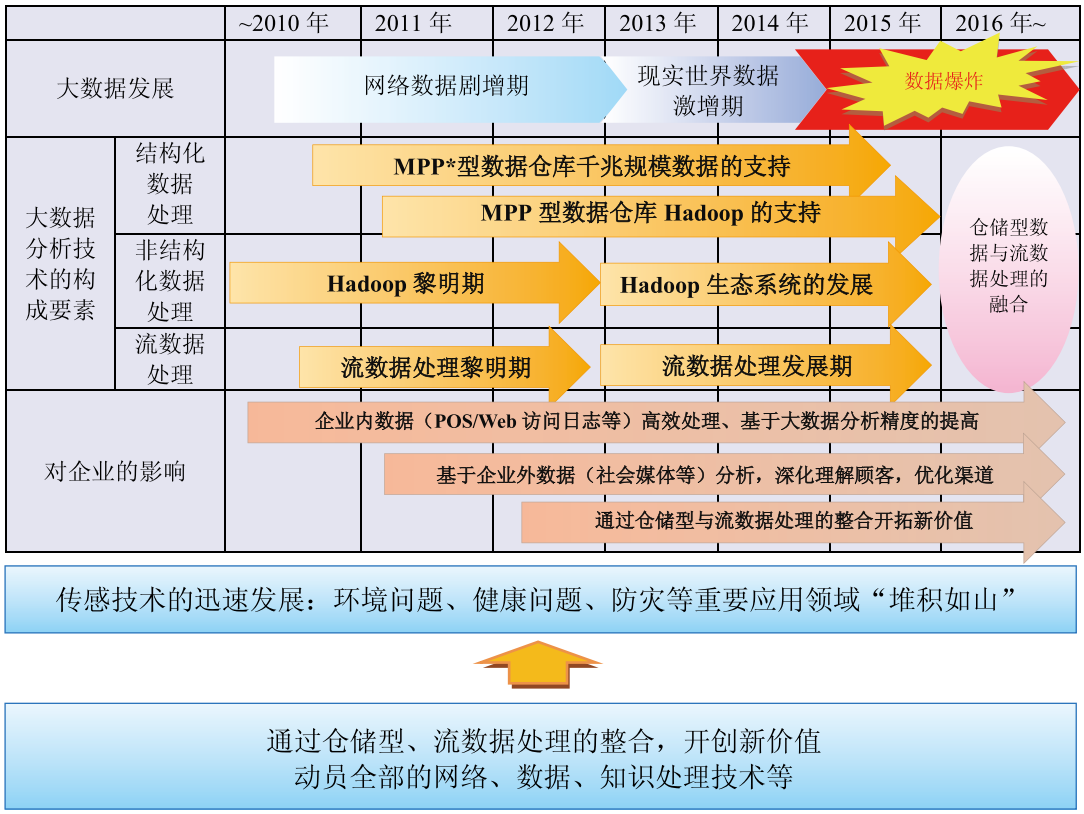

图2 大数据的发展纵观

* MPP: Massive Parallel Processing海量并行处理架构

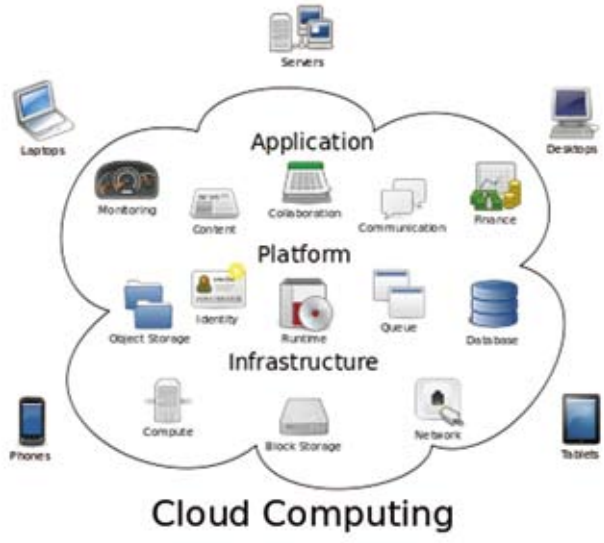

图3 云计算的概观

2) 广泛的网络访问Broad network access

3) 资源池Resource pooling

4) 快速灵活Rapid elasticity

5) 可量测的服务Measured service

3 种服务模式

1) 软件即服务Software as a Service (SaaS)

2) 平台即服务Platform as a Service (PaaS)

3) 基础设施即服务Infrastructure as a Service (IaaS)

4 种部署模式

1) 私有云Private cloud

2) 社区云Community cloud

3) 公用云Public cloud

4) 混合云Hybrid cloud

图4 云计算的发展史

2.2 云服务

基于云计算的各种服务即云服务，这也是“云”成为研究热点的重要理由。在数字图书馆领域，云服务就是利用云计算及其海量信息的存储能力，整合最丰富的信息资源，以极低的成本投入获取极高的运算能力，实现信息资源的共建共享，为用户提供优质全面服务的一种服务模式[7]。云计算的发展历史如图4所示。

典型案例1——WorldCat：全球图书馆联盟OCLC的WorldCat以数据库标准对所有数据进行分类存储，为各个服务提供一种统一的通信、发布、实现方式[8]。WorldCat根据数字图书馆的服务实现的流程分为5个层次：供应链层、数据整合层、数据库层、服务平台层、应用层。如图5所示：

1) 供应链层：作为WorldCat馆藏的最基础来源，是指成员馆的图书、期刊或数据库订阅授权、电子馆藏的购买与获取过程。

2) 数据摄入层：WorldCat将从供应链中收集回来的纸质与电子馆藏，通过OCLC统一的全自动在线编目体系Connexion进行编目组织，并将其编目信息与相关的元数据信息整理成与WorldCat编目数据库一致的数据格式之后，存储于WorldCat数据库中。

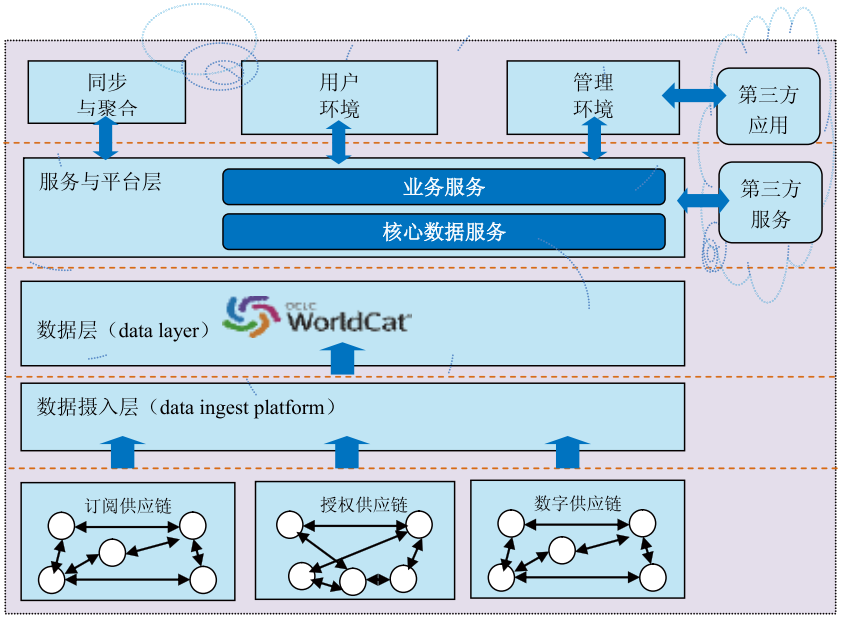

图5 OCLC的WorldCat云计算图书馆服务架构

3) 数据层：WorldCat联机目录数据库的主体，包括所有被编目与存储在WorldCat数据库中的书目数据与数字馆藏的信息。

4) 服务与平台层：基于数据层中的WorldCat联机目录数据库为用户提供SaaS与PaaS服务。SaaS服务的出口是WorldCat.org，该出口为用户提供一体化的数据查询服务，用户直接从该界面中查询WorldCat的书目信息。PaaS服务提供WorldCat API，供第三方程序开发商开发能与其他平台上调用

WorldCat数据库信息的程序。

5) 应用层：是指第三方或OCLC基于服务与平台层所开发的一系列能调用WorldCat信息资源的客户端与应用程序。如Amazon与Facebook上基于WorldCat API所开发的，让用户实现图书信息分享功能的插件、Mobile Library等。

典型案例2——CALIS：在云计算发展的浪潮中，中国高等教育数字图书馆（CALIS）经过多年的积累和建设，也在三期建设中采用了云计算技术作为自身的解决方案基础，搭建了由CALIS中心级云平台、共享域级云平台和本地平台组成的多云服务体系。CALIS提出了以“云计算”技术为基础的高校数字图书馆技术体系以及由全国中心、省中心、共享域中心和核心馆组成的服务体系。具体如图6所示。

典型案例3——NSTL：经过10多年的发展，国家科技图书文献中心（NSTL）已经成为国家科技基础条件平台建设的重要组成部分，逐渐演变为推动文献平台建设和促进科技文献共享的中坚力量，有力地承担了国家科技信息资源的战略保障任务。近些年，随着信息技术的迅猛发展，NSTL也在不断尝试加大为社会提供服务的深度和广度，用云计算、云服务理念规划未来架构，一方面有效利用资源，提高服务能力，另一方面将提升面向社会服务深度，从文献服务向知识服务、情报服务不断迈进。目前NSTL正在进行内部私有云的建设工作，为内部各成员单位提供基础设施服务，并为全国各服务站、用户管理平台提供平台服务，此外，NSTL也正逐步梳理业务流程，设计并构建信息服务及情报服务的框架，为将来面向全社会的深层情报分析与评价提供能力及技术支持。参见图7。

图6 CALIS国家中心级云服务平台[9]

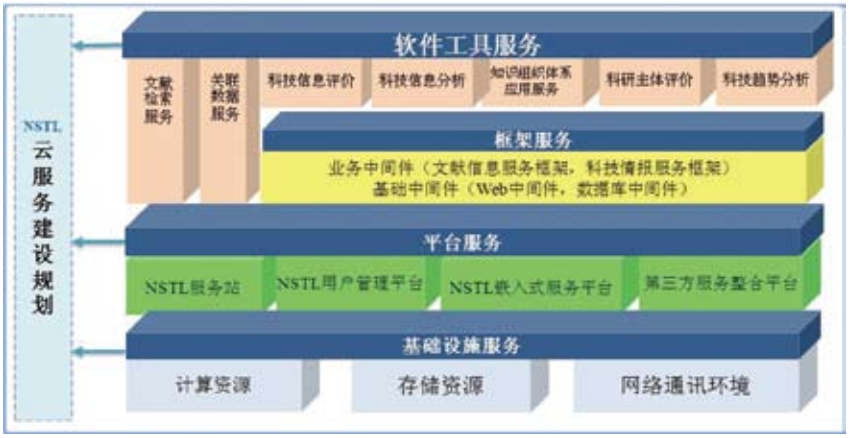

图7 NSTL云服务建设规划

3 大数据与云计算、云服务的关系

大数据与云计算是关系紧密的两个概念。广义上，大数据技术涵盖了从数据的海量存储、处理到应用等多方面的技术，包括海量分布式文件系统、并行计算框架、NoSQL（Not Only SQL，指非关系型的数据库）、实时流数据处理以及智能分析技术，如模式识别、自然语言理解、应用知识库等。狭义上，大数据技术是指从大量、多样、分散和异构的数据集中提取有用信息的核心技术，包括实时流数据处理以及智能分析技术等。

云计算之所以一经提出就得到广泛关注，是因为它使得人类“将计算能力作为公共事业设施来提供”的梦想变为现实。其关键技术是GFS、BigTable和MapReduce，这3项技术是Google为了巩固其搜索领域的核心地位而提出的。Google提出将文件和数据分割成块，以便支持分布式存储和并行处理，实现

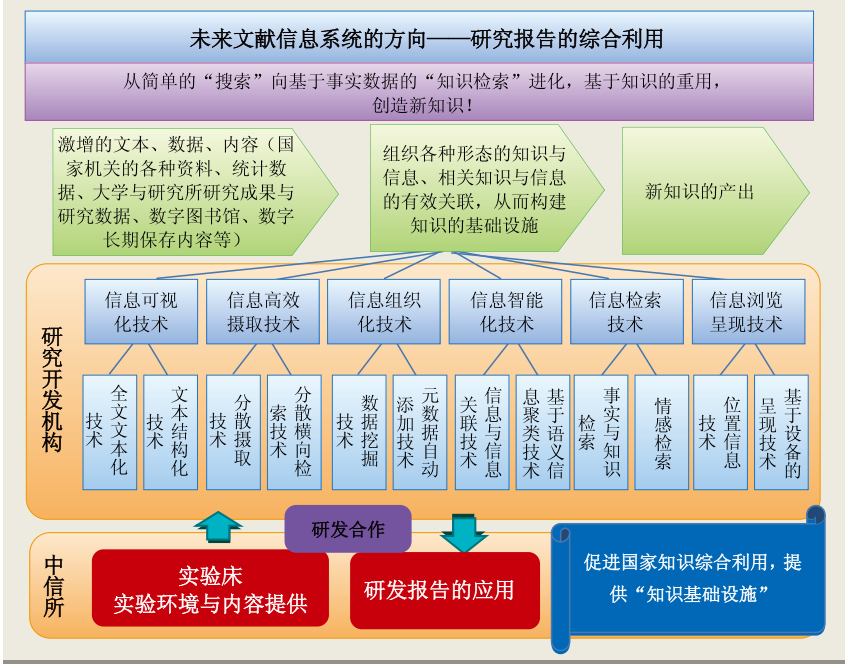

图8 基于研究报告综合利用的文献信息服务模式

海量数据存储并提升大数据量下的快速数据处理。因此可以说，云计算的核心是业务模式，本质是数据处理技术，云服务体系。

由此看出，云计算技术是广义大数据技术的一部分，也是狭义大数据技术的基础。可以说，大数据是资产，云为数据资产提供了保管、访问的场所和渠道。如何盘活数据资产，使其为国家治理、企业决策乃至个人生活服务，是大数据研究的核心问题。一方面，大数据离不开云计算，正因为有了云计算的超强计算能力，大数据才显示出了价值；另一方面，大数据处理的兴起也将改变云计算的发展方向，云计算正在进入以分析即服务（AaaS）为主要标志的Cloud 2.0时代。

4 基于大数据、云服务的科技情报工作机遇挑战思考

4.1 机遇和挑战

4.1.1 科技情报研究方法的突破

信息检索与知识组织可以说是情报工作的两个核心阵地，前者一直是计算机学科的擅长，后者则是图书馆学的根基。在研究方法上情报学一直紧跟许多相关学科的发展，不断地吸收利用。SCI创始人加菲尔德先生1955 年提出编制引文索引的设想，1963年编成出版《科学引文索引》，成为文献检索和引文分析的重要工具，不仅为文献计量学和科学学（Science of Science）的发展作出了重要贡献，也使得文献计量学的方法成为情报工作多年来发展的方法论根基。今天的情报用户面对着来自多种来源的庞杂信息，传统出版流程已经被Web 2.0解构，因此情报工作仅限于传统文献的计量分析已经无法满足用户需求，更不用说用户已经开辟了多种渠道来获取情报信息，因此文献计量学甚至于网络计量学都面临着巨大的挑战和机遇。如何突破创新，构建基于文献而又超越文献，用全新的方法论打造支撑创新科学研究体系的新型信息服务模式，是摆在我们面前的首要问题。基于此，我们提出了基于研究报告，即研究成果应用的情报新型信息服务系统。如图8所示。

4.1.2 科技情报分析模型工具的开发

当科技情报工作研究的对象——数据信息量从TB走向PB的时候，传统时代的软件工具自然陷入尴尬境地。尽管Google的MapReduce支持在大规模集群上的大数据进行并行计算，但是基于此类计算框架结合我们所拥有的数据，还需要大量面向应用的研究工具软件和分析模型，如果将数据范畴跨越文献，结合上海量的网络用户行为数据分析，这种情报业务模式就更加超越了传统研究范畴，因此面对大数据我们亟需研究开发多种分析模型和可共享的软件分析工具。

4.1.3 科技情报工作全新应用场景的设定

情报工作一直以来都是研究信息与人之间关系的学科，在信息量小、类型单一的历史时期可以通过对信息的组织、整序后提取情报激发用户认知。但面对海量的大数据，过去简单的以文献为主的应用场景越来越无法满足情报用户的深层需求，甚至连情报服务人员本身也可能会“望数据兴叹”，换句话说：情报工作过去的应用场景已经发展变迁为今天的大数据复杂环境了，用户的信息知识困惑更加突出，情报人员要处理的数据信息对

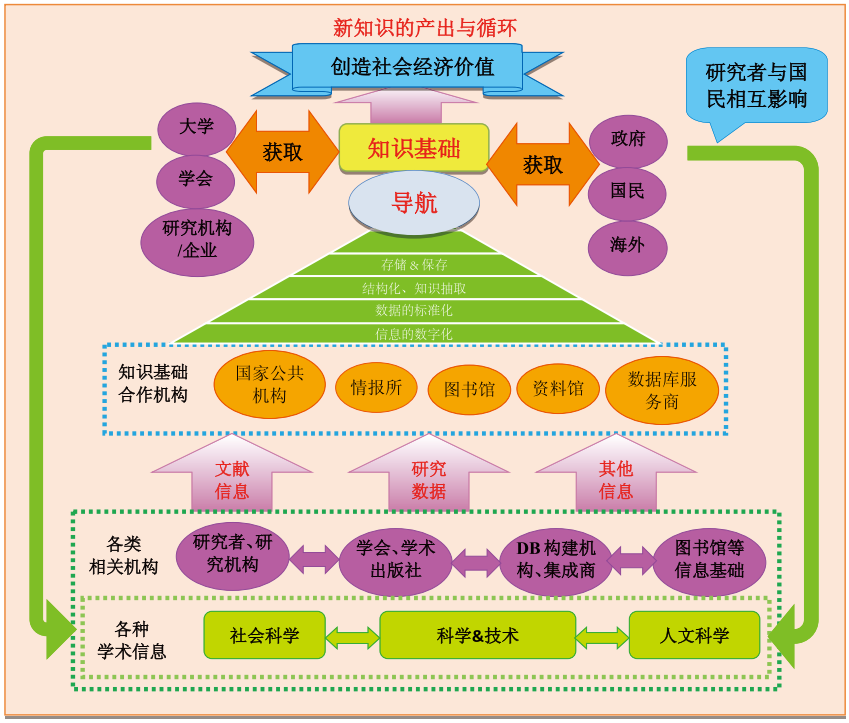

图9 大数据时代情报学国家创新服务体系

象更加复杂多样，但随之而来的却是各行各业对情报服务需求的如饥似渴。因此情报工作面临快速寻找和创新学科应用场景的现实问题，在大数据的机遇和挑战中完成学科的跃升。

4.1.4 科技情报工作与创新主体的结合

面对教育、科学、研究和产业发展两个完全不同类型的国家创新主体，情报工作必须深度融入教育、科学、研究以及产业发展的过程之中，承担起信息桥梁和成果转化的重任，唯有此才能打破研究与产业的深厚壁垒，逐渐消除研究与应用之间的信息鸿沟。但是，每一个细分领域都有其独特性，尤其是支撑国民经济发展的各产业，对信息情报的需求千差万别，如何在共性与个性之间求得平衡，既能真正为创新主体起到信息支撑的作用，又能不断发展学科的核心竞争力，值得思考。

4.1.5 国家科技情报服务体系的构建

经过多年努力，我国情报学教育、研究体系已经基本成型，全国形成了博士点、硕士点、图情档专业本科学校的教育网络，同时在全国各省、市、地区还有大约10万人的情报工作队伍，不包括企业中从事情报工作的不同岗位。如何以现有的基础为起点，借助大数据、云服务模式构建我国的情报信息服务体系也是当务之急，在顶层架构和战略驱动下，让国家科技情报服务体系成为创新战略的有效信息支撑，构想见图9。

4.2 科技情报工作未来发展方向的思考

4.2.1 学科定位

● 作为基础理论学科的情报学，应夯实理论基础，促进未来学术研究

面对新型超大规模系统的解决方案，以及开放性系统的分析方法等，情报学可以参考生物系统的开放性，追求一切可能的手段提升本学科的理论根基，结合科学研究的第3与第4范式，将情报学理论普及到其他广泛的研究领域中。

● 作为实践应用的情报学，应以“现实世界”为研究对象，切实体现情报学的社会价值

当整个社会进入信息化高速发展阶段之时，也是各行各业对情报学需求最为旺盛的时期，情报学应该抓住这千载难逢的机遇迅速转型，改变以往单纯以文献信息为主的研究对象思维，扩展竞争情报面向产业、面向企业实战的应用理论体系，开创学科新价值，在实践应用中探索新方法和新手段。以现实世界为对象的情报学需要安全的实时大型软件、高效开发系统的方法及机制，这些崭新的应用场景反过来也会将情报学的理论研究推升到新的阶段。

● 作为独立学科的情报学，应以“数据资产集中化”为手段，凸显情报学的学科地位

当数据已经被公认为资产和财富的时候，情报学就成为信息时代的资产管理者和财富创造或挖掘者。从创造空间价值的方法开发，到对社会经济有益数据的战略收集、保存和应用等，情报学可以通过“数据资产集中化管理”的理论、方法、工具、人才和众多最佳实践在多学科中凸显其战略地位，学习大数据、云服务的模式开创

属于情报工作的DAAS（data as a service）模式。

4.2.2 大数据应用思考

在全球数据爆炸性增长的同时，大数据中的大部分并没有在质与量上得到整理、结构化，无法获得有效的应用。情报工作者可以通过开展处理与应用的深度研发，以促进大数据的应用：①收集、存储、结构化各研究领域的多种观测及测量数据、模拟数据，研究开发大数据各个阶段的通用基础技术；②数字环境下科学研究相关的系统和机制研究；③以情报科学技术领域的研究者为中心，与环保、生命科学、防灾等有关领域的研究者在对等的立场上，建立领域间的合作，应用灾害领域有关的大数据来防灾减灾；④加强国际合作，建立培育具有大数据应用技能的数据科学家、数据应用专家、数据分析师的机制。通过这些工作，促进作为我国发展平台的大数据应用，努力解决新知识创造、科技创新以及社会与科学课题，通过创新来强化新型产业、新型市场的诞生和国际竞争力。

5 云计算关键技术与系统的国家863战略

为适应我国互联网用户规模和市场特征带来的技术挑战，以产业应用为目标，研究规划国家云计算发展的总体战略和整体部署，攻克云计算关键技术，形成标准，开展试点示范，确立了以科技文献服务为主的搜索引擎研制的国家863重大课题，以期在海量科技文献信息资源整合与信息组织、海量科技文献检索、海量知识挖掘分析、科技创新辅助决策分析、客户行为统计分析与个性化服务、云计算应用等关键技术方面取得突破，研制能有效支持大并发量、大用户量检索访问的基于云计算架构的科技文献搜索引擎，为政府、企业、科研院所、科技信息服务部门等提供科技文献云数据服务与云搜索服务，为科技人员提供科技文献搜索服务和科技创新辅助决策支持服务。经过一年多的研究实践，课题取得了一系列重要研究成果，在此组织专题与同行分享交流。

参考文献

[1] BUSH V. As We May Think [J]. The Atlantic,1945 (07).
[2] 化柏林,武夷山.数据为基础的情报学[J].情报学报,2012,31(6).
[3] 于艳华,宋美娜.大数据(1)[J].中兴通讯技术,2013,19(1):57-60.
[4] The NIST Definition of Cloud Computing [OL]. [2013-04-26]. http://www.nist.gov/itl/cloud/index.cfm.
[5] [OL]. [2013-04-26]. http://en.wikipedia.org/wiki/Cloud_computing.
[6] クラウドコンピューティング. [OL]. [2013-04-27].http://ja.wikipedia.org/wiki/.
[7] 王柏,徐六通.云计算(1-3)[J].中兴通讯技术,2010,6(1-3).
[8] OCLC: Achieving 'Web scale' with libraries[OL]. [2013-04-26]. http://www.oclc.org/content/dam/oclc/events/2010/files/Achieving_Web_Scale_With_Libraries.pdf.
[9] 王文清,等.CALIS高等教育数字图书馆技术体系[J].数字图书馆论坛,2013(1):29-36.

作者简介

贺德方，博士，中国科学技术信息研究所所长。E-mail: hedf@istic.ac.cn

The Development Trends of Informatics under Big Data and Cloud Service

He Defang / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: The article overviews the history of informatics and from the four paradigms of science, presents the relations between informatics, Big Data and cloud services. After briefly summarizing the characteristics of Big Data, cloud computing and services, it describes the challenge and opportunity of informatics facing. It presents our thinking of informatics under Big Data and cloud service with the cases of WorldCat, CALIS and NSTL.

Keywords: Informatics, Cloud service, Big data, Data science, Knowledge innovation

（收稿日期：2013-05-16）