气象大数据服务协同模式研究

聂峰英
(南京信息工程大学科技查新站，江南南京 210044)

摘要：契合大数据环境下气象服务的形势和需求，通过气象大数据服务协同模式来解决传统气象服务模式所遭遇的瓶颈。概述气象大数据特征，研究气象大数据的集成与服务，分析数据的采集聚合与处理，提出服务协同模式。构建面向用户体验支持可视化人机交互的气象大数据服务协同模型。依托大数据技术，通过协同优化服务，集数据采集聚合、专业化协同处理、用户管理于一体，实现气象服务的协同、整合、规模化，以解决气象服务“孤岛”现象。

关键词：大数据；气象大数据；数据集成；气象服务；协同模式
中图分类号：P409 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2015.05.010

A Study on Cooperative Service Modeling for Meteorological Big Data

Nie Fengying
(Nanjing University of Information Science &Technology, Nanjing 210044)

Abstract: There is an interest in the new situation and demands of meteorological services under big data, this paper established meteorological big data collaborative service model to solve the traditional meteorological services mode bottlenecks encountered. This paper overviews the characteristics of meteorological big data, analyze integration and service of meteorological big data, explore the collection, polymerization and processing of big data, propose service synergy model of big data. This paper constructed the services synergy model of meteorological big data which orients the user experience and supports the interactive visualization. In order to solve the meteorological service "isolated island" phenomenon, this paper proposed to use synergy and optimize services, integrated data collection and polymerization, professionally co-processing and user management, achieved synergy, integration, scale of meteorological services based on big data technology.

Keywords: Big Data, meteorological big data, Data Integration, meteorological services, Synergy Model

据中国气象报报道，“20世纪90年代及之前，气象资料大部分局限于地面及高空观测。当时，2000多个地面站以小时为单位收集气象信息，120多个高空站每天观测最多不超过4次。从数据量上看不算太多，即便考虑到卫星和雷达资料，其总体日增量也局限在GB量级。现在，地面观测站大约有4万个，每10分钟观测一次，未来还将加密至分钟级。在空间密度上，至少增

作者简介：聂峰英（1970-），女，南京信息工程大学副研究馆员，研究方向：大数据、情报信息服务。

基金项目：国家自然科学基金项目“基于供应链的产业绿色低碳多重耦合协同演进机制及政策研究”（71273140）；中国气象局软科学项目“大数据环境下气象信息资源协同创新机制研究”（气法函〔2014〕27号）。
收稿日期：2015年4月9日。

加20倍，频度将增加60倍，地面及高空观测信息总量增加了1200倍。而这些只占整个气象数据的30%，雷达、卫星以及数值预报数据占70%。目前，每年的气象数据已接近PB量级”[1]。气象部门每天的数据增长量有非常大的数据级，包括每天有2000多个地面站、120多个高空探测站、440多个雷达站、6颗在轨卫星、5万多个自动监测站、600多个农业监测站、300多个雷达站、90多个酸雨监测站[2]等。这些数据逐天逐小时甚至到逐分钟扫描着中国各种各样的天气数据，这些数据量大，且包括不同类型的数据类型。报告会专家表示，气象数据具备“大数据”的共性，即：数据体量巨大、数据增长速度快、数据类型多样等特点[3]。如何利用这些海量气象数据资源是当务之急。

1 气象大数据的特征

气象服务领域涉及比如工、农、渔、商、林、交通、运输、能源、水利、国土资源、海洋、环保、旅游、航空等多个行业和部门。可以说基本覆盖了国民经济建设和社会发展与国家安全各个领域。气象大数据具有如下特征。

（1）总量可控。这里对地面观测、气象卫星遥感、天气雷达和数值预报产品这四类体量最大的气象数据进行分析：地面观测资料数据量剧增的原因，是站点数的增加和观测频度的大幅加密。在观测台站达到一定密度，观测频度足以满足气象业务需求后，台站数不会无限制持续增加。因此，总量既是可预测的，更是可控的。对天气雷达而言，布网工作已基本完成，雷达总量不会成倍数地增加。而且，目前的天气雷达已基本实现7×24小时全天候不间断观测。因此天气雷达的资料量（年增量），将稳定相当长一段时间，而不会有倍数的增量变化。未来数年内，我国还将发射数颗气象卫星，每颗卫星都会产生数百TB级的数据年增量。为满足气象卫星资料的应用时效，国家卫星气象中心针对每一颗气象卫星，都建有相应专属的地面接收处理系统，已完全实现所有气象卫星遥测遥感资料的实时接收处理。因此，气象卫星数据目前虽以每年数百TB的量级增长，而且规模有可能继续扩大，但却始终处于可控可管和完全可用状态。数值预报模式产品资料是各级预报员最重要的预报参考资料，这些产品刚一生成，便即刻送达天气预报、气候预测专家的桌面，供其业务参考使用；同时，以满足业务需求的时效，分发至各省级乃至地市级气象部门，供其本地化应用。因此，与气象卫星资料相类似，数值预报产品资料体积虽大，却始终处于可控可管和可用的状态，未来也将始终如此。因此，气象资料体积虽大，在量级上算得上“大数据”,但却始终处于可控、可管、可用状态。

（2）内部信息单纯，来源单一。按照行业标准《气象资料分类与编码》，气象资料分为14大类，计有数百种之多。数百种的气象资料种类虽多，但每种资料所含信息却十分单纯：土壤持水量只记载某时某地某规定土壤深度中水的持有程度，“云能天”只记录某时某地的云量云状、能见度以及天气现象等信息。究其原因，海量气象数据是由气象探测系统以及数值预报业务系统产生的，来源比较单一。

（3）价值单一而明确。气象观探测业务系统只采集那些能够客观反映自然界气象状态的要素，所以气象观探测数据里包含且只包含丰富的气象信息，而以观探测数据为唯一数据和信息来源的气象数值模式，其生成的产品中所包含的信息也只能是局限于未来天气或气候状态的预测。因此，所谓“气象大数据”其自身的直接用途只能是气象业务，即天气预报、气候预测以及气象服务。

针对大数据环境下的气象服务，中国气象局公共气象服务中心高级工程师唐千红认为，“在看得见的未来，融入了地理信息、社会经济数据的气象服务，能够让人们知道任意时间地点可能会发生什么”。而中国气象局公共气象服务中心系统开发运行室主任惠建忠更是看到了大数据时代中气象部门的困境，“沿着气象服务社会化方向，光靠气象部门的数据很难满足各行各业及公众对气象服务的需求”。因此，大数据时代下的

气象服务绝不仅仅是提供气温、风力等简单天气信息，而是与民生息息相关的公共服务产品。但是，有效的信息协同共享公用机制还尚未建立，诸多基础性数据仍只在气象行业部门系统和数据库间使用，形成了大量信息服务“孤岛”。气象服务要主动跟上大数据时代的步伐，协同发展是未来趋势，便要借助大数据技术带来的契机，着力克服目前气象服务领域中存在的关键问题。

2　气象大数据下传统气象服务模式遭遇瓶颈

气象服务是一个大概念。黄宗捷等[4]认为，气象服务是指气象部门的劳动者对大气信息的采集、储存、加工、传输之后，所提供的超前服务。气象服务最基本的对象是政府和社会公众，此外，气象部门还针对不同行业的具体需求，针对经济社会发展的特定需求等提供气象服务[5]。在传统气象服务模式中，气象服务完全由政府投资，无偿提供用户使用，属单一总控投入模式（图1）。从图1可看出，传统气象服务模式严重缺乏实时性、动态性和交互性。

虽然传统气象服务模式为我国的气象事业快速发展做出了巨大的贡献，但随着大数据时代的来临，传统气象服务模式遭遇到了前所未有的瓶颈。

（1）传统气象服务模式主要集中在专业气象预报方法库、各行业专家经验预测库和专业气象预报产品库等预报服务项目上，而气象科技要求高的专业气象服务发展缓慢，还停留在以公共服务产品代替专业服务产品的水平上[6]。导致专业气象服务能力与用户需求差距越来越大，服务缺位的问题越来越突出。

（2）传统气象服务模式是政府单一向社会提供气象咨询服务，气象部门在保障基本气象业务正常运转的前提下，向社会提供最大限度的公共气象服务（常规气候资料检索服务及行业气候背景档案咨询服务）。而由于气象部门承担的政府职责是向社会提供均等的基本公共气象服务，不可能也不应该投入很多的资源提供全部的公共气象服务，这就使得基层台站普遍存在人员少，事务多而杂，各项工作疲于应付，很难深入开展各项资深咨询业务，导致公共气象服务能力与社会需求相差甚远。

（3）传统气象服务还存在气象信息不准确、产品供给不及时以及气象信息传递存在阻碍性等问题。预报气象变化的能力是气象科技服务的基础，但是这个过程很容易发生数值偏差，如果产品工作人员不能及时修订数据错误，就会使气象信息的精确度大打折扣。而当气象服务的供给不能及时，受到地域性的延迟和耽误时，导致气象

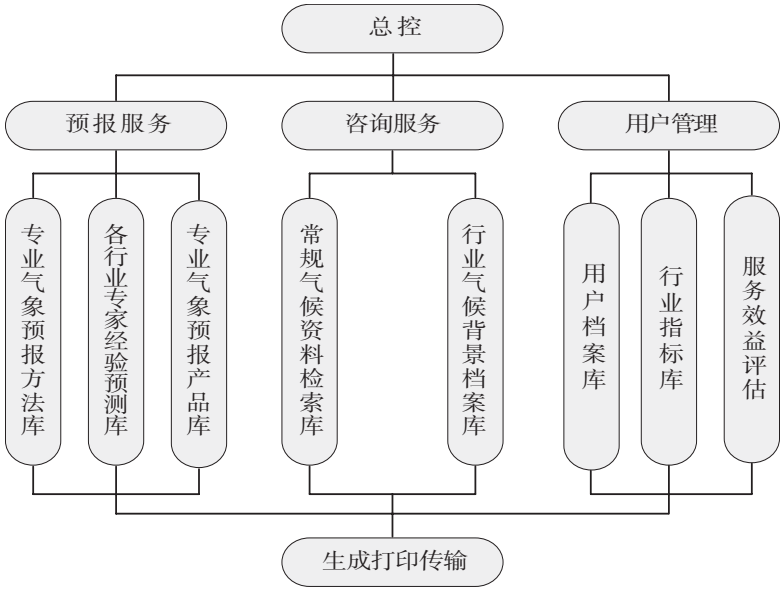

图1　传统气象服务模式

服务失去用户的信任。

显然，上述传统的单一总控投入模式，难以使气象服务满足大数据环境下各行各业快速发展的需要。气象服务应打破部门之间的信息垄断，通过政府数据开放应用，与用户一起形成协作网络，在共同分担社会责任的基础上共享公共资源，形成多元协同治理机制，构建一个以网络化互动为特征、大数据技术为支撑、纵横协调、多元统一的新气象服务协同模式。

3 气象大数据的集成与服务

据统计，当今世界结构化数据增长率约为32%，而非结构化数据增长率则是63%。至2012年，非结构化数据占有比例便已达到互联网整个数据量的75%以上。从这个数据统计中可以发现，结构化数据仅为大数据的一小部分，非结构化数据已为人们生活所依赖，更具持续性价值[7]。传统气象信息资源，其数据种类单一，以结构化数据为主。大数据时代，气象信息资源除具有基础数据通常的基础性、公益性、累积性等特点外，还具有全球性、多源性、实时性、海量性、分布式和异构性等复杂特点，并且非结构化数据的比例越来越多。非结构化数据的增长使得气象信息资源的数据采集、处理分析和服务方式都将得以改变，而传统的关系型数据库管理和机制不能很好地适应这种变化。因此，如何将这些气象大数据进行集成协同是急需解决的技术难题。

气象大数据不仅仅是指气象信息资源本身的庞大数据，还有气象关联活动及管理过程中所涉及到环境相关人财物的一切文件、资料、图表和数据等信息资源。也就是说气象大数据涉及气象监测、预报、预警和管理活动过程中所产生、获取、处理、存储、传输和使用的一切信息资源，贯穿于气象活动的全过程。如何有效地管理和组织好这些大数据？显然传统方式管理数据资源已不能适应大数据时代气象服务的要求，利用大数据技术进行气象信息资源服务协同便成为关键问题。因为，在当前大数据环境下，信息共享、交流互动已经不再是最迫切的需求，数据的采集融合和协同分析才是最大的挑战。目前，气象服务面对的是海量、模糊、复杂关联和动态发展的大数据，如何将不同类型、不同载体的气象大数据及其服务、系统进行有机结合，形成协同化、智能化的资源集合体，以提供更加便捷的气象服务？“全球有87.5%的数据未得到真正利用。其原因在于许多数据资源仅仅是简单汇聚而成，并没有形成真正的知识源”[8]。这就揭示了大数据时代气象信息资源的两个突出矛盾：一是面临着不断增长、大量的信息被“搁置”；二是面临着用户对检索结果提出更高要求，气象信息资源重新被发现和管理的挑战。面对海量的气象大数据，检索、分析和可视化是大数据面临的主要需求，如何将检索、分析和可视化以服务形式提供给用户是气象大数据服务需要解决的难题。

4 数据的采集聚合与处理

大数据技术是指设计用于高速收集、发现和分析从多种类型的大规模数据中提取经济价值的新一代技术和体系。涉及数据存储、合并压缩、清洗过滤、格式转换、统计分析、知识发现、可视呈现、关联规则、分类聚类、序列路径和决策支持等技术[9]。也就是说大数据技术可以忽略数据类型、物理和地理限制，实现信息资源逻辑上的联通和集中，轻易破解气象“信息孤岛”难题。从大数据的特征和气象领域不难看出，气象大数据的来源相当广泛，由此产生的数据类型和应用处理方法更是千差万别。因此，本文将重点针对气象服务领域对气象大数据资源的实际需求，研究气象大数据的数据采集、服务类型和功能，多源数据和专业数据集成、多类型数据集成、异构数据融合与海量数据集成，提出面向用户体验的、支持结构化和非结构化数据的、支持检索分析和可视化应用的气象大数据服务协同模型，以实现气象大数据价值的发现和数据洞察力的提升，最终为气象部门的管理和服务提供可靠的大数据决策支持。

大数据的“大”，即意味着数量多、种类复

杂。气象数据更是具有复杂性、多源性、地域差异性和丰富性等特点，因此，通过各种方法获取数据信息便显得格外重要，气象数据采集聚合便是气象大数据协同服务过程中最基础的一环。数据采集阶段的任务是以数字形式将信息聚合，也就是从气象领域知识中获得原始数据的过程，此过程中从多样的纵向或分布式数据源产生的大量的、多样的和复杂的数据集。通常，这些数据集和领域相关的不同级别的价值联系在一起构建覆盖面广泛、数据高度集成、数据关联表达的气象数据集。并与气象专业领域专家配合，构建不同数据集之间的数据聚合模型[10]，通过模型预测为各级用户提供咨询服务，将大数据关联的气象智能初步结果存入知识库，通过知识库技术进一步集成、提炼和分析最终形成大数据关联的气象中心知识库（图2）。

随着数据集的增长和实时处理需求的提出，对整个数据集的分析越来越难。需要设计一个集成技术实现跨数据集的智能处理模型，包含数据预处理及智能分析。因为气象数据来自不同的数据源，数据类型繁杂，而且由于数据集干扰、冗余和一致性因素的影响具有不同的质量,由此对数据的预处理和智能分析带来了一定的挑战。数据的预处理主要是完成对于已经采集到的原始数据进行去冗→过滤→清洗降噪→转换→集成为可信任的多源数据。在经过这一步骤数据的预处理与集成后，根据所需数据的应用需求对数据进行进一步智能分析，数据智能分析是整个大数据智能处理流程里最为核心的部分，是一个交叉学科研究领域，需要来自不同专业领域的专家应用各自的专业知识协作完成分析任务，通过融合→聚类→统计→挖掘→分析数据中隐藏的价值，最终形成专业数据（图3）。

5　协同服务模式

大数据环境下的气象信息服务具有分布、动态、实现多样性等特点，用户对专业气象服务的要求也越来越高，要求获取的气象信息更加丰富和精细化，气象服务产品更加及时、准确、直观、可视、可交互[11]。因此，智能化、个性化、互动化的气象信息服务将是大数据时代的努力方向。为避免形成一个个“服务孤岛”，通过协同优化服务数据，实现服务数据在不同部门应用之间的共享集成与服务协同，通过交互引擎和交互控制，实现气象大数据信息服务协同的人机交互服务。如图4所示，通过语音交互、信息呈现方式、多通道交互信息整合、人机交互软件和大数据可视化技术，重点依据服务请求研究服务动态组合→服务匹配→服务优化的信息服务协同过程、构建多类型－多层次－多水平用户服务模

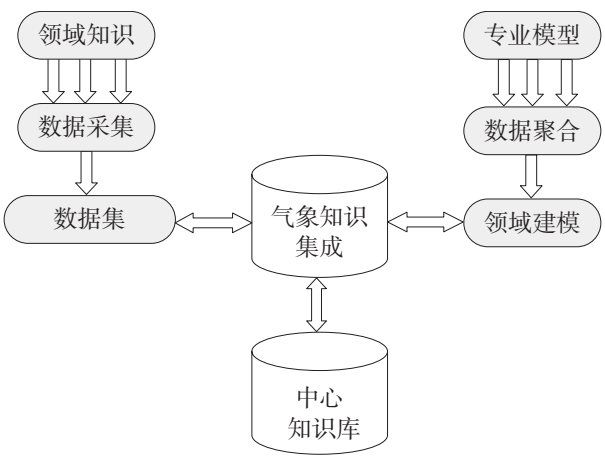

图2　气象大数据采集聚合模型

型，提供多级网络服务、普适终端自助互动服务和个性化新型的精准气象信息服务，最终利用可视化人机交互技术将结果形象生动的展示给用户。

6 结语

气象大数据服务协同是一个有现实前瞻性的复杂问题，涉及各种资源要素整合优化。尤其从气象数据采集聚合到数据预处理及智能分析，直到最后通过可视化人机交互技术的模型展示，均揭示了大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化协同处理。专业化协同处理不仅能够提供实时数据，供用户查看，按要求查询以及对数据能实时交互，更能缩短数据采集聚合处理周期，提高用户满意度。

传统的气象服务模式受技术条件限制，气象服务产品多以表格、文本等单纯的表现形式生成打印传输为主，如今依托大数据技术的发展，结合协同服务，便能形成直观生动的图形图像产品，并提供丰富的实时交互功能，能更精细化显示气象数据。同时，大数据技术也给气象服务协同提供了合理且有效的解决方法，服务协同模式集数据采集聚合、加工处理、显示发布、用户管理于一体，实现气象服务的协同、整合、规模化，能很好的解决气象行业部门条块分割、数据多源性、异构性等复杂问题。

但是在研究中发现技术上的差距并不难弥补，最大的差距是气象大数据思维。因此，后续研究惟有坚持资源、技术、思维三位一体协同发展，气象大数据的价值才能得到最大的发挥，用户的个性化需求也才能得到最大的满足。协同服务模式在气象服务中的应用，仅仅是一个雏形，

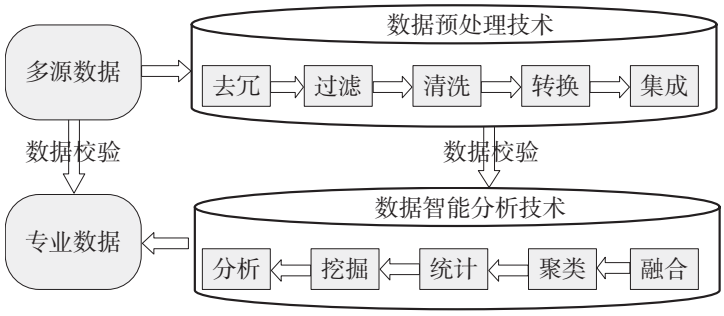

图3 气象大数据智能处理模型

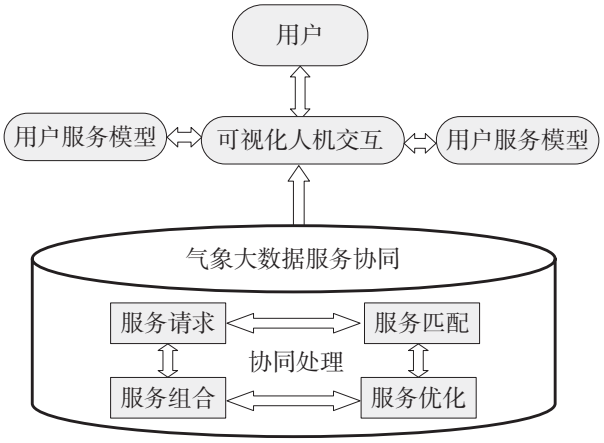

图4 气象大数据服务协同模型

仍然存在着一些问题。比如：新模式下气象服务效益评估？气象服务如何协同管理？气象服务模式如何根据气象服务发展水平适时改革调整？这些都有待于进一步研究和探讨。

参考文献

[1] 孙楠.气象部门怎样迎接大数据时代[N].中国气象报,2013-06-27(3).

[2] 张蕾,杨勇,湛莹莹,等.气象“云”气象万千[N].贵州日报,2014-03-26(9).

[3] 纪晓峰. 中国气象局：气象大数据的商业服务与研究[EB/OL].[2014-03-14]. http://www.ctocio.com.cn/cloud/73/12886573.shtml.

[4] 黄宗捷,蔡久忠.气象服务效益特征及其建模原则[J].成都气象学院学报,1996,11(1):33-39.

[5] 马鹤年,沈国权,阮水根,等.气象服务学基础[M].北京:气象出版社,2001:500.

[6] 矫梅燕.探索公共气象服务发展的体制机制创新[J].浙江气象,2009,30(4):3-6.

[7] 霍娜.非结构化数据来袭[EB /OL].[2013-12-19]. http://www2.ciw.com.cn /h /2562 /375443-17627.html.

[8] Big data: The Next Frontier for Innovation, Competition, and Productivity [EB/OL]. [2015-02-23].http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

[9] 邬贺铨.大数据思维[J].科学与社会,2014(1):1-12.

[10] Xie N F.Research on Agricultural Ontology and Fusion Rules Based Knowledge Fusion Framework[J].Agric. Sci.Technol,2012,13(12):2638-2641.

[11] 李超,胡耀文,甘建红,等.一体化专业气象服务集成系统[J].成都信息工程学院学报,2014(2):161-166.

中国科学技术信息研究所在“发明人名称消歧竞赛”中取得优异成绩

【本刊讯】 2015年9月24日，美国专利与商标局（USPTO）首席经济学家办公室在美国弗吉利亚州USPTO总部举办了旨在提高现有专利发明人名称数据精度的“PatentsView专利发明人名称消歧技术研讨会”。会议期间，举行了“专利发明人名称消歧竞赛”。此次竞赛的目的是通过设计专利发明人名称消歧算法，对USPTO收录的近40年（1976—2014年）的美国专利发明人数据（约1239万条记录）进行唯一标识，以改进现有的专利发明人标识算法。中国科学技术信息研究所派出代表队参加了这次竞赛，并凭着在预赛和复赛阶段的突出表现，取得了第二名的优异成绩。

参加本次“专利发明人名称消歧竞赛”的代表队分别来自美国、比利时、澳大利亚、德国、中国等国家的高等学校和科研机构。他们是宾夕法尼亚州立大学（美国）、马萨诸塞大学（美国）、加州大学圣巴巴拉分校（美国）、鲁汶大学（比利时）、斯文本科技大学（澳大利亚）、欧洲经济研究中心（德国）、中国科学技术信息研究所（中国）等。在竞赛中，中国科学技术信息研究所代表队提出了一套全新的发明人消歧混合算法（Mixed Method）。该算法融合了机器学习方法、概率记录链接方法、规则分类方法以及图聚类方法。其核心思想是：通过机器学习以及概率链接方法首先划定整个发明人名称匹配对的核心区域，然而通过加入分类规则逐步扩张发明人名称匹配对的外部边界，从而在保证计算结果的高准确性同时，兼顾了整体算法的稳健性。该算法在AWS平台C3.8xlarge实例上的运行时间为7小时。经过3轮共计20万数据集的测试，该算法的平均精准率（Precision）达到99.52%，平均召回率（Recall）为88.96%左右，平均F1值为93.94%。中国科学技术信息研究所代表队算法的最终测评结果也优于PatentsView平台目前正在运行的算法。

PatentsView（http://www.patentsview.org/web/）是由USPTO首席经济学家办公室主持开发的一个面向未来的专利检索与分析平台。该平台是以提高美国专利数据价值功能及实用功能为目的的可视化分析平台，是USPTO为实现其数据透明化，便利创新者、知识产权从业者、企业及个人利用专利数据而开发的搜索工具。专利发明人消歧问题是目前学术界关注的热点问题，通过对发明人名称进行消歧能够提升现有科研绩效评价、社会网络分析的准确度，也可以为国际人才流动、知识溢出等问题提供更为准确的数据支持。中国科学技术信息研究所代表团队取得的研究成果将为相关方面的研究工作提供更为精确的数据支持。

USPTO全程直播“PatentsView发明人名称消歧技术研讨会”。研讨会的视频已上传网站，敬请收看。视频地址为：http://www.uspto.gov/about-us/organizational-offices/office-policy-and-international-affairs/patentsview-inventor。（**杨冠灿**）