



知识产生方式和科技决策支撑 的重大变革

面向大数据和开源信息的科技 态势解析与决策服务*

文 / 王飞跃

中国科学院自动化研究所复杂系统管理与控制国家重点实验室 北京 100190



中国科学院

【摘要】 文章简述了面向大数据和开源信息的科技态势解析与决策服务的理念、概念及基本的系统框架和技术体系。主要强调科技态势及决策支持的相关工作应借鉴情报安全信息学的最新进展,加快变革科技信息的获取、分析、转播及影响方式,实现从科技信息向科技情报,并迅速向科技解析的转化,进而及时和变革性地提升科学知识的产生与教育方式,以及科技决策的制定与实施手段。本文讨论的方法亦适用于支撑经济社会其他领域的重大决策。

【关键词】 大数据,开源情报,科技情报,研究解析学,ACP方法

DOI: 10.3969/j.issn.1000-3045.2012.05.001

随着 Cyberspace 的兴起,网络化及其相关技术的普及与深入,我们迎来了一个大数据的时代。人类社会的主流,已从利用地表资源的农业社会历史、挖掘地下资源的工业社会现实、正迈向开发数据与智力的智能社会未来。

显然,Cyberspace 和大数据对于人类社会的重要性,将远大于任何一项技术的出

现,因为两者的结合,已带来人类的智力爆炸,并将变革性地重新界定人类智力的空间、内涵及其作用。我们必须从这一个角度和高度重新审视未来科技的发展,因为一个社会的科技水平是这个社会智力程度的直接体现,当智力发生重大变革时,科技首当其冲,必将随之而变。因此,我们必须努力,抓住这一历史性的变革时机,用更广、更深、更新的思路来考虑面向 Cyberspace 和大数据之科技发展方针的制定与实施,创新科技态势的分析与科技决策的支持方法,更好地服务国家科技体系的建设与发展,尽可能地

* 本研究得到国家自然科学基金委 平行管理 重点项目(71232006)和中科院决策科技支持系统创新团队建设资助
修改稿收到日期 2012年9月13日

发挥人类智力的积极作用,促进人类社会迈向一个崭新美好的辉煌时期。

从表象而言,Cyberspace下的大数据一般可用4V概之^[1]:(1)从体量(Volume)上,极大极小的数据都要考虑,而且,正因为极小数据的海量汇集,成就了大数据,故大数据之大,本质在小;(2)从类型(Variety)上,数据品类将极其混杂,关联度一般极低,而且在相当长的时期内非结构化数据会占据大数据的主体;(3)从时速(Velocity)上,数据的快速和实时化是其主要特征,但更重要的是数据的全时全速覆盖,历史的、现实的、未来的、极慢的、极快的数据等等;(4)从价值(Value)上,通讯是大数据意味着极高的价值,但同其体量一样,正是因为极小价值的海量汇合,才铸成了大数据的高价值,故大数据之大价值,实质在小,是典型的长尾效应,更是长尾效应的常态化。总之,大数据之大,本源于小。

从本质而言,我们认为,Cyberspace背景之下的大数据时代的实质可借用现代管理科学两位奠基人的名言概之:首先是戴明(W. Edwards Deming)所言:除了上帝,其他任何人都必须用数据说话,其次是德鲁克(Peter F. Drucker)之言:预测未来最好的方法,就是去创造未来。Cyberspace和大数据的出现,引发限制智力的认知、信息及资源等要素发生重大变革,并使这两句管理的格言成为可行的技术目标。

以此为指导,本文在已有实践的基础上,针对面向开源信息和大数据的科技态势评估与决策支持,抛砖引玉,围绕下列问题展开讨论:

- 如何变革科技信息的获取方式?
- 如何变革科技信息的分析方式?
- 如何变革科技信息的影响方式?
- 如何变革科技知识的产生方式?
- 如何变革科技决策的制定、评估、实施方式?

2 国内外发展现状

400多年前,中西文化交流的先驱徐光启试图将西方科学引入中国,遭遇封建士大夫的阻力,许

多人认为科学无用,更无助于科举;在刻《几何原本》序中,徐光启的回答是:盖不用为用,众用所基。以网络信息为主形式的开源信息,虽然被认为是印刷出版发明后知识之产生与传播的又一场革命,但其价值仍未被充分地认识,在过去相当长的时间里,对于正式渠道的科技态势评估和决策支持而言,基本还处于“无用”状态。

然而,综观国内外相关研究近况,指导思想就是:无用之用,众用之基;即以开源信息为主,汇集海量数据,然后通过定量的方式来描述、分析、评判科技发展的态势,服务于科技决策。而且,评估科技态势的手段及工具近年来呈跳跃性发展,目标要求也越来越高,正从科技信息向科技情报,进而迅速向科技解析(Academic Analytics或Research Analytics)转化(图1)。



图1 科技态势分析的发展趋势 科技信息、情报、解析

首先是国际上科技态势与科技决策的结构变化,即群体科学(Team Science)及相应的群体研发科学(Science of Team Science,SciTS)的兴起^[2]。SciTS是一个跨学科交叉研究领域,专注于科学研究群体如何发起、组织、交流、从事科研活动等的过程,其核心任务是理解认清促进或阻碍各种各样科技合作的情况和阶段,并加以有效管理和利用:从确定大规模科技合作研究、培训、转化的有效性,直到掌握研发群体是如何协力攻关取得靠

个人或递增式研究难以实现的突破和跨越。目前,推动该领域的主要团队是美国一批跨学科交叉研究的学者,他们认为:SciTS是促进21世纪科学合作的一盏指路明灯(A Beacon for 21st Century Scientific Collaboration)。

国际上,学者们确认迄今为止已有180多篇与SciTS相关的核心文献。其中除了17篇发表于1944—2000年之间外,其余皆为2001年之后发表的。这一发现充分表明:SciTS作为一个新的领域正在兴起,关于SciTS的研究热潮即将或已爆发。100年前,人类经历了一场由几位天才个体物理学家诱发的科学风暴;百年后,我们或许正处在另一场由群体唱主角的科技风暴之前夜,而SciTS正是指引这一风暴路径的灯塔。

美国学者对从1945—2008年全世界发表的2100万篇论文进行研究,结果发现几乎所有科学分支都呈现出一个根本性和普遍性的变化:在产生高影响、高引用科学成果时,群体比个体科学家越来越居优势,群体变得越来越大,群体越来越跨界超域。他们还就世界范围内发表的专利进行研究,结论也十分相似。学者们认为:造成这一变化或转移的原因是我们研究的问题越来越复杂,涉及交叉学科知识,超越传统学科边界。因此,解决这些复杂问题必须从许多不同的学科视角去考量,而关键与核心就是跨学科交叉的研究群体,其采用的方法多半是多学科(Multidisciplinary)、交叉学科(Interdisciplinary)、跨学科(Transdisciplinary)的,有时还是它们的综合。评估科技态势或进行科技决策时,必须考虑这一时代特征。

其次,SciTS的研究必定要沿着多学科、多层次、多方法的思路和框架进行,而且与传统的方式有所不同:(1)其研究必然涉及人文、社会、科学、技术和工程等领域的知识

和方法;(2)可以在以群体的集合为对象的宏观层次,以群体本身为对象的中观层次,以及以群体的成员为对象的微观层次展开研究,在不同层次上发现问题、寻找相应的解决方案、常见模式、普遍规律和进一步的挑战等等;(3)在各个层次和不同问题上可采用多种方法,如海量数据挖掘、统计学习、复杂网络、模式分类、问卷调查、背景分析、可视化等等,对群体的形成、活动、过程、结构及功能从时间(When)、地点(Where)、任务(What)、人员(Whom)、如何(How)和为何(Why)等方面进行研究考察。在设计、研发新型的科技态势评估和科技决策服务的方法和系统时,我们必须充分考虑这些因素和相关模型。

此外,美国近年来已展开对SciTS相关科技项目的大力支持。比如,NSF启动了网络化支撑的发现和发现(Cyber-enabled Discovery and Innovation)项目CDI,围绕:(1)从数据到知识;(2)认识自然、人工、社会系统的复杂性;(3)虚拟组织等主题长期地对优秀的跨学科交叉研究群体进行支持,目的是在计算思维的引导下,通过大胆的跨学科交叉,诱发能够根本性改变科技水平的研究成果。NIH对此的支持力度更大,其临床和转化科学基金(The Clinical and Translational Science Awards, CTAs)的主要目标就是培育众多能将实验室的科学发现转化成临床治疗手段的跨领域研究者群体。正是CTAs对以美国西北大学为主的有关群体之资助,才使SciTS具体化成为一个正在兴起且实实在在的跨学科交叉研究领域。2010年4月,CTAs还资助了首届SciTS国际大会。此次大会是在2006年NIH癌症研究所的SciTS会议的基础上召开的,与会者来自许多不同的领域,极大地促进了SciTS向一个独立研究领域方向的发展。



中国科学院

在欧美,与SciTS直接相关的几个正在开展且较有影响的大规模群体研究项目包括:(1)构建国家甚至国际层面科学家网络的VIVO项目,由NIH资助,以康奈尔大学为主的7家单位、4个群体、120位研究人员参加;(2)构建植物科学网络化基础合作设施的iPlant项目,由NSF资助,以亚利桑那大学为主的6家单位、6个群体构成核心,强调项目和环境的开放性;(3)通过网络化、虚拟化、社交媒体化,构建能够使知识的产生、传播、评价、维护过程发生范例转移般革命的LiquidPub,由欧盟第七框架下的未来与新兴技术(FET)计划支持,以意大利Trento大学为主的8个机构参与,涉及出版企业、多个国家和国际学术团体。类似于iPlant,法国也开启了PI@ntNet的研究。这些项目利用自然语言处理、本体建模方法、语义网技术等,针对科技知识产生过程中的人、物及出版过程,以跨学科交叉群体的方式,从SciTS的视角,正在开展扎实且有成效的工作。我们认为,这些工作将在科学知识的获取、分析、影响及产生方式等方面诱发重大变革。

中国是世界上最早开展SciTS实际工作的国家之一。而且,我国的管理体制更有利于群体的组织与发展,国家自然科学基金委员会的创新研究群体就是一个典型案例。与VIVO、iPlant、PI@ntNet、LiquidPub相似,我们也开始了iCAN、cPlant、PI@ntWorld和AI 3.0等项目^[1-2]。然而,我们的工作源于情报安全信息学(Intelligence and Security Informatics, ISI)^[3],起步于社会态势的评估,并发展成为关于社会计算和平行系统的研究。2009年,在中科院战略规划局的支持下,我们开始研发科网系统,力图在科技情报的获取、科技信息的传播、科技成果的影响、科技知识的产生方式上实现根本性的变化。

3 面向大数据和开源信息的科技态势解析与决策服务

互联网和信息技术的迅猛发展,使得科研工作者的互动和交流愈发便捷,同时开源科技情报亦呈现出动态性、海量性和多源异构性等特征,对开源科技情报的有效监测与分析,将有助于掌握科研现状,挖掘出科研项目在开展过程中的关键影响因素,构建出现代科技创新的倍增器和科学决策的智囊团,进而从根本上改变传统科技的采集、传播、影响和知识的产生模式。

因此,我们融合社会科学、情报科学、信息科学等学科理论提出基于ACP方法的开源情报解析理论框架(图2),并以天网(科情信息监测)、天眼(科情信息分析与挖掘)、天鹰(科情信息管理与决策支持)为指导流程(图3),构建面向大数据和开源信息的科情监测与协作创新平台,即科网。在异构、多模态的科技情报资源之上,生成科情获取传感网络、科情语义网络、科研协作社会网络3个核心网络,实现智能采集、处理、分析与挖掘科技情报等功能,有效地组织并发挥研究群体的效率,实现科技的跨越性发展。

科网提供以下4方面服务(图4):

(1)对科技领域科情动态、研究报告、专利摘要等资料信息一搜到底,并通过科情传感网络以

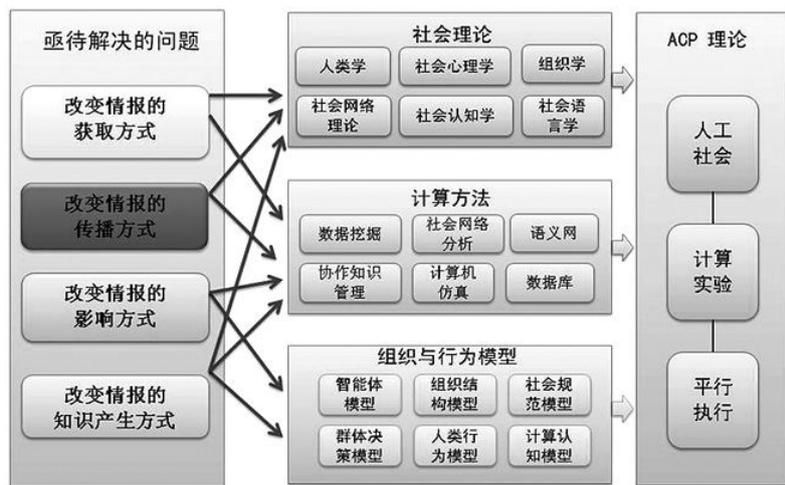


图2 基于ACP方法的开源情报解析框架

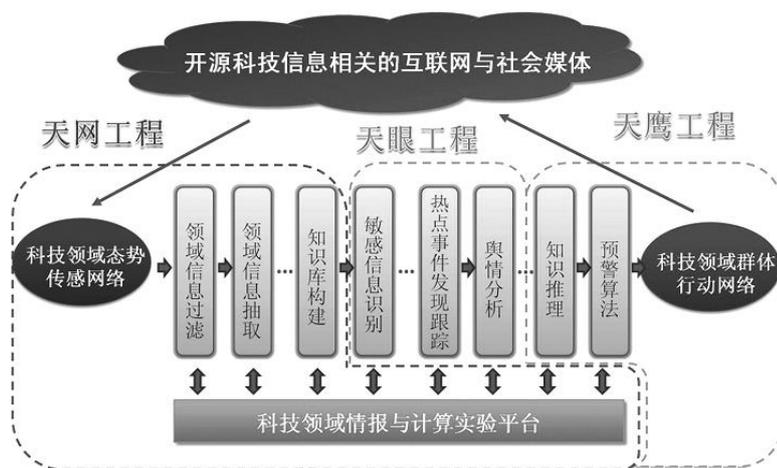


图3 面向开源信息和大数据的科技态势解析与服务流程

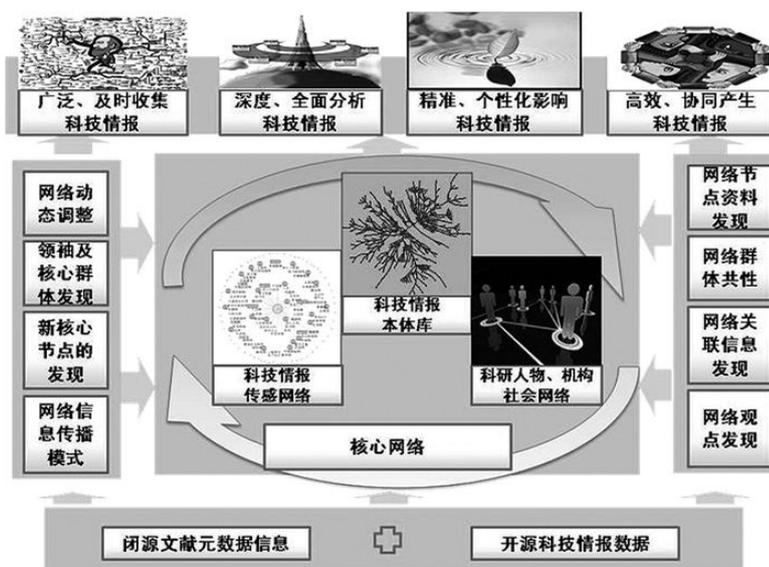


图4 科技态势解析与服务系统结构

及 ASKE (Application Specific Knowledge Engine) 机制,自动构建领域专家资料库,有效获取各类科技情报信息;

(2)对科技情报进行深度分析及挖掘,构建科技情报领域人物、组织、机构的复杂网络,分析挖掘其核心群体、关键人物,依照时空多维度分析科技传播态势,构建全方位的情报分析指标体系,对科技事件、科技要闻以及科技态势进行全方位分析,以辅助科研群体进行决策;

(3)对科技态势信息与分析结果,可提

供每日科技要闻日报、科技要闻专报等个性化服务,实现可定义、可配置、可交互、可快速上报的自动采、编、分析、发布一体化简报平台,也可提供国内外每日科技要闻简报的定制服务;

(4)整合各领域、各学科科研人员关注的信息资源,为科技工作者和学术机构提供个性化推荐,支持科研人员学术互动,使科研人员能够在线发表学术观点,交流创新思想,对科研各个阶段性产物进行信息共享、发布、版本控制,实现快捷高效的协同创新。

4 科情监测与协调创新平台的原型建设及讨论

4.1 变革科技信息的获取方式

根据中科院决策科技支持系统创新团队正在建设中的科情监测与协调创新平台原型系统,我们按下面5个方面就其关键技术和主要功能简要地描述并讨论如下。

ASKE (Application Specific Knowledge Engine, 图5)是将领域专家和丰富的互联网络领域资源相融合的知识库构建的有效机制。通过该机制,与用户交互生成网络领域知识配置文件(KCF, Knowledge Configure

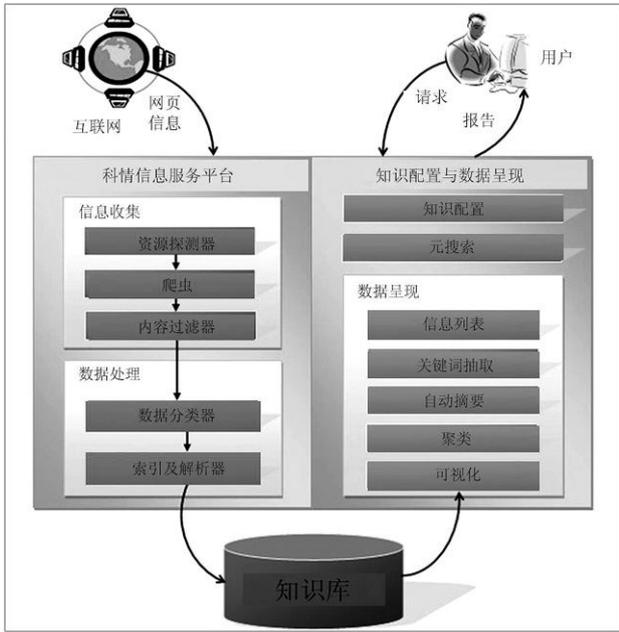


图5 领域专用知识引擎 ASKE (Application Specific Knowledge Engine)

File), 准确获取领域科技文献、科技信息等数据资源信息后,应用自动向量识别技术进行特征科情相关实体抽取与分类,在人机结合的少量干预下形成分类的语义特征向量列表,然后应用有监督的统计学习方法学习语义进行科情实体间关联关系的自动构建,最终构建科情相关领域专家知识库。

面向各类不同领域科情的开源情报获取与处理的需求,我们将物理传感器概念推广到网络空间,传感网络中应用垂直搜索和聚焦爬虫、深度网和动态网信息的采集、网络信息的可配置搜索等技术,构建科研机构、任务、组织等各类不同科情传感网络(图6),分层、分簇感知科技要闻、文献、专利等涉及科技最新动态的网络开源信息,并且该科情网络不断扩充、丰富,同时各个传感点根据信息内容不同动态调整网络中的关联与位置,以达到最快感知领

域科情关注焦点信息的功能。

4.2 变革科技信息的分析方式

科研事件从最初发现到广泛传播,再到逐渐消退中的各环节,需要多层次、多粒度的分析方式将整体态势进行刻画。我们提出科情监测、科情云热点、科情云数据、科情云指数的一体化分析解决方法(图7)。其中科情监测实时汇总海量科研情报数据、科情云热点主动发现热点词簇并体现为滚动可视化形式,科情云数据则根据热点检错呈现相关人物、机构、成果等信息,科情云指数则以定量的方式进行多维热点的分析,通过上述方式可有效、全方位跟踪科技热点事件及热点研究话题的产生、发展、消亡全过程态势。

构建一套可行的、有效的、易于后期计算评估的指标体系是研究科情监测与决策支持的前提条件,熵的概念可以从作用结果的角度来反映科情发展态势,从而有助于观察个体事件对全局局势的影响,科研态势解析提出了四熵的分析方法(图8)。其中,量熵是对科技开源情报信息总量的度量,反映互联网科技事件、科技成果网络反应总体态势,情熵是对科技情报支持度的分析,反映科技人物、科技实验在网络中情绪的分布,强熵是

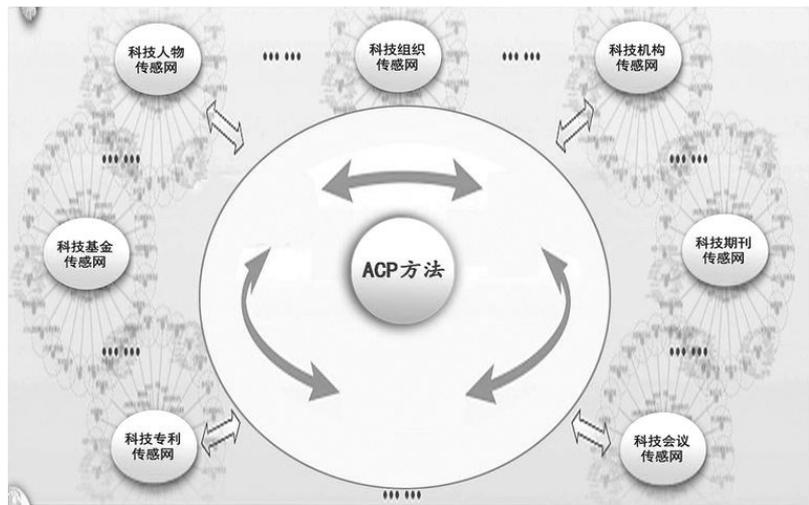


图6 科情传感网络

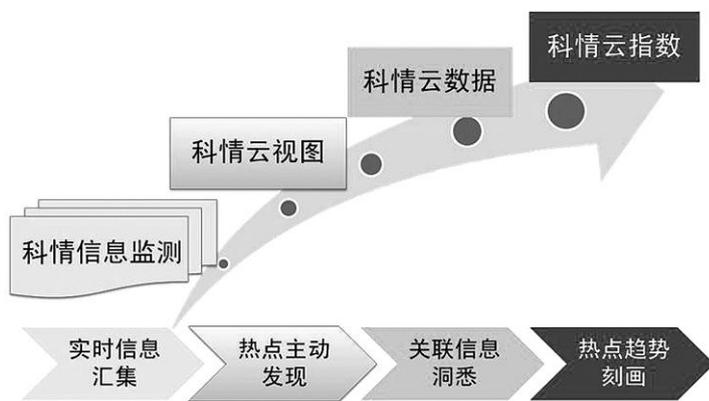


图7 科技态势解析的主要任务和方法

的联系紧密程度,或通过在科情网络中移除某些人物与机构,分析其对网络的连通性的影响,分析其重要性信息,或基于机器学习中的相关排序算法,发掘科情网络中的关键人物或组织。

4.3 变革科技信息的影响方式

科情专题事件分析与简报自动定制平台通过对科情事件的发现和追踪,自动筛选出热度较高的事件话题,并对该话题事件进行分析探测,对事件的发展状态、趋势等相关信息进行推荐,以引起用户的重视和关注,并提供事件报告自动生成工具,该工具

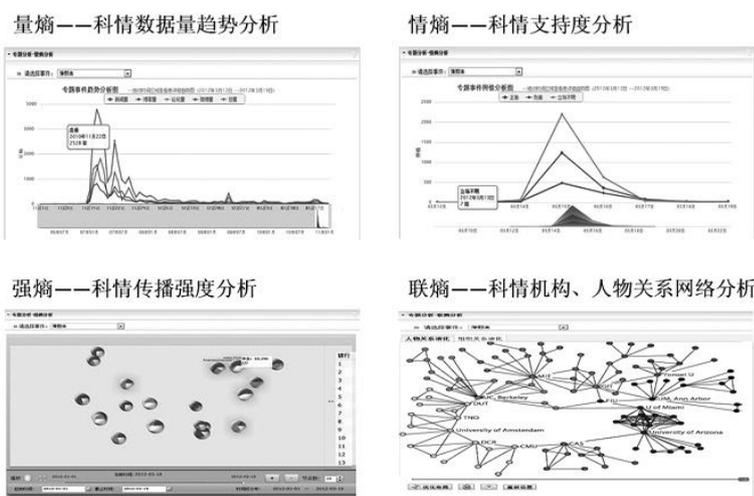


图8 科技态势分析的 四熵 方法

对网站活跃度的分析,反映不同网站在不同科技事件传播中的影响力,联熵是对科情机构和人物关系网络的直观可视化,可以直观地呈现机构或人物中活跃人物、兴趣群体等,展现领域学术研究的宏观合作态势。

科研合作网络中存在意见领袖、关键人员、活跃群体、兴趣网络等特性,科研合作网络在有效抽取网络实体及其特性后,可视化构建不同的合作网络、兴趣网络(图9),同时采用中心性度量的方法,分析人物与组织在科情网络中的关联关系,提取科情网络中最活跃的一部分节点,研究这部分节点之间

支持量变图、舆论图、关联图、人物、组织、事件关联图的分析 and 自动生成,而且能添加到事件专报中进行发布。

平台的具体工作流程如图10所示,分为配置层、服务层、系统集成层3部分,包括数据配置、分析配置、数据采集服务、数据分析服务、特定事件情报库和系统服务引擎等方面。

同时,简报平台提供领域知识的科技信息智能推荐,形成可定义、可配置、可交互、可快速上报科技信息的自动采、编、分析发布一体化系统,并根据用户的特点和类别的



中国科学院

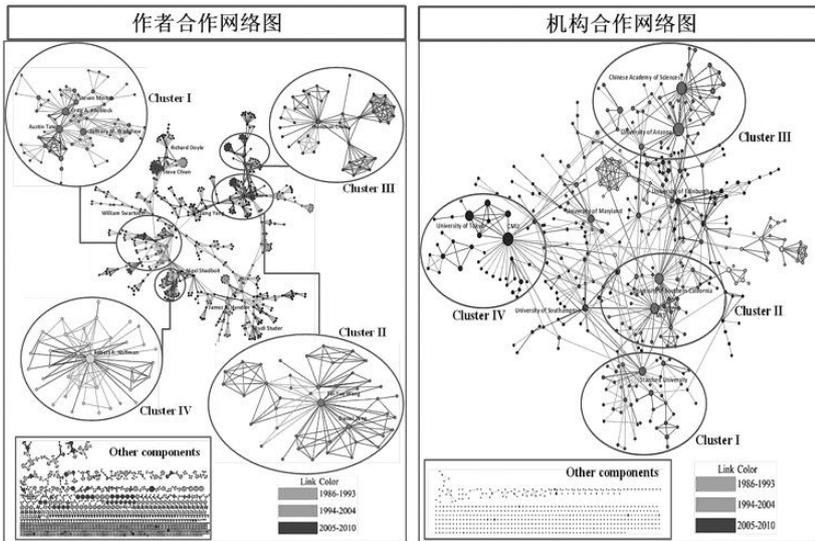


图9 科技文献作者与机构的社会网络示例

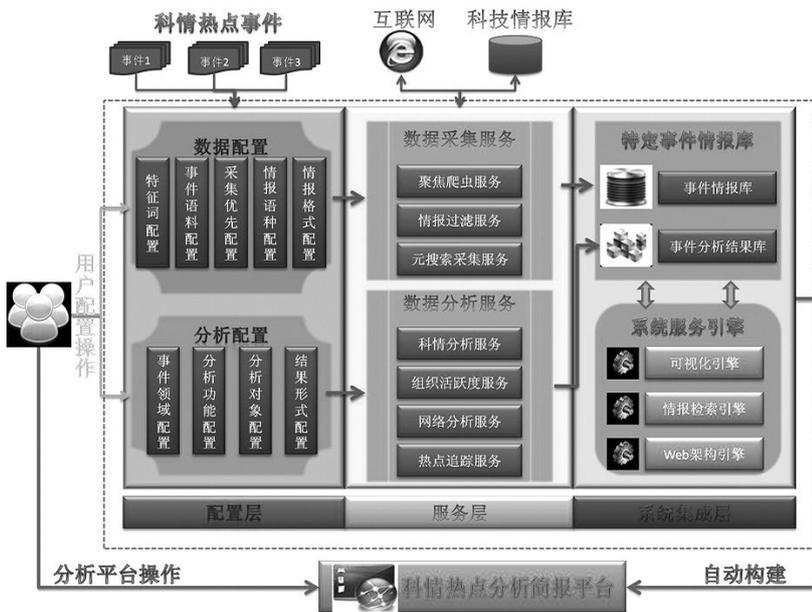


图10 科情专题事件分析与简报自动定制

不同,实现简报的个性化定制和智能推送,如图11所示。

4.4 变革科技知识的产生方式

面对互联网中海量的科技情报、文献知识创新素材,科研工作的发起、组织、交流、成果应用等过程除需要精准的分析与挖掘结果外,更需要融合群体智慧对知识素材进行深度加工与提炼(图12)。众包作为一种群体智慧或群体智能,使得更精细的知识从众多科研人员、领域专家的合作与

竞争中涌现出来,更开放、更专业、更低成本地为科技协同创新提供新动力,极大地辅助大规模科技合作研究、培训、转化,从而取得个人或递进式努力难以实现的科研突破和跨越。

开源科技情报中蕴含着大量的可提炼知识,对闭源知识起到了良好的补充、对比和验证作用,但开源科技情报局部相关、总体无序状态,需要进行结构化地抽取提炼,借助众包和领域专家指导相结合的方式,建立与闭源知识对象的索引和相互关系,组建一个以人员、组织、机构、科研成果为核心的科情领域知识库,构建科研人员专用的 Academicpedia,从而得到更广泛、更深层的知识,同时根据保密的需要将平台分为公共共享平台和闭源共享平台,以便于科研人员之间的交流、协作,实现科研成果的快速生产、转换和共享(图13)。

4.5 变革科技决策的制定、评估、实施方式

基于 ACP 的科研决策支持与服务系统由三步曲组成:第一步,利用人工科技决策系统对复杂科技决策环境进行建模;第二步,利用科技决策计算实验对复杂科技决策现象进行分析和评估;第三步,将实际科技决策系统与人工科技决策系统并举,通过实际与人工之间的虚实互动,以平行执行的方式对科技知识创新的运行进行有效地控制和管理。图14给出两系统进行平行互动的基

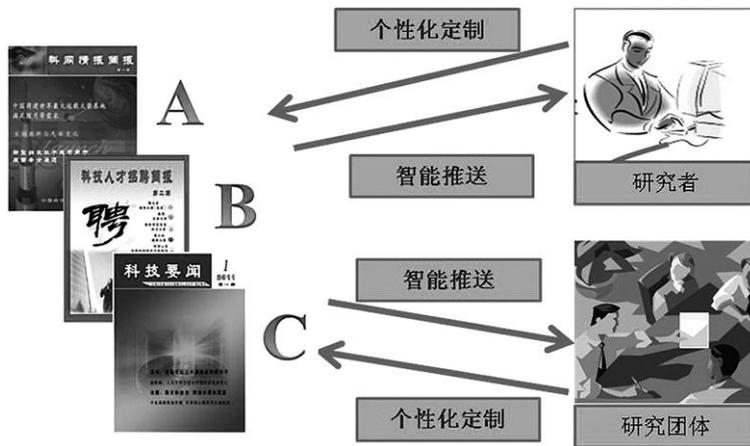


图11 科情简报个性化定制与智能推送

人工科技决策系统第一时间了解相关研究课题的全貌;(2)实验与评估,科研人员可在人工科技决策系统中进行科研创新影响力计算实验、科技协同创新实验等等实验,并对其效果进行评判和预估;(3)控制与管理,科研人员及时获得自身科研成果的评估,通过人工系统与实际系统的互动,互相借鉴,以完成对科研决策支持全流程的有效控制与管理。

5 展望

印刷术的发明和普及,是人类历史,特别是知识传播和产生方式的一场重大革命,而社交媒体特别是互联网的出现,是另一场更加强烈的革命,必将引发知识和信息在传播、获取、影响、作用等方面的重大变革,我们必须积极应对,这就是本文讨论的基本出发点。

大数据之大,源于信息的开源。随着大数据的海量般地不断增加,相信不久的将来,每个人都必须依靠特定的深度精确的情报系统来

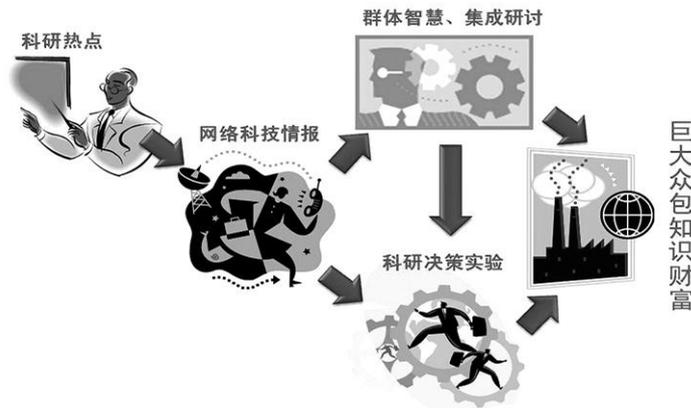


图12 利用知识众包协作进行知识创新

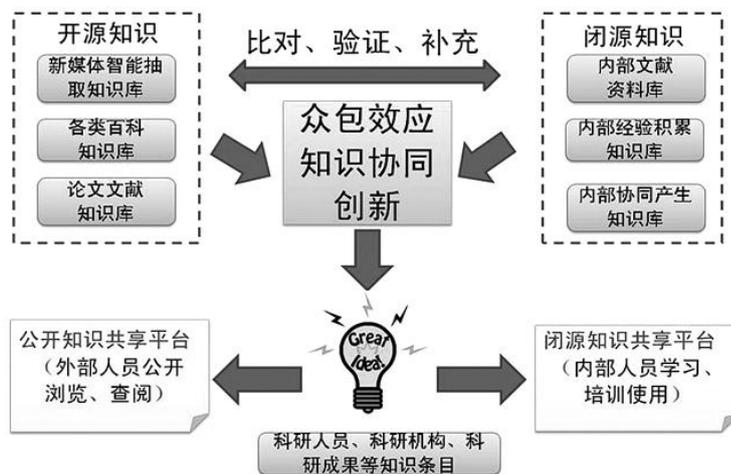


图13 科研情报百科知识库

框架。在此框架之下,可有3种主要的运作模式,即:(1)学习与交流,科研人员可通过

了解外部世界并与之互动,而不是靠简单的网上搜索系统。在大数据时代,科技态势的

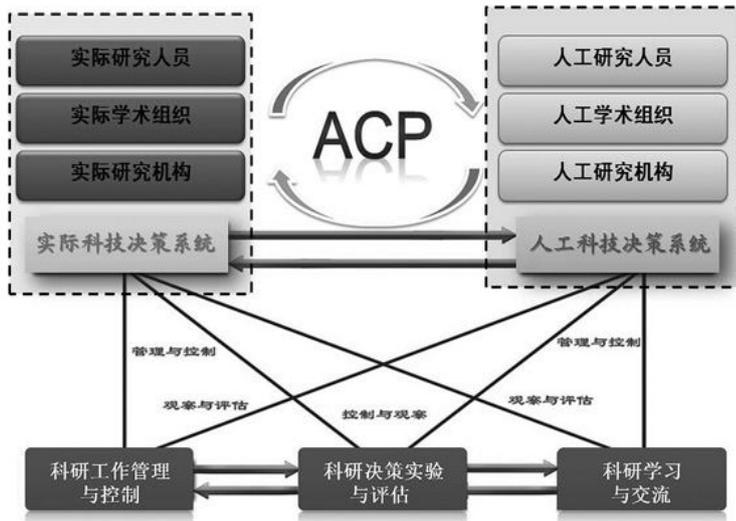


图 14 基于 ACP 的科研决策支持与服务系统

评估必须从科技信息、科技情报向科技解析(Academic Analytics)转化,其中 Descriptive Analytics 以描述现状为主, Predictive Analytics 以预测趋势为主,而 Prescriptive Analytics 以规划目标为主。总之,无论是事实、可能、希望,都必须以数据说话,而且,最终的目的,是实现预测未来,不如创造未来,这就是我们采用 ACP 方法研究和开发面向大数据和开源信息的科技态势解析与决策服务系统的动机。

显然,本文讨论的方法亦适用于支撑经济社会其他领域的重大决策。

致谢 本研究得到国家自然科学基金委 平行管理 重点项目(71232006)和中科院决策科技支持系统创新团队建设项目资助,部分论述直接采用了作者的有关内部工作报告和未发表论文的内容,特此说明。

主要参考文献

- 1 王飞跃. 科研情报与社会媒体 学术情报、会议情报、杂志情报. 中科院复杂系统与智能科学重点实验室工作报告, 2008 年.
- 2 王飞跃. SciTS :一座 21 世纪科技合作的灯塔? .科技导报, 2011, (12).

- 3 Wang F Y. From AI to SciTS: Team Science and Research Intelligence. IEEE Intelligent Systems, 2011, 27(4).
- 4 王飞跃. 复杂性与情结 vs 智能化与情报. 复杂系统与智能化, 2005, (2).
- 5 Zhang Q, Wang F Y, Zeng D et al. Understanding Crowd-Powered Search Groups: A Social Network Perspective. PLoS ONE, 2012.
- 6 Lai G, Zhang Q, Wen D et al. A Prototype of the Next Generation Journal System for ITS: Academic Social Networking and Media based on Web 3.0. IEEE Transactions on Intelligent Transportation Systems, 2012, 13(2): 1 078-1 087.
- 7 Zhang Q, Feng Z, Li X et al. 25 Years Collaborations at IEEE Intelligent Systems. IEEE Intelligent Systems, 2010, 25(6): 67-75.
- 8 Wang F Y, Zeng D, Hendler J et al. A Study of the Human Flesh Search Engine: Crowd-Powered Expansion of Online Knowledge. IEEE Computer 2010, 43(8): 45-53.
- 9 Wang T, Zhang Q, Liu Z et al. On Social Computing Research Collaboration Patterns: A Social Network Perspective. Frontiers of Computer Science, 2012, 6(1): 122-130.
- 10 Zheng X L, Ke G Y, Zeng D D et al. Next-Generation Team-Science Platform for Scientific Collaboration. IEEE Intelligent Systems, 2011, 26(6): 72-76.
- 11 Li L J, Li X, Cheng C J et al. Research Collaboration and ITS Topic Evolution: 10 Years at T-ITS. IEEE Transactions on Intelligent Transportation Systems, 2010, 11(3):517-523.
- 12 Wang F Y, Lai G, Tang S M. An Application Specific Knowledge Engine for Researches in Intelligent Transportation Systems. Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems, 2004.

Decision Service and Academic Analytics for Development of S&T based on Open Source Intelligence and Big Data

Wang Feiyue

(The State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, CAS Beijing 100190)

Abstract This article addresses concepts, frameworks, methods, architectures, processes and prototype systems for decision service and academic analytics in science and technology based on open source intelligence and big data. The impact and significance of the Cyberspace, open source intelligence, and big data are investigated. The focus of discussion is on the needs for rapid and revolutionary changes in the acquisition, analysis, dissemination, influence, and generation of academic information for the purpose of evaluation and decision support in science and technology development.

Keywords big data, open source intelligence, academic intelligence, research analytics, ACP method

王飞跃 中科院自动化所复杂系统管理与控制国家重点实验室研究员。主要研究领域为复杂系统的智能控制理论与方法。自20世纪80年代,提出并建立了智能系统的协调结构和理论、语言动力学理论、代理控制方法、ACP方法等,发起并开拓了社会计算、平行控制、平行管理等新的交叉学科研究领域。现任IEEE会刊 *IEEE Transactions on Intelligent Transportation Systems*、*IEEE Intelligent Systems* 和《自动化学报》的主编。E-mail: feiyue@gmail.com



中国科学院