

论计算型情报分析*

李广建 江信昱

摘 要 受大数据环境和第四范式的影响,情报分析走向计算的趋势已经越来越明显。在信息设备低廉化、处理对象数字化、计算方法智能化的共同作用下,情报分析逐步走向计算型情报分析。计算型情报分析是情报分析与以计算机技术为核心的信息技术相结合的产物,涵盖目标、功能、实现、分析方式和知识体系五个层面的内容;以计算机及相关信息技术为工具,以机器学习与知识理解等为核心技术,以情报分析方法和数学模型方法为组织分析手段,通过对数据内容及其关系、模式的深度解析、挖掘和发现,帮助分析人员解决情报问题,完成情报任务。计算型情报分析的思维理念包括定量思维、自动化思维、融合思维、容错思维等,在理论、方法和技术、系统、应用实践等四个方面都有诸多研究课题等待我们探索和挖掘。参考文献 51。

关键词 情报分析 大数据 计算型情报分析 计算化 深度分析

分类号 G250.2

On Computational Information Analysis

LI Guangjian & JIANG Xinyu

ABSTRACT

Due to the impact of the big data environment and the fourth paradigm of scientific research, the trend of computability in information analysis has become more obvious. First, as information equipment becomes cheaper, the cost of tools for information analysis decreases and the path for the computability of information analysis widens. Second, most of the objects to be processed are in digital forms which enable a wider span of data resource to be analyzed. This creates more opportunities for computing in information analysis. Third, computing methods have become more intelligent, which brings about the ability to carry out an in-depth analysis and provides a deep understanding of the information. These three factors play a decisive role in making the transition from information analysis to Computational Information Analysis(CIA).

CIA is both a theory and an approach which derives from information analysis and information technology of which computer technology is at the center. It is proposed to address the limitation of human's ability in collecting, processing and analyzing large amount of information. In CIA, computers and related information technology are adopted as basic tools and intelligent technologies like machine learning and knowledge understanding are incorporated as core technology. Information analysis methods and mathematical models

* 本文系国家自然科学基金重点项目“大数据环境下的计算型情报分析方法与技术研究”(编号:14ATQ005)的研究成果之一。(This article is an outcome of the key project “Research on Computational Information Analysis Method and Technology in Big Data Environment”(No. 14ATQ005) supported by National Social Science Foundation of China.)

通信作者:李广建,Email:ligj@pku.edu.cn, ORCID:0000-0002-2897-6246 (Correspondence should be addressed to LI Guangjian, Email: ligj@pku.edu.cn, ORCID: 0000-0002-2897-6246)

are used to organize and analyze information. Through a deeper analysis and mining in the content, relationship and pattern in the data, CIA helps analysts solve analysis problem and tasks and make decisions.

While human is inadequate in gathering, processing and analyzing large amount of data, CIA makes up for that shortcoming with the ability to harness the fast computing power of computers and is also more objective when compared with human. CIA provides a more comprehensive, faster, accurate, objective and deeper analysis result for the target problems thanks to its advantages in harnessing the power of computer technology and quantitative analysis methods. To be more specific, CIA is driven by data mining and machine learning techniques on the analysis environment which is built upon computers. The analysis of CIA is mainly quantitative and places an emphasis on methods that can be processed by machines to undergo quantitative, computational and testable process. Besides, CIA also contributes to the knowledge system as it not only inherits theories and methods from information analysis but also borrows theories and methods from computer science, data science, mathematical statistics and other fields and develops new theories and methods with computational characteristics.

CIA has its own thinking and ideas. Quantitative thinking is to abstract quantitative mathematical forms from objective things and phenomena. Automatic thinking is to transform data and analysis method in a form that can be processed by a computer. Integrative thinking places an importance on the relevance of different data sources. By fusing multiple data sources, it is possible to not only reveal correlations but also complement or cross-validate using different kinds of data. Fault-tolerance thinking means that precision loss at a micro-level and a certain amount of errors and confusions is tolerable.

Finally, this paper raises research topics which deserve more attention under different perspectives, i.e. theory, method and technique, system, and application practice, including but not limited to: constructing a theory of data advanced analytics based on the integration of information science theory and related disciplines, discussing on automatic implementation and improvement of different information analysis methods in CIA, designing an intelligence analysis system within task-data-method which is interrelated, combining with new computing technology or multiple data sources for deep data content mining, etc. 51 refs.

KEY WORDS

Information analysis. Big data. Computational Information Analysis. Computability. Advanced analytics.

0 引言

近年来,大数据所带来的思维理念、方法技术等一系列变化,对社会生活的方方面面都产生了巨大的影响,研究人员也在思考应对的策略。例如,MIT 人类动力学实验室 Pentland 教授^[1]提出“数据驱动型社会”(Data-Driven Socie-

ty),哈佛大学定量社会科学中心 Lazer 教授等人^[2]提出“计算社会科学”(Computational Social Science)这一概念,强调用基于数据的分析方法与技术从海量数据中找出社会发展或科学研究对象的演变规律和潜在关系。在情报研究与分析领域,众多研究人员也在呼唤新的分析模式、方法、技术或服务,以应对大数据环境下情报研究工作面临的挑战。在理论上,相关研究特别

强调量化、可计算或自动化的思维理念,以数据为基础,以计算机技术为支撑,广泛使用定量或计算方法,重视数据化决策或数据驱动分析,分析技术的智能化应用,等等^[3-6]。在实践上,情报研究与分析工作正从注重定性分析向注重定量分析转变^[7-8],强调利用多种数据并注重数据之间的相关性及与分析结果的相互补充^[9];同时,分析过程越来越重视数据分析技术(如数据挖掘、自然语言处理、可视化分析等工具或软件)在数据预处理和数据内容分析方面(如相关分析、共现、网络分析、建模仿真等)的深度整合^[10-12],以减少人工经验判断的主观性问题。换句话说,未来的情报分析会明显呈现出以数据为中心、量化或数据化、分析方法和技术的智能化、多源融合、机器可处理等计算特征。

我们将这种具备上述计算特征的情报分析概括为“计算型情报分析”。本文的目的在于:梳理计算型情报分析的产生背景,界定计算型情报分析的基本概念和研究内容,分析计算型情报分析的思维理念,总结计算型情报分析与传统情报分析之间的关系,以期能对计算型情报分析这一概念有一个清晰的认识。

1 计算型情报分析的产生背景

大数据环境下,情报分析工作历经从经验科学向数据密集型科学过渡的范式转变^[3],情报学学科大量引入数理统计、计算机科学的领域知识和分析技术^[7, 13],基于综述或定性分析的情报成果已经无法满足日渐复杂的决策需求和情报任务,转而需要综合利用数据或信息进行更精细、准确的情报分析并提高情报服务质量^[10, 14]。情报分析的这种变化源自信息设备成本低廉化、处理对象数字化、计算方法智能化三个因素的共同作用,这三个因素对于情报分析向计算型情报分析的发展有决定性的促进作用。

1.1 信息设备成本低廉化

信息设备成本低廉化是指计算机、存储设

备、网络设备等支撑计算型情报分析的物质基础的性能显著提高,而成本急剧下降。例如,英特尔公司在1993年推出奔腾微处理器,集成320万个晶体管,价格超过2 000美元;现在的酷睿i系列处理器,集成超过5亿个晶体管,价格下降到200—300美元。一个集成电路集成越来越多的晶体管,意味着系统的处理性能越好,且CPU价格也越来越便宜。又例如,2002年,256MB、USB2.0接口的U盘价格超过600元人民币,到近年来16GB、USB3.0接口的U盘价格约100—200元人民币,不只数据的传输速度更快,单位价格的数据容量也大幅提高。不断下降的存储和传输成本以及不断增大的数据存储空间,促使数据和信息的处理和存储能力成百万倍地提升。近年来,云计算、分布式运算平台等新技术设备的出现,又进一步改变了数据的存储、访问和计算方式。也就是说,在这些新技术出现之前,多数的数据分散存储在个人计算机或服务器中,而现在可以把数据上传到公有云或私有云中,通过分布式文件系统和计算技术把数据和计算资源分布到大量的计算机中,使数据存储和计算获得较强的可扩展能力,解决了传统单一物理机器或并行计算技术存在的CPU性能不足、内存计算慢、高能耗等问题。

在情报分析过程中,绝大部分计算工作不是简单的统计分析,而是要进行复杂的语义分析,这类分析工作涉及多种数据,计算量大,时间复杂度高,必须要有高速的计算和传输能力作为支撑,才能在有限的时间内完成相应的情报任务,使得情报分析的计算化成为可能。

1.2 处理对象数字化

处理对象数字化是指情报分析的处理对象由以往的纸质存在形式转变成数字存在形式。随着互联网的发展,各种类型信息及其载体的数字化已经成为一种常态,这不仅为情报分析提供了更多的可计算机会,也极大地拓展了传统意义上的情报分析处理对象的范围。例如,科技文献是情报分析的传统处理对象,近年的

数字出版与出版物数字化,已经使得科技文献普遍呈现数字形式。如,DOAJ 网站已有超过 9 000 本开放获取期刊及超过 240 万篇论文,arXiv 预印本公开网站已有超过 120 万篇预印本论文,WoS 中的 Data Citation Index 数据库可以提供近 200 万条论文中使用的研究数据纪录,WorldCat 数据库收录自 1971 年建库以来近两万家图书馆收录的近 19 亿条馆藏纪录,高校或科研单位的机构知识库也提供机构成员的各种科研成果。又例如,情报分析在研究电子政务或政府政策时,会利用公开的政府信息进行数据分析或文件解读。随着政府信息化建设和开放数据平台的推展,各种数字化形式的政府数据不仅可以在线访问及获取,也允许再利用。如美国政府数据开放平台已累积超过 20 万组数据集,北京市和上海市政务数据服务网也提供超过 900 类数据集等,这些信息资源也已经成为现代情报分析的研究对象。

此外,Kleiner 等人^[15]在谈及社会科学的计算化时,将计算社会科学可用的处理对象总结为四类:第一类是管理型数据,是指社会管理机构(例如政府及其他机构)以管理为目的而收集的数据,它们记录了公共领域的日常活动,包括税收、劳工、教育、社会保险、重大事件、健康等,这类数据可以用来分析社会政策制定的得失和效果;第二类是社交媒体数据,是指人们在社交媒体平台如微博、微信上活动所产生的数据,这些数据可用来了解人们的真实行为与想法;第三类是事务数据,是指相关业务系统如客户管理系统、信用卡交易系统等在运行中所产生的数据,这些数据可用来了解系统用户的活动;第四类是文本数据,主要是指数字形式的文章、讲演稿、图书以及各种文本数据库。Kleiner 等人认为,上述四类数据都是计算社会科学研究素材,并且已经有成功的分析案例。笔者认为,Kleiner 等人所得到的结论,同样适用于情报分析。这些数字化的数据资源都是情报研究进行全领域分析或多源知识融合分析的基石,主要特征在于机器可读和可处理,为计算型情报分析提供了计算基础。

1.3 计算方法智能化

计算方法智能化是指现在的软件及算法越来越智能,自然语言理解、知识表示、知识推理、机器学习等技术越来越成熟,软件与算法的功能早已超出以往统计计算的范畴,向更高层次的建模计算方向发展。在情报分析领域,海量数据隐藏了大量有价值的知识或情报,只利用简单的统计方法是无法将它们找出来的,只有通过智能化的计算方法或分析方法(指方法或技术能适应不同数据条件,自动地进行高级、复杂的数据处理与分析工作)^[3,4,10],才能发现更具价值的知识、揭示潜在关系或规律或回答更前沿、更复杂的问题,也才能代替情报分析过程中的人的脑力劳动。近年来,广泛地利用自然语言处理、数据或文本理解、深度学习、语义计算、可视化分析、知识图谱分析等相关方法和技术来完成情报课题任务,已经成为情报分析的一个热门趋势,例如,过去分析学科领域的主题变化多以计量方法进行词频或共现分析,较难识别出文献中其他的隐含主题关系,现在转而利用主题模型(如 LDA、AT、DTM 等模型算法),把每个主题看成词的概率分布,用以识别文献中的潜在主题或是结合其他特征进行主题演化分析^[16-17]。又例如,在生物医学领域的情报分析中,2015 年多伦多大学计算机工程学系组织生物医学相关团队,利用深度学习技术对全基因组数据进行基因与癌症的关联挖掘,其中的一项结果除了找出与结肠癌、胰腺癌和脊髓性肌萎缩症相关的新突变基因外,还发现了 39 个未曾发现的与自闭症关联的基因(以往文献只找出 CTNND2、PTEN 等几个基因与自闭症相关)^[18]。

情报分析从简单统计计算朝向高级建模应用的发展,说明过去强调量的规律和变化的分析已经远远不能满足大数据环境对情报分析提出的新要求,随着数据量的急增,除了数据的属性和复杂性增加之外,也隐藏了更多以往观察中不易发现的点或关联^[3,19]。也就是说,在数据不断增加的情况下,只靠人脑的判断力或简单

统计方法无法完整揭示全部的数据关系并满足决策者对深刻洞察的需求。因此,需要借助机器可执行的计算方法或模型,帮助人们更快、更容易地发现海量数据的所有特征和关系,以支持对处理对象内容的深度分析或深度理解。可见,计算方法的智能化为计算型情报分析提供了手段上的支撑。

2 计算型情报分析的内涵

计算型情报分析并不是一个全新概念。相关研究已注意到,大数据时代和科学研究第四范式的发展,需要现有的情报研究融入新的思维,因而都提出了与表述情报分析向计算化方向演进的类似概念^[10, 20, 21]。但是,学者们各自研究视角不同,在用词和表达方式上存在差异,因而也未就这一问题进行专门的讨论。本文把这些类似的概念统一总结为“计算型情报分析”,并在相关研究基础上,进一步说明计算型情报分析的概念内涵。

本文认为,所谓的“计算型情报分析”,是情报分析与以计算机技术为核心的信息技术相结合的产物,它弥补了人脑在面对庞大信息时搜集、处理和分析上的自然局限,以计算机及相关信息技术为工具,以机器学习与知识理解等智能分析技术为核心技术,以情报分析方法和数学模型方法为组织分析手段,通过对数据内容及其关系、模式的深度解析、挖掘和发现,帮助分析人员解决情报问题、完成情报任务、辅助决策。

以上认识包括以下五层含义。

第一,从目标上看,计算型情报分析是为弥补人脑在面对庞大信息时搜集、处理和分析上的不足而产生的,这里所说的不足包括但不限于以下方面:用户需求表达不明确或者对用户需求理解的不到位^[22];不完全或有限的资料搜集范围;难以在有限的时间里对大量数据和资料进行快速有效的整理、筛选以及阅读理解;人工分析过程中存在主观性干预或非理性判

读^[23-24];对同一现象或结果进行分析,不同背景的专家研究结论差异过大,等等。因此,为了避免人脑的自然局限带来的各种问题,需要做到全面地占有与情报任务有关的信息,充分利用计算机的快速处理和客观分析的计算能力,支持分析人员从不同视角去探索、挖掘或发现数据中的内容、关系、模式或异常,以避免人的主观性。

第二,从功能上看,计算型情报分析在本质上是情报分析,是情报分析的一种特定类型。它和一般情报分析一样,基本流程都是通过对数据或信息的搜集、整序、分析、浓缩、综合、评价等一系列研究活动,从中找出对决策有用的内容、知识或经过检验评估的可靠信息。也就是说,计算型情报分析最终要解决的问题和情报分析是一致的,都是要解决情报问题、完成情报任务,只是计算型情报分析借助计算机技术和定量化分析方法的优势,力图提供比一般情报分析更全面、快速、精准、客观以及有深度的情报分析结果。

第三,从实现上看,计算型情报分析的基础是计算机技术,它以计算机强大的计算能力为依托,运用数据挖掘与机器学习技术实现分析。因此,计算型情报分析的实现必须要广泛地应用网络传输、网络数据库、文本处理、数据融合、数据挖掘、机器学习、语义处理、可视化等相关技术,支撑分析过程中的数据处理、计算、建模或表示,同时也会借鉴或移植计算机科学、数据科学、数学或统计学等相关理论知识,用以指导分析过程的操作与结果诠释。也就是说,计算型情报分析的核心分析环节需要基于机器或计算机技术的分析环境,通过机器或计算机技术代替或补充人的分析或计算,其目的是为了

提高分析的科学性,保持结果的可靠性,使结果逼近真实情况,对情报问题或情报任务提供系统性、多角度和可信的分析结果或解决方案。

第四,从分析方式上看,计算型情报分析以定量分析为主,包括情报分析方法中的计量方

法(文献计量、信息计量、科学计量、网络计量、知识计量等)、引文分析、网络分析等,或是运用数学模型方法(指采用形式化的数学语言,概括地表达研究对象的主要特征、关系及结构)和其他数理统计方法,针对不同情报问题或情报任务组织或集成各种专门方法或一般方法,对数据内容及其关系、模式进行深度解析、挖掘和发现。因此,计算型情报分析除了强调用数据说话、让研究始于数据等思维之外,在具体实践中,它基于机器或计算机技术的分析环境,重视运用机器可读取、处理、理解的分析方法,来实现可定量、可计算、可检验的过程,其目的是减少人对分析结果生成的直接干预并确保分析结果可再现,为情报问题或情报任务提供研究对象的整体数量分布、规律或关系变化,模拟仿真的最优结果或发现新的知识。

第五,从知识体系上看,计算型情报分析不仅仅是计算机技术及算法在情报分析中的应用,它应该有自己的理论基础。即在情报学基础上借鉴或移植计算机科学、数据科学、数学统计等相关理论知识来定义或论证计算型情报分析的若干基本问题(如概念与内涵、范畴与边界、性质与作用、分析方法、应用模式等);同时,它也应该有自己的分析思维与方法技术,即以数据为基础,注重定量计算方法,在基于机器或计算机技术的分析环境上执行机器可处理的分析方法。这些都体现出计算型情报分析与以往情报分析不同的理念及模式,它帮助分析人员以“数据驱动”取代传统“假设驱动”的研究方式,借助计算机技术实现分析过程的自动可控,并减少人为干预,借助定量分析方法,结合数据挖掘和机器学习技术,揭示数据内容中更多的变化规律、结构或关系特征,或是对不同的模型、关联与模式提供相应的仿真模拟与检验。所以,计算型情报分析不仅仅是一种实践,也是一种理论与方法,它既继承了情报分析的理论与方法,也发展出带有计算特征的情报分析理论与方法。

3 计算型情报分析的思维理念

传统的情报分析更多是建立在分析人员的“智慧”之上,是业务导向的分析研究工作,要求情报分析人员具备深厚的业务知识以及构建解决问题方法及逻辑框架的能力^[3, 14, 25]。业务导向的分析过程侧重于定义业务问题,通过分析人员自身深厚的业务背景及业务知识积累,对相关信息进行采集、选择、评价、分析和综合等,形成解决问题的方案或预定问题的答案,并在必要的情况下,用数据来证实自己的分析结论。

计算型情报分析则是一种以数据为导向的分析工作。计算型情报分析当然需要分析人员的智慧支持,但它更强调对数据的获取、处理和利用,注重对分析对象的量化,关注数据之间关系的发现,强调在各种类型数据中持续探查,深入挖掘,从而得到相关结论。因此,计算型情报分析有自己的分析思维 and 理念。

3.1 定量思维

定量思维指一切事物及事物运动的状态皆可用定量的方式进行描述^[19],也是把被量化的事物和数字、数据建立映射(Mapping)的过程^[26]。也就是说,把客观存在的事物或现象抽象概括为数量、模型、规律或公式、指标或变量等可量化的数学形式,例如,学术影响力可概括为论文发表数、被引量、H指数、期刊影响力模型等内容;测量学科知识结构变化可通过三大定律、引文指标、老化规律等揭示学科变化程度;信用评分涉及了人的住房、收入、文化程度、年龄等指标。当我们把事物及其运动状态抽象成可量化的数学形式时,就可以进行数字统计、指标组合、模式发现、仿真模拟等计算过程,或者可以把这些量化的数学形式与计算机上的表达和执行联系起来^[27],实现后续分析的可计算过程,进而提供客观结果来精准解释事物或现象变化的发展和关系,也有助于对未知事物或现象给出一个清楚可行的定义范围。

3.2 自动化思维

自动化思维指通过数字化技术将数据、原始材料(如文本、视频、图片等)、分析方法或已量化表示的模型、指标、变量或量表转变成机器或计算机可理解、读取、存储、处理的形式^[10, 28]。计算型情报分析强调人尽可能少地干预分析的过程,且人的主观意识不直接参与最终结果的生成,是基于对数据和任务复杂性的考虑,并追求自动化本身的计算效率。例如,情报分析中的科学计量、引文分析和网络分析,在一定程度上都可以通过科学知识图谱工具来实现,用户只涉及数据导入、方法选择和参数调整,这就大大减少了人的劳动并提高了分析的效率。又例如,日本前七次的技术预见主要使用了德尔菲方法,每次可获得约2 000—4 000份问卷(有两轮调查),其数据处理和分析是重复性的劳动过程,且无法实时反馈结果^[29];到了第十次技术预见,除了增加调查专家数量,他们还在德尔菲方法基础上构建了一个在线调查系统,借助数据挖掘和可视化技术完成在线分析,既可短期收集到大量数据,也可实时更新和反馈结果^[30]。

3.3 融合思维

融合思维指通过多源数据融合揭示同一个事实或规律的不同侧面。大数据环境下,同一个事实或规律可以同时隐藏在不同的数据形式中,也可能是每一种数据形式分别支持了同一个事实或规律的某一个或几个侧面^[10, 31],这就要求情报分析要充分利用各种信息源并对它们做融合分析。例如,对学科领域发展研究来说,需要综合利用期刊论文、学位论文、项目、专利、会议等信息来描述和预测某项研究的整体发展;对于竞争对手分析来说,除了关注对手网站或产品信息,还要掌控相关领导讲话、行业招聘信息、相关微博、行业数据等各类信息^[9]。又例如,在交通大数据中经常会有出行者分析的应用场景,为了揭示不同出行者的行为特征,需要收集GPS数据、公交或地铁刷卡数据、手机移动

数据、出行者活动的娱乐场所数据、道路数据等多源数据,利用数据挖掘分析出行者的出行链并进行活动诱导,进而降低路网交通压力^[32]。也就是说,融合思维强调多源数据隐含不同的相关性,通过协同利用多种数据,使其相互补充或印证,用以获得对同一事物全面的客观认识,这样才能提高数据分析的科学性和准确性。

3.4 容错思维

容错思维指绝对的精准不再是追求的主要目标,允许适当忽略微观层面上的精确度,容许一定程度的错误与混杂^[19, 26]。也就是说,在大数据环境下,数据分析不要求绝对精准,而是交换更多数据,通过多次检验,利用交叉印证来“逐步减少”问题的不确定性,使概率估计变得准确、结果逼近真实,而不是为了“完全消除”不确定性而舍弃相关数据或分析结果。对传统的情报分析来说,多数情况下并不需要使用大量数据,而是要获取与任务相关的关键数据,因此在适当的数据规模下,会要求数据质量与真实性^[11]。对计算型情报分析来说,随着采集、存储和分析数据的成本降低(包括数据处理的物理限制和人脑分析的自然限制逐渐消失),分析人员可以不用再为了采集20%的核心数据而费时费力地“筛选”,同时还要冒舍弃的80%数据中包含有重要价值内容的风险。例如,过去网络舆情研究通常是随机选取某个时间段的社交数据,且分析结果距离数据产生已有一段时间。在大数据环境中,则可以使用流式计算建构可实时处理的网络舆情分析模型,其流处理系统可以稳定采集各种渠道的实时数据并主动分析,且系统的低延迟计算能力也有助于快速进行数据分析并及时反馈结果^[33]。

4 计算型情报分析中值得关注的研究课题

计算型情报分析的研究刚刚起步,我们认为,在理论、方法及技术、系统以及应用等方面,

都有一些值得关注的研究课题,说明如下。

4.1 理论研究

面对大数据和第四范式对情报分析的影响,现有研究不只是梳理总结大数据环境下情报分析的特征和变化^[3, 5, 6, 10, 21],也有从学科角度来思考情报学科、计算机科学、数据科学等相关学科的交叉或共同发展^[34-36]。其中,深化原有理论基础是这些文献的共同看法,主要包括现有情报分析所使用的理论(如情报交流理论、计量学的三大定律、相关性理论等)需要结合计算机、网络、数据科学等学科的相关理论,作为未来情报分析走向计算、拓展应用研究的重要支撑。例如,王芳等人^[37]分析了十年的情报研究文献中的理论变化,发现我国情报研究除了使用传统理论,也采用了向量空间模型、粗糙集、复杂网络、模糊理论等作为情报分析的知识基础和论据。笔者认为,这些理论的引入有助于情报学科在数据密集型科研下的发展,也有利于情报分析在计算方法和挖掘技术方面的提升。

随着第四范式对情报分析的研究路径、分析视角、分析方法、分析对象等影响的加深,传统情报分析思想和理论内容可能不适用于大数据环境。想要解决这个问题,我们需要进一步研究:①在概念上,剖析情报分析走向计算的特征与作用、大数据分析情报分析的关联和边界差异等;②在知识体系上,建构一套基于情报学理论且融合相关学科的数据深度分析理论,研究数据驱动分析的机制或模型等;③在工作模式上,探讨现有情报分析模式和决策模式在大数据环境下的转变,培养分析人员的数据思维,挖掘用户对数据决策的任务需求等基本内容,进而解释情报分析计算化的现象,指导计算型情报分析的实践。

4.2 方法和技术研究

理论研究是认识计算型情报分析的基础,具体实践则涉及方法和技术的使用。对情报分

析来说,情报分析方法是解决情报问题或情报任务的手段总和,而情报分析方法体系反映了情报研究的知识体系^[25]。由于学者对情报方法的认识差异,构建的方法体系及其方法种类也不尽相同^[38-40],如三分法或四分法、面向分析流程、面向分析对象等方法体系,这说明情报分析方法体系尚未达成共识,且现有体系都存在方法的颗粒度或逻辑层次不一致的问题。

由于现有方法体系的结构表示和方法种类并没有揭示到可计算的指标层次,因此较难适用于计算型情报分析。要解决这个问题,我们需要进一步研究:①在概念上,厘清情报分析方法的本质和特征,辨析方法、技术和工具三者的关系,探讨情报分析方法和计算型情报分析的联系与衔接等;②在体系构建上,研究各种方法间的逻辑关系及相互作用机理、应用原理、适用范围、使用流程、前置条件和后续要求等,研究方法体系的容纳性和结构表示等;③在方法选用上,关注机器可处理的计算方法或技术、情报分析方法的自动化实现途径和适应性改造,调研支持数据深度分析的技术方法或算法,分析技术或算法的集成表示、面向情报任务或情报分析方法的算法调度、不同分析技术或算法的衔接或交叉应用等基本内容,进而指导分析人员解决情报任务。

4.3 系统研究

在计算型情报分析过程中,以机器或计算机代理人进行分析和计算、使用机器可处理的情报分析方法等是重要特征。因此,我们亟需数据分析平台,以便顺利完成上述工作,进而提高分析效率。现有研究大都是运用现成数据分析工具中的分析功能进行数据处理和计算^[41-45],例如,用知识图谱工具来实现文献计量和引用分析,用网络分析工具来实现基于路径、关联或网络的相关分析等。

但是,这些现成工具都只能解决单一情报任务,而随着情报任务的日渐复杂,就需要综合利用多种数据、方法、工具解决任务中的不同情

报问题,且要在有限的时间内及时反馈分析结果。其中,技术工具的分散、多源数据的融合处理、不同方法的衔接等问题,都会大大影响情报分析结果的全面性和时效性。要解决这个问题,我们需要进一步研究:①在系统需求上,调研支持大数据或多源数据存储、处理和计算技术在分析系统中实现的解决方案,设计情报分析流自动化实现机制,研究人和机器如何协作完成情报任务等;②在功能设计上,调研现有情报分析系统已实现的方法功能与指标参数,分析“任务—数据—方法”关联模式作为系统建置的指导思想,规划基于任务问题分解的可配置分析流程,完善系统基础资源(如情报任务库、情报方法库和内置资源库)等基本内容,进而支持数据分析的全过程(从数据导入到结果输出),并有效支持用户对不同情报课题的分析需求。

4.4 应用实践研究

计算型情报分析和情报分析的本质是一致的,所解决的情报任务也是相同的。其中,除了可以应用在业务驱动的情报分析之外,计算型情报分析更强调数据驱动的情报分析,目的是对数据内容进行深层的挖掘,把情报分析从定量提升到计算层次^[46],向用户提供基于知识的深刻洞察和分析服务。这方面,已经有了一些很好的尝试,例如,融合特定技术利用多种数据进行技术洞察和评估发展潜力^[47],利用社交数据进行用户情感分析或舆情跟踪^[48-50]等。又例如,商务管理、生物医学、政府治理或军事情报等领域也有基于大数据的情报分析应用,如竞争对手的实时动态跟踪分析、客户的精准营销策略分析、老药新用组合分析、公共舆情分析或

城市问题管理等^[51]。

由此可见,各领域都已经开始注重计算型情报分析的应用了,但这种应用在总体上看,多是利用数据来回答案实性的问题,即回答过去、现在、未来发生什么的问题,对于更复杂的技术及其应用规律还缺乏研究,为此,我们需要进一步研究:①在技术应用上,如何运用深度学习、情景计算、语义处理等新型计算技术对数据内容进行更细粒度的特征或是多维度的属性提取和处理,并在全面揭示事物的所有关系的基础上进行知识识别或知识融合等;②在结果适用性上,挖掘多种数据或情报产品中不同用户对情报课题需求的共性表现或差异特征,这样既有利于控制情报研究的质量,也可以增强情报分析结果的针对性和重用性;③在应用效率上,如何利用实时数据和近实时(Near Real-time)技术进行建模,通过多次数据计算改进模型仿真的精确性,进而降低因为历史数据所产生的分析滞后问题。

5 结语

本文概略地探讨了计算型情报分析的提出背景、内涵、思维理念、值得关注的课题等几个方面,从中可以看出,计算型情报分析和情报分析的本质和功能是一致的,计算型情报分析应该是情报分析的“深度分析”表现形式,是情报分析为适应当今社会环境和社会需求的必然趋势。因此,探索及深入研究计算型情报分析的理论、方法、工具和应用,为改善和提高情报分析效率与效果提供理论依据和解决方案,将是促进情报学学科发展和情报行业地位提升的关键。

参考文献

- [1] Pentland A. The data-driven society[J]. Scientific American, 2013, 309(4): 78-83.
- [2] Lazer D, Pentland A, Adamic L, et al. Computational Social Science[J]. Science, 2009, 323(5915): 721-723.
- [3] 徐敏,李广建. 第四范式视角下情报研究的展望[J]. 情报理论与实践, 2017, 40(2): 7-11. (Xu Min, Li

- Guangjian. Prospects of information research from the fourth paradigm [J]. Information Studies: Theory & Application, 2017, 40(2): 7-11.)
- [4] 张志强,冷伏海,刘清,等. 知识分析及其应用发展趋势研究[J]. 情报科学, 2010, 28(7): 1100-1107. (Zhang Zhiqiang, Leng Fuhai, Liu Qing, et al. Study on the development of knowledge analysis and its application [J]. Information Science, 2010, 28(7): 1100-1107.)
- [5] 吴晨生,李辉,付宏,等. 情报服务迈向3.0时代[J]. 情报理论与实践, 2015, 38(9): 1-7. (Wu Chensheng, Li Hui, Fu Hong, et al. Intelligence service towards the era of 3.0 [J]. Information Studies: Theory & Application, 2015, 38(9): 1-7.)
- [6] 官思发,朝乐门. 大数据时代信息分析的关键问题、挑战与对策[J]. 图书情报工作, 2015, 59(3): 12-18, 34. (Guan Sifa, Chao Lemen. Research on key issues and challenges of information analysis and its countermeasures in big data era [J]. Library and Information Service, 2015, 59(3): 12-18, 34.)
- [7] 刘红煦,曲建升. 情报学定量研究现状与趋势分析[J]. 情报理论与实践, 2015, 38(12): 10-14, 36. (Liu Hongxu, Qu Jiansheng. Analysis of the status and trends of quantitative research on Information Science [J]. Information Studies: Theory & Application, 2015, 38(12): 10-14, 36.)
- [8] 孙鸿飞,侯伟,周兰萍,等. 近五年我国情报学研究方法应用的统计分析[J]. 情报科学, 2014, 32(4): 77-84. (Sun Hongfei, Hou Wei, Zhou Lanping, et al. Statistical analysis of application of research methods of Information Science in our country in recent five years [J]. Information Science, 2014, 32(4): 77-84.)
- [9] 化柏林,李广建. 大数据环境下的多源融合型竞争情报研究[J]. 情报理论与实践, 2015, 38(4): 1-5. (Hua Bolin, Li Guangjian. Research on multi-source fusion competitive intelligence in the big data environment [J]. Information Studies: Theory & Application, 2015, 38(4): 1-5.)
- [10] 李广建,杨林. 大数据视角下的情报研究与情报研究技术[J]. 图书与情报, 2012(6): 1-8. (Li Guangjian, Yang Lin. Intelligence analysis and intelligence technology in view of big data [J]. Library and Information, 2012(6): 1-8.)
- [11] 李广建,化柏林. 大数据分析 with 情报分析关系辨析[J]. 中国图书馆学报, 2014(5): 14-22. (Li Guangjian, Hua Bolin. Relationship between big data analysis and intelligence analysis [J]. Journal of Library Science in China, 2014(5): 14-22.)
- [12] 赵纪东,张志强. 从主流应用工具看科技情报的定量化研究[J]. 情报杂志, 2011, 30(10): 4-8. (Zhao Jidong, Zhang Zhiqiang. Analysis of scitech information's quantitative studies from popular application tools [J]. Journal of Intelligence, 2011, 30(10): 4-8.)
- [13] 肖勇. 论情报学的数学计量化研究[J]. 图书情报工作, 2004, 48(6): 30-40. (Xiao Yong. Research on the mathematicization and metricization of Information Science [J]. Library and Information Service, 2004, 48(6): 30-40.)
- [14] 徐峰,张旭. 面向决策的情报研究与服务探析[J]. 情报学报, 2012, 31(11): 1124-1130. (Xu Feng, Zhang Xu. Decision oriented intelligence studies, service [J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(11): 1124-1130.)
- [15] Kleiner B, Stam A, Pekari N. Big data for the social sciences; FORS working paper series[EB/OL]. [2016-09-10]. http://forscenter.ch/wp-content/uploads/2014/06/FORS_WPS_2015-02_Kleiner.pdf.
- [16] 张晗,徐硕,乔晓东. 融合科技文献内外部特征的主题模型发展综述[J]. 情报学报, 2014, 33(10): 1108-

1120. (Zhang Han, Xu Shuo, Qiao Xiaodong. Review on topic models integrating intra- and extra-features of scientific and technical literature [J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(10): 1108-1120.)
- [17] 王萍. 基于概率主题模型的文献知识挖掘[J]. 情报学报, 2011, 30(6): 583-590. (Wang Ping. Literature knowledge mining based on probabilistic topic model [J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(6): 583-590.)
- [18] Xiong H Y, Alipanahi B, Lee L J, et al. The human splicing code reveals new insights into the genetic determinants of disease[J]. Science, 2015, 347(6218): 1254806.
- [19] 维克托·迈尔-舍恩伯格, 肯尼斯·库克耶. 大数据时代: 生活、工作与思维的大变革[M]. 盛杨燕, 周涛, 译. 杭州: 浙江人民出版社, 2013: 45-47, 97-102. (Mayer-Schonberger V, Cukier K. Big data: a revolution that will transform how we live, work and think [M]. Sheng Yangyan, Zhou Tao, trans. Hangzhou: Zhejiang People's Publishing House, 2013: 45-47, 97-102.)
- [20] 贺德方. 工程化思维下的科技情报研究范式——情报工程学探析[J]. 情报学报, 2014, 33(12): 1-13. (He Defang. The research paradigm of science and technology intelligence under the engineering thinking style-tentative discussion on intelligence engineering [J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(12): 1-13.)
- [21] 张志强. 论科技情报研究新范式[J]. 情报学报, 2012, 31(8): 788-797. (Zhang Zhiqiang. New paradigm for S & T intelligence studies [J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(8): 788-797.)
- [22] 细野公男, 邹永利. 情报需求及其表述中的知识结构与情报系统设计[J]. 图书情报工作, 1999, 1(1): 1-7. (Kimio Hosono, Zou Yongli. Knowledge structure for information needs and their representation: implication for information system design [J]. Library and Information Service, 1999, 1(1): 1-7.)
- [23] 王大龙. 基于情报决策者非理性决策的情报失察研究[J]. 现代情报, 2011, 31(8): 18-20. (Wang Dalong. Research in intelligence failure due to intelligence decision maker's irrational decision [J]. Modern Information, 2011, 31(8): 18-20.)
- [24] 沈固朝. 情报失察——西方情报研究的重要课题及其对我们的启示[J]. 图书情报工作, 2009, 53(2): 34-37. (Shen Guchao. Intelligence failure—an important topic in western intelligence studies and its inspirations to its practices in China [J]. Library and Information Service, 2009, 53(2): 34-37.)
- [25] 包昌火. 情报研究方法论[M]. 北京: 科学技术文献出版社, 1990: 12-22. (Bao Changhuo. Methodology of intelligence study [M]. Beijing: Science and Technology Literature Press, 1990: 12-22.)
- [26] Hubbard D W. How to measure anything: finding the value of intangibles in business[M]. New Jersey: Wiley, 2014: 48-50, 148-152.
- [27] Wing J M. Computational thinking[J]. Communications of the ACM, 2006, 22(9): 45-49.
- [28] Wing J M. Computational thinking and thinking about computing[J]. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2008, 366(1881): 3717-3725.
- [29] 范晓婷, 李国秋. 日本技术预见发展阶段及其未来趋势分析[J]. 竞争情报, 2016, 12(3): 37-42. (Fan Xiaoting, Li Guoqi. Analysis on the development stage and future trend of technology foresight in Japan [J]. Competitive Intelligence, 2016, 12(3): 37-42.)

- [30] 张峰, 邝岩. 日本第十次国家技术预见的实施和启示[J]. 情报杂志, 2016, 35(12): 12-15. (Zhang Feng, Kuang Yan. The implementation and enlightenment of Japan's 10th science and technology foresight [J]. Journal of Intelligence, 2016, 35(12): 12-15.)
- [31] 化柏林, 李广建. 大数据环境下多源信息融合的理论与应用探讨[J]. 图书情报工作, 2015(16): 1-6. (Hua Bolin, Li Guangjian. Discussion on theory and application of multi-source information fusion in big data environment [J]. Library and Information Service, 2015(16): 1-6.)
- [32] 吕明, 程歆. 多源数据融合的出行者画像研究[J]. 科技创新导报, 2017, 14(10): 143-144. (Lü Ming, Cheng Xin. Research on traveler profiling based on multi-source data fusion [J]. Science and Technology Innovation Herald, 2017, 14(10): 143-144.)
- [33] 高欢. 基于流式计算的网络舆情分析模型研究[J]. 情报学报, 2016, 35(7): 723-729. (Gao Huan. Research on model of network public opinion analysis based on stream computing [J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(7): 723-729.)
- [34] 叶鹰, 马费成. 数据科学兴起及其与信息科学的关联[J]. 情报学报, 2015, 34(6): 575-580. (Ye Fred Y, Ma Feicheng. Data science: its emergence and linking with Information Science [J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(6): 575-580.)
- [35] 曾建勋, 魏来. 大数据时代的情报学变革[J]. 情报学报, 2015, 34(1): 37-44. (Zeng Jianxun, Wei Lai. The changes of Information Science in big data era [J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(1): 37-44.)
- [36] 周晓英. 数据密集型科学研究范式的兴起与情报学的应对[J]. 情报资料工作, 2012, 33(2): 5-11. (Zhou Xiaoying. The rise of a data-intensive scientific research paradigm and the response of Information Science to it [J]. Information and Documentation Services, 2012, 33(2): 5-11.)
- [37] 王芳, 史海燕, 纪雪梅. 我国情报学研究中理论的应用: 基于《情报学报》的内容分析[J]. 情报学报, 2015, 34(6): 581-591. (Wang Fang, Shi Haiyan, Ji Xuemei. The use of theory in Chinese Information Science research based on the content analysis of the *Journal of the China Society for Scientific and Technical Information* [J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(6): 581-591.)
- [38] 化柏林. 网络海量信息环境下的情报方法体系研究[J]. 情报理论与实践, 2012, 35(11): 1-5. (Hua Bolin. Research on system of intelligence method under the network massive information environment [J]. Information Studies: Theory & Application, 2012, 35(11): 1-5.)
- [39] 化柏林, 李广建. 面向情报流程的情报方法体系构建[J]. 情报学报, 2015, 35(2): 177-188. (Hua Bolin, Li Guangjian. Constructing methodology of intelligence analysis based on intelligence process [J]. Journal of the China Society for Scientific and Technical Information, 2015, 35(2): 177-188.)
- [40] 余波, 温亮明, 张妍妍. 大数据环境下情报研究方法论体系研究[J]. 情报科学, 2016, 34(9): 7-12. (Yu Bo, Wen Liangming, Zhang Yanyan. Research of system of information research method in big data [J]. Information Science, 2016, 34(9): 7-12.)
- [41] 张雯雯, 许鑫. 文本挖掘工具述评[J]. 图书情报工作, 2012, 56(8): 26-31. (Zhang Wenwen, Xu Xin. Review of text mining tools [J]. Library and Information Service, 2012, 56(8): 26-31.)
- [42] Cobo M J, Lopez-Herrera A G, Herrera-Viedma E, et al. Science mapping software tools: review, analysis, and cooperative study among tools[J]. Journal of the American Society for Information Science and Technology, 2011,

- 62(7): 1382-1402.
- [43] 侯丽,李亚子. 情报分析方法与情报分析软件的适用性探讨[J]. 情报科学,2014, 32(4): 114-121. (Hou Li, LI Yazi. Discussion of applicability on information analysis method and software of information analysis [J]. Information Science, 2014, 32(4): 114-121.)
- [44] 刘玉琴,彭茂祥. 国内外学术关系分析方法与工具研究综述[J]. 情报科学,2013, 31(11): 137-142. (Liu Yuqin, Peng Maoxiang. A summary of studies on analysis methods and tools about academic relationship at home and abroad [J]. Information Science, 2013, 31(11): 137-142.)
- [45] 杨振瑜,王效岳,白如江. 国外主要可视化数据挖掘开源软件的比较分析研究[J]. 图书馆理论与实践, 2013(5): 89-93. (Yang Zhenyu, Wang Xiaoyue, Bai Rujiang. Comparative analysis of foreign open source software for visual data mining [J]. Library Theory and Practice, 2013(5): 89-93.)
- [46] 李广建,江信昱. 情报分析计算化:背景、作用及关键问题[J]. 图书情报工作,2017, 61(16): 24-30. (Li Guangjian, Jiang Xinyu. The computation of intelligence analysis: background, functions and key issues [J]. Library and Information Service, 2017, 61(16): 24-30.)
- [47] 乔晓东,朱礼军,李颖,等. 大数据时代的技术情报工程[J]. 情报学报,2014, 33(12): 38-53. (Qiao Xiaodong, Zhu Lijun, Li Ying, et al. Introduction of technology intelligence engineering in big data era [J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(12): 38-53.)
- [48] Golder S, Macy M. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures[J]. Science, 2011, 333(6051): 1878-1881.
- [49] 黄晓斌,赵超. 文本挖掘在网络舆情信息分析中的应用[J]. 情报科学, 2009, 27(1): 94-99. (Huang Xiaobin, Zhao Chao. Application of text mining technology in analysis of net-mediated public sentiment [J]. Information Science, 2009, 27(1): 94-99.)
- [50] 陈忆金,曹树金,陈桂鸿. 网络舆情意见挖掘:用户评论情感倾向分析研究[J]. 图书情报知识,2013(6): 90-96. (Chen Yijin, Cao Shujin, Chen Guihong. Online public opinion mining: user's sentiment analysis [J]. Document, Informaiton & Knowledge, 2013(6): 90-96.)
- [51] 李广建,江信昱. 不同领域的情报分析及其在大数据环境下的发展[J]. 图书与情报,2014(5): 7-12. (Li Guangjian, Jiang Xinyu. Intelligence analysis in different domains and its development under the environment of big data [J]. Library and Information, 2014(5): 7-12.)

李广建 北京大学信息管理系教授,博士生导师。北京 100871。

江信昱 北京大学信息管理系博士研究生。北京 100871。

(收稿日期:2017-10-11)