DOI:10.11992/tis.201610025
网络出版地址:

大数据情报分析发展机遇及其挑战

黄河燕[1,2],曹朝[1,2],冯冲[1,2]
(1.北京理工大学 计算机学院,北京 100081; 2. 北京市海量语言信息处理与云计算应用工程研究中心,北京 100081)

摘 要:大数据时代,情报信息的分析处理面临着前所未有的机遇和挑战。本文从情报学发展范式的角度阐述了情报分析的现状;以事实数据、工具方法和专家智慧相融合的情报处理理念为指导,剖析了大数据情报分析在大数据融合、大数据处理技术与工具、信息深度挖掘方面的应用需求和面临的挑战;最后以大数据情报分析过程中的数据采集、预处理、分析和应用为主线展望了大数据情报分析的应用发展机遇和技术趋势。

关键词:大数据;情报分析;情报学;机遇与挑战;云计算
中图分类号: TP18 **文献标志码**:A **文章编号**:1673-4785(2016)06-0719-09

中文引用格式:黄河燕,曹朝,冯冲. 大数据情报分析发展机遇及其挑战[J]. 智能系统学报, 2016, 11(6): 719-727.
英文引用格式:HUANG Heyan, CAO Zhao, FENG Chong. Opportunities and challenges of big data intelligence analysis[J]. CAAI Transactions on Intelligent Systems, 2016, 11(6): 719-727.

Opportunities and challenges of big data intelligence analysis

HUANG Heyan[1,2], CAO Zhao[1,2], FENG Chong[1,2]
(1. School of Computer Science, Beijing Institute of Technology, Beijing 100081, China; 2. Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, Beijing 100081, China)

Abstract: In the era of big data, information and intelligence analysis is facing unprecedented opportunities and challenges. This paper describes the status of intelligence analysis from the perspective of the information science development paradigm. With the guidance of information processing concepts, which is an integration of factual data, tools, methods and expert wisdom, the application requirements and challenges of big data intelligence analysis were analyzed in terms of big data integration, big data processing technology, tools and deep information mining. Finally, because the big data intelligence analysis process consists of data collection, pre-processing, analysis and application as the main components, the application development opportunities and technical trends of big data intelligence analysis were forecasted.

Keywords: big data; intelligence analysis; information sciences; opportunities and challenges; cloud computing

大数据时代,随着数据的爆炸式增长,海洋一般浩瀚的数据已成为一种类似于矿藏的战略资源。Gartner 公司的报告提出大数据是大容量、高速和多样化的信息资产,它们需要新的处理方式,以提高决策能力、洞察力并进行流程优化。另外,如何从这些海洋一般浩瀚的数据中挖掘出有价值的信息、提炼出知识规律、提供正确的决策如同矿产资源探测、采矿、冶炼一般需要数据科学家和领域专业人员的共同努力。情报工作是对情报进行科学地、有组织地搜集、整理、加工、存储、检索和研究,及时而准确地进行传播交流,达到充分有效提供使用的目的的一种业务活动。美国政府已经将大数据技术应用到实际运作中,比如:美国中央情报局(CIA)首席技术官透露美国已经将大数据技术应用于恐怖分子追踪和社会情绪的监控;在“阿拉伯之春”过程中,通过大

收稿日期:2016-10-24.
基金项目:国家重点研发计划项目(2016YFB1000902).
通信作者:黄河燕.E-mail:hhy63@ bit.edu.cn.

数据分析可以了解多少人和哪些人正在从温和立场变得更为激进,并预测出谁可能会采取对某些人有害的行动。由此可以看出,大数据的价值链与情报工作的价值链完全一致[1]。

大数据时代的来临,给各个学科带来了前所未有的机遇和挑战,尤其是以数据采集和信息处理与分析为基础的情报分析,其发展也随着大数据技术的发展面临着前所未有的机遇和挑战。本文结合情报分析的发展现状以及当前大数据情报分析的应用需求,阐述大数据技术的发展给情报分析带来的重大影响和变革,并且从大数据情报分析过程中涉及到的数据采集、处理、分析和应用各个阶段对大数据情报分析的未来应用发展和技术发展趋势进行了详细的分析和展望。

1 情报分析的发展范式及其现状

情报分析也称为信息分析或者情报研究,是指根据社会用户的特定需求,以现代信息技术和软科学研究方法为主要手段,以社会信息的采集、选择、评价、分析和综合等系列加工为基本过程,形成新的、增值的情报产品,为不同层次科学决策服务的社会化智能活动[2]。情报分析是社会重大决策规划和实施中的"耳目和尖兵",它研究的重点也始终关注于数据的采集、处理、分析及深层次挖掘,探索从复杂的数据中找到知识之间有效关联及知识发现的最佳方法。

从情报学发展范式来看,情报学发展经历了 4 个阶段:

1)基于信息的事实型情报学发展范式(20 世纪 40~60 年代),这个阶段提出了情报学的研究内容和研究方法,形成了最初的情报学思想,也是标志情报学的产生和确立的重要时期;

2)基于信息管理的综述型情报学发展范式(20 世纪 70~90 年代),这一时期情报学研究对特定的学科选题进行了分析,具备了明显的管理学特征;

3)基于智能的智慧型情报学发展范式(1995~2010 年),情报学的研究表现出了智能深度挖掘、数据信息关联的特征。但是这一时期的情报学研究也有一些限制,比如主要聚焦于单一领域,考虑的数据源和数据类型主要局限于结构化数据,智能情报分析对分析人员要求过高(模型选择、各种繁杂的参数),需要大量的人工辅助或者人工处理,智能化程度有待进一步提升;

4)基于大数据的情报学发展范式(21 世纪初至今),在大数据技术蓬勃发展的背景下,本阶段情报学的研究范畴明显符合了数据量巨大、信息源多、数据类型复杂等大数据的典型特征。IBM 公司定义的大数据的 4V 特性:大数据量(Volume)、高数据速率(Velocity)、多样性(Variety)和真实性(Veracity),在大数据情报学发展范式中有明显的体现[3]。

情报分析发展到基于大数据的阶段,大数据技术的应用对情报学的理念、研究内容、主要技术方法等方面产生了深刻而重要的影响,一方面各国的政府机构逐步重视大数据在情报分析方面的应用;另一方面也产生了专门进行情报大数据分析的商业化公司。以美国的 Palantir 公司为例,Palantir 公司主营情报分析业务,也是将大数据技术应用于情报分析的典型代表,它的主要客户包括:中央情报局(CIA)、国土安全部(DHS)、国家安全局(NSA)、联邦调查局(FBI)、疾病防控中心(CDC)等美国政府机构。有消息称:"本拉登的行踪线索是通过情报软件 Palantir 确定的"。

目前大数据情报分析仍然处于初步且快速发展的阶段。以 Palantir 公司为例,随着应用于情报分析的大数据技术不断成熟,Palantir 与客户的合作模式也在发生转变。在 2010 年之前,外派工程师需要花费数十天时间对客户的大规模数据进行人工预处理,然后通过该公司的产品将凌乱的数据转换成直观的图表,借助先进的软件和算法进行分析。而在 2010 年以后该公司逐步形成软件对大数据集成、安全等进行统一管理和进一步的分析。由 Palantir 公司成功的经验可以看出,大数据情报分析首先需要有高质量的数据基础,因此数据的清理、预处理也是大数据情报分析重要而且必须的一个环节。

2 大数据情报分析的应用需求和面临的挑战

情报学研究的重点始终关注数据的处理、分析及深层次挖掘,探索从复杂的数据中找到知识之间有效关联及知识发现的最佳方法,大数据情报分析作为其中的一种发展范式也不能例外。情报分析中传统的基于"事实数据+工具方法+专家智慧"的研究方法和需求与大数据分析历年不谋而合:1)事实数据在大数据情报分析中表现为对来自于多个数据源的大量数据的整合和融合利用;2)工具方法体现于大数据情报分析中对各种大数据工具和自动化处理技术的需求;3)而专家智慧则具体体现为通过智能关联、数据挖掘、深度学习等机器学习方法对数据和信息进行深层挖掘的需求。这些需求印证了大数

据分析技术的进步能够促进情报分析的发展。

2.1　多种数据的整合和融合利用

在大数据的环境[4]下,情报分析的数据来源和数据类型表现出空前的多元化特征,其中涉及的数据量越来越大,数据的类型变得更加复杂,尤其是非结构化数据所占的比重明显增大,数据的处理和分析难度增加,随之而来的对智能型数据分析工具和数据可视化工具等的要求也越来越高。大数据情报分析中的数据特征明显符合大数据的“数据量大(Volume)”、“多样性(Variety)”、“数据速率快(Velocity)”和“真实性(Veracity)”特性[5]。

2.1.1　数据量大(Volume)

1)大量数据源。数据的来源多种多样,而不同的数据源产生出的数据价值密度不尽相同甚至差异巨大,因此要从中筛选出高价值的数据源,或者根据价值密度的高低对不同的数据源设置不同的数据更新采集频率;另外,每一种数据源内的数据采集点巨大,以社交网络为例,每个用户作为一个采集点,Twitter 有 3 亿以上的用户,新浪微博有注册用户 5 亿以上、活跃用户 2 亿以上,因此要从这些潜在的采集点中找到有价值的采集点是一个巨大的挑战。

2)数据量大。由于大数据情报分析中数据量的巨大,对于大数据情报分析系统来说,一方面需要高效的数据存储方式作为基础,另一个重要方面就是必须支持对海量数据进行高效快速地处理和分析,提供对情报分析数据的全生命周期管理,同时需要支持对数据的离线批处理和实时在线分析。

3)冗余/无关数据量大。大数据情报分析的各个数据源每时每刻都在产生大量的数据,其中很可能会包括冗余、无关紧要的数据记录,正确地判断并且清除无关数据,消除多数据源之间信息冗余对于数据的高效存储、有效而准确地分析都显得非常有必要。

2.1.2　多样性(Variety)

1)数据来源的多样性。从传统的图书报纸等纸质出版物到网络化时代的电子出版物,互联网产生的政府、机构、公司等主页信息,互联网新闻信息,各种开放存取数据,近年来涌现出的大量社交网络(FaceBook、Twitter、微博、微信等)和电商网站信息使得情报分析的数据来源变得前所未有的丰富。

2)数据类型的多样性。一方面,由于数据来源的多样性,不同来源通常使用不同的数据类型,比如出版物多采用 PDF 格式并辅助以一定的元数据、社交网络数据通常是文本数据和视频数据的混合、门户网站和论坛通常是网页数据;另一方面,不同的行业通常采用的数据格式不同,比如制造业中有大量的 CAD 绘图文件、出版业中有对老书籍的扫描件等。各种各样的数据类型通常包括文本、网页、图片、PDF、CAD 绘图、视频、音频、扫描件等[6]。

3)行业多样性。除了门户网站、搜索引擎(百度、谷歌等)、电子商务网站(淘宝、亚马逊等)这些流量巨大、产生数据量也巨大的企业为代表的互联网数据外,大数据情报分析还涉及诸如医疗卫生、航空、地理信息、专利标准、影视娱乐、机械、科学研究等行业,情报大数据分析过程中需要统筹考虑来自于各个行业以及互联网的数据[7-8]。

4)语言多样性。语言的多样性源于大数据情报分析需要处理来自于不同国家、不同语种的信息,比如汉语、英语、德语、法语、韩语、西班牙语等;另外,我国是一个多民族的国家,也要充分考虑民族语言的多样性,比如藏语、维吾尔语、蒙语等不同民族所特有的语言。需要对来自于这些语言的情报信息处理和分析在统一的框架下进行。

2.1.3　数据速率快(Velocity)特性需求

1)流式数据处理。在大数据时代,数据的变化、变动或者产生的速度非常快,比如从服务器日志到各种各样的传感器每时每刻都在源源不断地产生新数据。大数据情报分析需要对这些流式数据进行实时采集和分析处理。另外,流式数据的高速率导致大数据量,从而难以对完整的数据流进行存储,因而需要对数据流进行在线分析并对数据进行摘要后存储。

2)高时效性分析。根据采集到的数据进行处理分析得到结果以快速地响应环境的变化和需求,特别是对于一些应用来说需要在很短的时间窗口内返回分析结果,超过一定时间窗口后返回的结果将失去应用意义。比如在金融情报分析系统中需要根据市场数据的变化实时快速分析出结果并做出决策。对于另外一些应用来说则需要对实时增量更新的数据进行分析得到结果。

2.1.4　准确性(Veracity)需求

1)歧义/冲突多。大数据情报分析由于其数据源多、数据多样、数据量巨大的特点,不同的数据源或者不同时刻采集到的数据会产生相互矛盾和冲突的数据记录,因此智能地消除信息的歧义,自动且智能地处理信息源之间的内容冲突的功能也变得不可或缺。

2)信息互补。单一数据源的数据有时仅提供了情报信息中的某一个侧面,如果要获取完整的情报信息需要融合多个信息源提供的互补信息或者对

多个信息源提供的信息进行相互印证。比如:通过一定蜂窝数据能够分析出我们的住所以及工作单位位置信息,而纳税信息能够推断出一个人的收入状况,通过诸多信息源信息的互补能够还原一个人的多方面信息。

2.2 大数据处理与分析工具和自动化处理

大数据情报分析需要采集海量的情报素材,然后对对海量的素材进行存储、预处理和分析,其中数据的存储包括对结构化和非结构化的数据的存储。对于不同来源的数据也需要能够对采集到的数据进行转化、冗余或者冲突数据的清除,以及对不同来源的数据进行融合,都需要大数据情报分析系统能够自动地完成,这就对大数据工具以及工具间作业流转的自动化提出了要求。总体来说,大数据情报分析对大数据工具和自动化处理技术的需求主要体现在大数据情报素材采集、大数据分布式存储、大数据并行计算平台、大数据分析算法和流程自动化方面。

大数据情报素材采集方面的需求主要包括:1)针对不同的数据源采用不同的采集方法;2)可配置、自适应的大数据情报素材采集系统,比如采集系统能够适应新的社交媒体内容或者经过简单配置后能够处理新的媒体内容;3)对于一些受限的信息源,能够突破这些限制。

大数据分布式存储、并行计算平台、分析算法、流程自动化的研究和发展为大数据情报分析提供了坚实的技术基础。目前,已经有很多的大数据技术服务提供商、互联网企业、研究机构和开源组织(比如 Apache Hadoop 和 Spark)致力于大数据的处理和分析技术研究与开发,提出了新的大数据存储与分析的方法和技术,并且开发除了具备相应功能的大数据存储和计算处理工具以及完整的通用大数据开源云计算平台 Hadoop、Spark 等[9-10]。而且,随着开源社区的不断发展壮大,这些开源软件的功能不断完善并增加。从大数据情报分析的角度来看,主要的需求是充分的利用开源社区的成果,针对大数据情报分析的特定需求开发或定制相应的模块。

2.3 大数据情报深度分析

深度分析是在预处理后的数据基础之上借助复杂的机器学习、信息关联、智能分析与可视化工具通过智能的方法将其转换为信息和知识的能力,这种能力主要体现在信息抽取、多元信息融合和深度挖掘 3 个方面[11-13]。

在信息抽取方面,在情报研究对象大幅度扩展的情况下,其中可能包含 Twitter、微博等社交媒体信息,由不同的用户产生不同呈现形式的数据,如数值型、文本型、图形图像、音频类型和视频类型,这些大量涌入的非结构或半结构化数据,必然需要通过预处理技术将这些数据转化为结构化数据,以供后续分析[14]。

在多元化信息方面则需要根据分析需求加以融合[15-16]。多源异构是大数据的基本特征之一,多元数据的融合也成为大数据分析处理的重要环节。根据实际的问题场景,多元信息的融合有利于进一步挖掘数据的价值,提升信息分析的有效性和准确性的作用;通过多元信息交叉印证,可以减少信息错误与疏漏,提供决策的准确性。对于大数据情报分析来说,多元化信息的融合已经成为一个重要的理念和必不可少的需求,具体的表现形式包括传感数据与社会数据的融合、历史数据与实时数据的融合、线上数据与线下数据的融合、内部数据与外部数据的融合等。

深度挖掘方面,针对海量的包含丰富而复杂信息的数据,简单的统计分析已不能满足决策需求,为了从中发现潜在模式以及关系,需要利用的算法包括简单方法、基于概率论的方法、基于模糊推理的方法以及人工智能算法等[17-21]。简单的算法包括加权平均、单元或者多元线性回归等[21]。基于概率的算法则有贝叶斯估计、贝叶斯滤波、贝叶斯推理网络和 D-S 证据理论等。基于模糊推理的方法则有处理数据模糊性、不完全行和不同粒度的模糊集和粗糙集方法[22-24]。人工智能计算方法如神经网络、遗传算法、蚁群算法、机器学习、深度学习算法可以处理不完善的数据,在处理数据的过程中不断地学习与归纳,从海量的数据中学习知识和发现规律。大数据情报分析的数据具有关系复杂、数据漂移、超高维、噪声多以及属性稀疏等特点,导致传统的数据挖掘和机器学习算法难以有效地进行数据处理和情报分析,为此需要研究新的机器学习理论和方法。另外,需要研究适合大数据分布式处理的数据挖掘编程模型和分布式并行化执行机制,支持数据挖掘算法中迭代、递归、聚合、集成、归并等复杂算法编程,以及在现有的并行计算平台上设计和实现复杂度低、并行性高的分布式并行化机器学习与数据挖掘算法。

3 大数据情报分析应用展望和发展机遇

大数据技术给情报分析的发展带来了深刻的影响和变革,也给情报学研究带来的前所未有的机遇,

如图 1 所示。在海量情报知识库构建管理平台以及高效能情报大数据存储与并行计算云平台的支撑之下,本文从情报大数据素材采集、数据预处理、数据分析和应用过程中的各个环节展望大数据情报分析将会发生的巨大变化。

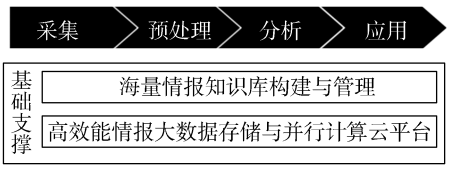


图 1　大数据情报分析展望示意图

Fig.1　Big data intelligence analysis outlook diagram

3.1　大数据情报素材采集

在大数据情报分析的数据和素材的采集阶段,海量网络信息采集系统将是一个具备以下功能和特征的智能系统:

1)通过智能的信息源发现与管理技术筛选并甄别有价值的信息源。不同的数据源包含的信息价值密度也不尽相同,过滤掉无价值或者价值过低的数据源可以有效地减少数据的存储与处理开销,更进一步提高后续分析的效率和准确度。

2)大规模网络信息获取需要支持实时、高并发、快速的网络内容获取。目前从网络产生的日志信息到机器传感器监测到的设备数据产生的速度非常快,大数据情报分析系统需要能够近实时快速地获取相关的数据。

3)通过受控信息源突破技术获取受控或者管制的信息,这些受控或受管制的信息可能会蕴含更大的价值,从而为后续分析提供更全面、更有价值的信息。信息系统中记录的主要是结果数据,实际上存在大量的过程数据并没有在数据库中记录,而这些过程数据以及中间结果信息对于情报信息分析具有重要作用,智能信息采集系统能够获取掩盖在业务应用系统之下的过程数据。

4)使用预处理技术移除冗余、无关信息。在采集到的素材经过大数据情报分析系统之前,通过清除无关信息以及不同数据源之间采集到的冗余数据,可以有效地减少下一阶段中数据处理的负担。

3.2　大数据情报预处理

不同的数据来源甚至同一数据来源都会产生格式不尽统一的数据。比如对同一个情报主题,情报数据可以由不同的网站和不同的用户产生,不仅不同的网站产生的数据模态不一致,即使同一个网站的每一个用户所产生的信息也可能会包含不同呈现形式的数据,如音频、视频、图片和文本等格式。这些结构化、半结构化甚至非结构化的多模态数据组合在一起导致大数据情报分析中的数据呈现出明显的异构性。数据融合以数据提取、转换、聚合为基础的核心技术,完成各异构数据源之间的数据分享与数据归并。利用异构信息融合技术,实现统一的数据检索和数据展现,将相互关联的分布式异构数据源融合后进行提取、转换、聚合,实现自动化构建专题数据库、领域数据仓库等功能。

专题数据库是以某一种产品或某一类技术为主题,对全部信息进行检索、下载、存储,收集到的专题信息数据的集合。发展专题信息提取技术,实现基于专题的高效检索、数据提取、数据归并等功能,根据用户需求对专题数据进行筛选。专题数据库将筛选后的专题数据集合进行归并入库,实现数据的检索、统计、分析等功能。

来自于分散的操作型数据,按照一定的主题域(领域)被抽取出来,进行加工与集成,统一与综合之后形成数据仓库。领域数据抽取时需要利用领域概念建模方法——需要运用实体建模法从纷繁的数据背后抽象出实体、事件、说明等抽象的实体,从而找出实体间的相互的关联性。这种方式可以保证数据仓库所需的数据能按照数据模型达到一致性和关联性。这些数据定义直接输入系统中,作为元数据存储,供数据管理和分析使用。

在数据的预处理阶段,由数据中间层在程序应用层与底层数据源之间构建统一的数据层,该层提供一个统一的数据逻辑视图来隐藏底层数据源的数据细节,使用户可以把各异构数据源看为一个统一的整体,能够用透明的方式访问各类数据。统一的数据中间层可以使得大数据情报分析对类型繁多、结构各异的多模态数据的访问和分析更加方便。这些不同类型的信息从不同的角度反映出事物的特征和信息,通过统一的数据接口将这些数据汇聚融合到一起,能够更加深刻全面地揭示事物之间的联系,挖掘出新的关联和模式等有价值的知识和情报信息。多模态数据的融合可以说是大数据情报分析的固有特征,也是其发展的必然趋势。

在数据预处理阶段需要进行的另一项重要工作是数据歧义消除和语义标签的计算。同一个词在不同的上下文中有不同的含义,以“apple”为例,在谈论公司的语境中的语义是生产计算机、手机等设备的美国苹果公司,在饮食相关语境中的含义则为水果。

3.3　数据分析

大数据情报分析的数据分析阶段主要涉及以下几个方面。

1)大数据情报信息挖掘。以大数据情报信息

挖掘理论、方法与工具为基础,比如数据抽取、聚类分析、时间和空间的序列模式分析、关联规则分析以及分类分析等,根据应用需求和数据基础,构建并综合应用上述各种模型,从经过预处理的情报素材中有目的地挖掘有价值的信息。并且在此过程中对于情报信息挖掘的共性问题分析逐步减少人工干预,提供探索式大数据情报挖掘环境,将情报信息挖掘方法与语义技术相结合,提升挖掘深度和准确度。在大数据情报信息挖掘理论的基础之上,利用大数据情报分析的方法和工具,可以进行包括主题情报聚合分析、趋势演变分析、社交媒体倾向性分析、线索挖掘以及情报预警等基于大数据情报分析的信息挖掘。

2)新型社交媒体分析。社交媒体服务的兴起产生了各种各样的社交媒体数据,比如:微博类网站的文本信息流数据、媒体分享网站的多媒体数据、社交网站的用户交互数据、签到网站的地理位置数据、购物网站的消费数据等[25]。这些社交媒体多源数据从不同角度记录着人们的网络生活,并映射着物理世界。社交媒体的多源主要体现在不同社交媒体网络所关注的异构用户行为信息,理解社交媒体多源现象对于社交媒体分析和社交媒体大数据的深度应用具有重要意义。社交媒体数据处理的重点方向包括社交网络中的多语信息处理(具有数据规模大、口语化严重、需要支持多种语言、社会群体特征明显等特点[26])、社交网络多语机器翻译、社交网络跨语检索以及社交网络情感分析。新型社交媒体的大数据情报分析是深度利用社交媒体大数据的关键,随着大数据情报分析技术的成熟,可以从社交媒体的数据中进行分析并从中挖掘宝贵的信息并为大规模的社交媒体应用提供有效使用的解决方案。

3)认知计算。情报学的分析方法将从原来的计算机辅助分析为主体转变为计算机认知为主体的智能分析,从而形成类似于 IBM Waston 的大数据情报认知计算及分析平台[27]。认知计算是综合了多种新兴技术的一个领域,并且将会对情报科学的发展产生深远的影响,比如认知情报学已经成为了情报学领域理论的一个重要研究方向,在情报分析方法、情报检索和信息资源建设领域,认知计算的相关技术也在起到日益重要的最用[28]。随着大数据情报分析技术的发展,传统的基于数据计算的挖掘技术正在向基于内容的知识发现技术发展,认知计算技术的发展可以有效的解决情报分析过程中知识处理的困难。

3.4 情报分析应用

大数据情报分析中,在前面数据采集、存储和处理分析技术的飞速发展的基础之上,如何让海量的数据集的应用变得简单和易于理解,可视化无疑是最有效的途径,所以可视化分析也将在大数据情报分析中得到极大应用。情报可视化技术主要以信息可视化分析系统为核心,能够自动化地实现多维信息可视化、领域知识可视化、情报预测评估可视化。能够提供强大的图形展现功能,将大量的、分散的、低关联的数据抽取整合,转化为图形中的节点数据,再由平台后台提供的丰富的图形分析算法,挖掘出数据之间隐藏着的关联关系,对各种维度、多层次、时空、动态、关系等类型的情报信息进行可视化展现。

可视化分析广泛应用于对于不易形成固定的分析流程或模式的场景,可视化数据分析平台,可辅助人工操作将数据进行关联分析交互式可视化分析能够引导数据探索、自动化实现预测分析,对数据加以可视化解释。典型的情报可视化分析包括多维信息可视化、领域知识可视化和预测分析的可视化[29-32]。实现可视化技术在海量信息组织方面的应用,能够利用二维或三维的概念图、认知地图、思维导图、趋势图、语义网络等图形化方式呈现情报信息,满足对热点情报、技术趋势的聚类信息展示和分析预警,及时感知行业最新动态和热点事件,为快速应对和采取措施提供直观的判断与决策依据。

3.5 高效能情报大数据存储与并行计算云平台

高效能的情报大数据存储与计算云平台是整个大数据情报分析系统的基础和支撑,提供的主要功能是基于云计算的多源异构大数据存储和管理,大规模增量实时数据的并行计算方法和面向异构数据的大规模并行处理体系结构。

高效能的大数据存储与并行计算云平台主要包括两个方面,一方面是是大数据情报分析中需要的海量数据的存储,另一方面是在大数据情报分析过程中的对海量数据进行并行分析计算的框架或者平台[33-36]。

对于大数据情报分析中的数据来说,传统的关系型数据库在处理此数量级的数据时候已经开始变得吃力,而分布式的存储系统可以用来存储如此海量的数据并对其进行管理。海量的数据系统选择将数据放在多个机器中,在解决存储容量问题的同时,也带来了许多单机系统不曾出现的问题,目前已经出现了很多的分布式数据存储解决方案,其中包括 Hadoop、Spark,各种非关系型数据库系统(比如

HBase、Cassandra、MongoDB 等)[37]。这些不同的解决方案针对不同的应用需求解决了满足了特定的要求,在应用到大数据情报分析中可以根据不同情报分析的具体需求采取不同的解决方案,或者将不同的解决方案组合在一起以满足特定的需求,随着大数据技术的发展,越来越多并且更加成熟的分布式数据存储解决方案会涌现出来并且被应用于大数据情报分析中去[38]。

大数据情报分析的核心在于对收集到的数据进行分析,从中获取有价值的信息和情报。对于海量数据的分析必然涉及各种复杂的计算,对于高效的并行计算的需求不言而喻。伴随着海量数据的存储方案的出现,各种不同的大数据分布式计算框架也被提出来,其中 Hadoop MapReduce、Spark 和 Storm 是目前最重要的三大分布式计算框架,这 3 种不同的框架侧重点不同,解决的问题也不相同[39-40]。Hadoop MapReduce 常用于解决离线的复杂的大数据处理,Spark 常用于进行离线的快速的大数据处理,而 Storm 常用于进行实时在线的大数据处理。不同的计算框架具有各自不同的优点和缺点:Hadoop MapReduce 易于编程、具有良好的扩展性、高容错性、适合 PB 级以上的海量数据的离线处理,但是不支持实时计算和流式计算;Spark 是一种基于内存的迭代计算框架,通过将中间数据放置于内存中,获得了更高的迭代计算效率,弹性分布数据集(resilient distributed dataset, RDD)对于数据的抽象更高级,通过 Checkpoint 实现容错,Spark 的编程模型比 Hadoop MapReduce 更加灵活,但是 Spark 并不适合那些需要异步地对数据状态进行细粒度更新的应用,也就是说,Spark 并不适合需要增量修改的应用模型;Storm 适合于流数据处理,可以用来对源源不断流进来的消息进行处理,并且将处理之后的结果写入到制定的存储设备中去,Storm 另一个主要应用便是实时对数据进行处理,数据不需要写入到磁盘等存储设备中,延迟很低一般在毫秒级,特别适合于大数据情报分析中需要实时在线分析得到结果的场景。

高效的存储解决方案以及并行计算框架是大数据情报分析的重要基础支撑,可以保证海量数据的高效存储,同时支持对海量数据的离线批处理分析以及实时在线交互计算,为情报分析人员提供了强大的分析工具[41]。

3.6　海量情报知识库构建与管理维护

知识库是知识的集合,知识库系统是现代许多智能系统的关键基础部件[42-44]。情报知识库是基于信息技术建立的情报知识管理系统,是情报分析系统的重要组成部分,特别是对于大数据情报分析来说,完善高效的海量情报知识库显得尤为重要[45-46]。海量情报知识库主要分为 3 个组件:语言学相关知识库、行业情报知识库和知识库管理系统。

1)语言学相关知识库包括语言知识库,翻译语料库和分类语料库,主要用于获取语言知识比如词性标注、词义标注、搭配规则和语法规则等,为行业情报知识库分析提供基础。

2)行业情报知识库包括领域本体库、机构知识库和叙词库等,存储了海量情报知识库的数据本体。

3)知识库管理则主要是通过海量数据根据一定的规则进行自动学习,从而达到自动动态更新知识库的效果。知识库管理还需要对知识库的访问接口(如 API 等)标准化,以便于知识库中内容的共享,提高知识库的利用效率。

海量情报知识库的高效维护和管理也为大数据情报分析提供坚实的基础。同时,随着信息技术以及各个行业数据的不断扩充演化,需要知识库管理系统能够动态地自适应学习扩充已有的知识。

4　结束语

在大数据时代,情报分析的发展正在发生着重大的变革,大数据情报分析已经在各个方面对传统的情报分析产生深刻的影响。本文在阐述了大数据情报分析的发展范式以及现状以后,对大数据情报分析的所面临的应用需求和挑战从多种数据的整合和融合利用、大数据情报分析的方法和工具以及对深度分析方面进行了详细的分析,最后从大数据情报分析具体过程中数据的采集、处理、分析和应用各个阶段对大数据情报分析在技术和发展机遇方面进行了展望。随着大数据技术的不断发展,大数据情报分析也会越来越成熟、越来越向智能化的方向发展,从而更好地迎接更加复杂情报分析需求带来的挑战。

参考文献:

[1]GINSBERG J, MOHEBBI M H, PATEL R S, et al. Detecting influenza epidemics using search engine query data[J]. Nature, 2009, 457(7232): 1012-1014.

[2]包昌火. 情报研究方法论[M]. 北京: 科学技术文献出版社, 1990.
BAO Changhuo. Information research methodology[M]. Beijing: Science and Technology Literature Publishing House, 1990.

[3]WEISS G. A Modern approach to distributed artificial intelligence[J]. IEEE transactions on systems man & cybernetics

-part c applications & reviews, 1999, 22(2).

[4] MANYIKA J, CHUI M, BUGHIN J, et al. Big data: the next frontier for innovation, competition, and productivity [R]. McKinsey Global Institute, 2011.

[5] ETEMADPOUR R, MURRAY P, FORBES A G. Evaluating density-based motion for big data visual analytics [C]// Proceedings of IEEE International Conference on Big Data. Washington, DC, USA, 2014: 451-460.

[6] SONG Jingkuan, YANG Yang, YANG Yi, et al. Inter-media hashing for large-scale retrieval from heterogeneous data sources [C]//Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. New York, NY, USA, 2013: 785-796.

[7] RAGHUPATHI W, RAGHUPATHI V. Big data analytics in healthcare: promise and potential [J]. Health information science and systems, 2014, 2: 3.

[8] PIRES A J M. Big data analytics in healthcare: are end-users ready [D]. Braga: Universidade Católica Portuguesa, 2014.

[9] SHVACHKO K, KUANG Hairong, RADIA S, et al. The hadoop distributed file system [C]//Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies. Incline Village, NV, USA, 2010: 1-10.

[10] ZAHARIA M, CHOWDHURY M, FRANKLIN M J, et al. Spark: cluster computing with working sets [C]//Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing. Berkeley, CA, USA, 2010: 10.

[11] JUNG K, KIM K I, JAIN A K. Text information extraction in images and video: a survey [J]. Pattern recognition, 2004, 37(5): 977-997.

[12] SODERLAND S. Learning information extraction rules for semi-structured and free text [J]. Machine learning, 1999, 34(1/2/3): 233-272.

[13] ZHANG Yongmian, JI Qiang. Active and dynamic information fusion for facial expression understanding from image sequences [J]. IEEE transactions on pattern analysis and machine intelligence, 2005, 27(5): 699-714.

[14] SU Xueyuan, SWART G. Oracle in-database hadoop: when mapreduce meets RDBMS [C]//Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. Scottsdale, AZ, USA, 2012: 779-790.

[15] TAHANI H, KELLER J M. Information fusion in computer vision using the fuzzy integral [J]. IEEE transactions on systems, man, and cybernetics, 1990, 20(3): 733-741.

[16] WANG Jun, HU Yiming. WOLF-a novel reordering write buffer to boost the performance of log-structured file system [C]//Proceedings of the 1st USENIX Conference on File and Storage Technologies. Monterey, CA, USA, 2002: 4.

[17] 孟小峰, 慈祥. 大数据管理: 概念、技术与挑战 [J]. 计算机研究与发展, 2013, 50(1): 146-169.
MENG Xiaofeng, CI Xiang. Big data management: concepts, techniques and challenges [J]. Journal of computer research and development, 2013, 50(1): 146-169.

[18] WU Xindong, ZHU Xingquan, WU Gongqing, et al. Data mining with big data [J]. IEEE transactions on knowledge and data engineering, 2014, 26(1): 97-107.

[19] KOVAR L, GLEICHER M. Automated extraction and parameterization of motions in large data sets [J]. ACM transactions on graphics, 2004, 23(3): 559-568.

[20] LAZER D, KENNEDY R, KING G, et al. The parable of Google flu: traps in big data analysis [J]. Science, 2014, 343(6176): 1203-1205.

[21] FAN Jianqing, HAN Fang, LIU Han. Challenges of big data analysis [J]. National science review, 2014, 1(2): 293-314.

[22] SCHMIDHUBER J. Deep learning in neural networks: an overview [J]. Neural networks, 2015, 61: 85-117.

[23] CARLSON A, BETTERIDGE J, KISIEL B, et al. Toward an architecture for never-ending language learning [C]// AAAI 2010 Twenty-Fourth AAAI Conference on Artificial Intelligence. Atlanta, Georgia, USA, 2010: 529-573.

[24] BLUM A L, LANGLEY P. Selection of relevant features and examples in machine learning [J]. Artificial intelligence, 1997, 97(1/2): 245-271.

[25] JIN Songchang, LIN Wangqun, YIN Hong, et al. Community structure mining in big data social media networks with MapReduce [J]. Cluster computing, 2015, 18(3): 999-1010.

[26] TANG Jiliang, LIU Huan. Unsupervised feature selection for linked social media data [C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 904-912.

[27] CASSIDY A S, MEROLLA P, ARTHUR J V, et al. Cognitive computing building block: a versatile and efficient digital neuron model for neurosynaptic cores [C]//Proceedings of the 2013 International Joint Conference on Neural Networks. Dallas, TX, USA, 2013: 1-10.

[28] PREISSL R, WONG T M, DATTA P, et al. Compass: a scalable simulator for an architecture for cognitive computing [C]//Proceedings of the 2012 International Conference on High Performance Computing, Networking, Storage and Analysis. Salt Lake City, UT, USA, 2012: 1-11.

[29] KEIM D, QU Huamin, MA K L. Big-data visualization [J]. IEEE computer graphics and applications, 2013, 33(4): 20-21.

[30] MEYEROVICH L A, TOROK M E, ATKINSON E, et al. Superconductor: a language for big data visualization [M]. Shenzhen, China: ACM, 2013.

[31] HACHET M, KRUIJFF E. Guest editor´s introduction: special section on the ACM symposium on virtual reality software and technology[J]. IEEE transactions on visualization and computer graphics, 2010, 16(1): 2-3.

[32] CHILDS H, BRUGGER E, BONNELL K, et al. A contract based system for large data visualization[C]//Proceedings of VIS 05. IEEE Visualization. Minneapolis, MN, USA, 2005: 191-198.

[33] KANOV K, PERLMAN E, BURNS R, et al. I/O streaming evaluation of batch queries for data-intensive computational turbulence[C]//Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis. Seattle, WA, USA, 2011: 1-10.

[34] FRASCA M, PRABHAKAR R, RAGHAVAN P, et al. Virtual I/O caching: dynamic storage cache management for concurrent workloads[C]//Proceedings of 2011 International Conference on High Performance Computing Networking, Storage and Analysis. Seattle, WA, USA, 2011: 1-11.

[35] 张建勋, 古志民, 郑超. 云计算研究进展综述[J]. 计算机应用研究, 2010, 27(2): 429-433.
ZHANG Jianxun, GU Zhimin, ZHENG Chao. Survey of research progress on cloud computing[J]. Application research of computers, 2010, 27(2): 429-433.

[36] WANG Guojun, LIU Qin, WU Jie. Hierarchical attribute-based encryption for fine-grained access control in cloud storage services[C]//Proceedings of the 17th ACM conference on Computer and communications security. Chicago, Illinois, USA, 2010: 735-737.

[37] CHANG F, DEAN J, GHEMAWAT S, et al. Bigtable: a distributed storage system for structured data[J]. ACM transactions on computer systems, 2008, 26(2): 4.

[38] ARMBRUST M, FOX A, GRIFFITH R, et al. Above the clouds: a Berkeley view of cloud computing[R]. Technical Report No. UCB/EECS-2009-28. Berkeley: EECS Department University of California Berkeley, 2009: 50-58.

[39] DEAN J, Ghemawat S. MapReduce: simplified data processing on large clusters[C]//Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation. San Francisco, CA, USA, 2004: 107-113.

[40] IQBAL M H, SOOMRO T R. Big data analysis: apache storm perspective[J]. International journal of computer trends and technology, 2015, 19(1): 9-14.

[41] WANG Cong, CHOW S S M, WANG Qian, et al. Privacy-preserving public auditing for secure cloud storage[J]. IEEE transactions on computers, 2013, 62(2): 362-375.

[42] KATSUNO H, MENDELZON A O. Propositional knowledge base revision and minimal change[J]. Artificial intelligence, 1991, 52(3): 263-294.

[43] HOFFART J, SUCHANEK F M, BERBERICH K, et al. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia[J]. Artificial intelligence, 2013, 194: 28-61.

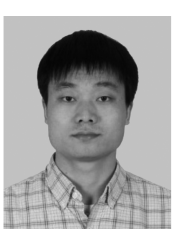
[44] LEHMANN D, MAGIDOR M. What does a conditional knowledge base entail[J]. Artificial intelligence, 1992, 55(1): 1-60.

[45] BARBARá D, GARCIA-MOLINA H, PORTER D. The management of probabilistic data[J]. IEEE transactions on knowledge and data engineering, 1992, 4(5): 487-502.

[46] KOUBARAKIS M, SKIADOPOULOS S, TRYFONOPOULOS C. Logic and computational complexity for Boolean information retrieval[J]. IEEE transactions on knowledge and data engineering, 2006, 18(12): 1659-1666.

作者简介:

黄河燕,女,1963 年生,教授。任中国人工智能学会和中国中文信息学会副理事长。主要研究方向为机器翻译、自然语言处理、社会计算。曾获国家科技进步一等奖、中国科学院科技进步一等奖和北京市科学技术一等奖等奖励。发表学术论文多篇。



曹朝,男,1982 年生,副研究员,博士,中国计算机学会数据库专委会委员。主要研究方向为数据库管理系统、分布式系统、智能信息处理。发表学术论文多篇。

冯冲,男,1977 年生,副研究员,博士,中文信息学会社会媒体处理专委会委员、语言与知识计算专委会委员。主要研究方向为网络信息抽取和多语机器翻译。曾获部级科技奖励 3 项。发表学术论文 30 余篇、编著 1 部,申请专利 10 余项。