数据挖掘视角下的情报分析研究*

王翔[1,2]，侯威[1]
（1. 安徽省科学技术情报研究所，合肥 230011；2. 合肥工业大学计算机与信息学院，合肥 230009）

摘要：随着大数据技术的深入发展及其在图书情报领域的积极应用，大数据时代的情报分析已成为热点研究问题。通过对大数据环境下的情报搜集、情报来源融合、数据存储、数据挖掘算法以及分析结果可视化等问题的研究，从数据挖掘视角分析了大数据时代情报分析领域存在的挑战与机遇，并提出了一种适用于大数据环境的采用数据挖掘技术的情报分析模型。该模型的建立为情报分析提供了一种新的理论模型及具有较好可操作性的方法。

关键词：数据挖掘；情报分析；大数据
分类号：G350.7
DOI：10.3772/j.issn.1673-2286.2015.09.008

1 引言

随着广东、上海、北京等地纷纷启动大数据战略以推动政府转型，政府推动的大数据变革给情报分析研究带来了巨大的挑战与机遇，如何应用大数据技术支持情报分析成为当前学术研究的热点。如，Bonnie Hohhof[1]指出从社交网站及其他内部网络收集来的大数据会对竞争情报分析产生巨大挑战；Hsinchun Chen[2]提出了对商业智能应用及新兴研究领域进行划分的框架，并通过文献计量方法探讨高质量论文如何影响商业智能应用；Michael Minelli[3]综合分析了大数据技术与信息管理及商业应用的结合点，认为将大数据技术应用到商业情报分析中可获取巨大的收益及更高的效率；Ping-Tsai Chung[4]等通过对比数据挖掘工具与传统商业分析方法，认为数据挖掘将成为今后竞争情报系统建设的重要组成部分；贺德方[5][11]分析了大数据给传统情报学带来的挑战与机遇，认为情报学研究应积极与大数据研究和发展结合；黄晓斌[6][9]等人分析了大数据时代企业竞争情报研究的创新与发展，提出了基于大数据的企业竞争情报系统模型；吴金红[7][10]等人探讨了大数据环境将如何影响技术竞争情报分析的服务理念、模式与方法，基于此提出应对策略；王晓佳[8]等从情报分析实践出发，阐述了大数据时代情报分析与挖掘技术结合的建模机理，并予以验证；张玉峰[12]等构建了基于数据挖掘的企业竞争情报分析模型，实现了语义智能挖掘与分析，在数据挖掘与情报分析结合方面做出有益尝试。

本文从数据挖掘视角，分析大数据时代情报分析领域存在的挑战与机遇，从大数据环境下的情报搜集、情报来源融合、数据存储、数据挖掘算法以及分析结果可视化等方面，提出一种基于数据挖掘算法的适用于大数据环境的情报分析模型，以提高竞争情报源数据预处理能力，提升竞争情报分析及结果可视化展现的时间性能。

2 大数据环境下情报分析面临的挑战与机遇

传统的情报分析较多基于人的智力加工，为了提高情报分析的真实性及精准性，往往要求情报分析人员

* 本研究得到2014年度安徽省自然科学基金项目“基于海量科学文献的科研合作社团发现及评价关键问题研究”（编号：1408085QF136）资助。

收集全面的数据或信息，进而从信息中整理、提取、加工成有价值的情报。这一过程中，也会使用统计学工具来处理结构化数据，减少工作量。随着大数据时代的到来，传统的情报分析在应对这些数据量庞大、变化快、类型多样且价值非常稀疏的数据时，往往无法有效收集、存储、分析大数据，在挖掘与辅助决策方面也存在较多问题。

数据挖掘研究的内容就是如何从大量的、不完全的、有噪声的、模糊的、随机的数据中提取隐含在其中的、事先不知道的，但又是潜在有用的信息和知识。应用好数据挖掘技术，就可以解决传统分析方法无法深入挖掘隐含在各种信息背后的知识这一难题，进一步提升情报分析的效率与深度。从数据挖掘的视角看，大数据环境下的情报分析面临挑战的同时，也存在较好的发展机遇。

2.1 大数据环境下情报分析的挑战

（1）大数据的有效存储

大数据时代，为了提高情报分析真实性，提升情报分析精准度，需要收集和处理大量详实的情报资源，其数据量远远超过传统科技情报数据存储规模，甚至无法用传统的数据库去管理（如视频、非结构化事实性数据等）。而收集、存储和维护这样庞大的数据对于一般的单位或部门来说也是很大的负担，如何有效存储信息资源成为大数据环境下情报分析需要解决的首要难题。

（2）数据统一表示与数据源融合

数据收集的完备程度直接影响了情报分析的结果。能否将不同来源的数据通过不同主题进行划分，建立数据仓库，为决策分析之用，成为判定情报分析系统优劣的关键因素之一。

传统的情报分析多局限于文本数据源，如文献资源、Web信息、专利资源等，而现实中的数据源还包括查新报告、科技报告、科技计划项目立项书、政府公文等非结构化文档，甚至是图片、视频等文件，因此必须建立统一的数据表示模型，即采用统一的模型融合多个异构数据源中的数据。如对于给定的一种半结构化或非结构化数据（图像），如何把它转化成多维数据表、面向对象的数据模型或者直接基于图像的数据模型，这是大数据时代情报分析的重要环节之一。同时，由于大数据的存储量多达TB甚至是PB级别，必须设计算法从上述情报资源中抽取所需的有效信息，避免无用的信息，从而提高抽取信息的效率与质量。因此，数据统一表示与多数据源融合成为大数据环境下情报分析的一个重要研究内容。

（3）适应大数据环境的情报分析数据挖掘算法

大数据环境下的情报分析必须满足确保时效性、处理增量式数据、处理流数据、处理分布式数据的要求，特别是将数据挖掘方法与情报分析结合时，更要注意满足以上要求。

传统的数据挖掘算法（如分类算法、关联规则等）把挖掘结果精确度放在第一位。在大数据时代，对分析结果的准确度要求往往弱于时效性要求。传统的情报分析方法或数据挖掘算法多采取集中式处理情报资源的方式，而在大数据时代，情报分析需要处理螺旋式爆炸增长的数据，很难采用传统的情报分析或数据挖掘模式去处理问题，采用数据挖掘与情报分析结合的方法，必须能处理增量式数据。

动态数据流是大数据的主要特征之一，有了分布式的文件系统支撑之后，也必须进行数据流处理才能发挥其效用。但是目前情报分析工具基本不具备分布式流处理的功能，对许多实时数据的处理无能为力。采用数据挖掘思想的情报分析方法能对流数据进行抓取、分析和挖掘，由于单一节点很难完成大量情报资源的分析工作，必须保证分析方法可移植到分布式环境或并行计算环境。

（4）脏数据与丢失信息处理

传统的情报分析多局限于较为“纯净”的情报资源，通过人工分析进行情报资源的清洗。随着大数据时代的来临，传统的人工清洗数据方式很难在大规模数据中发现可能存在的“脏数据”，而这些“脏数据”的存在可能影响到最终分析结果的真实性，必须在进行情报分析之初就对“脏数据”进行清洗。

由于传统情报分析将半结构化数据或非结构化数据转化为结构化数据再进行处理，这一过程可能导致丢失非结构化数据中隐含的关系，进而导致分析结果的不确定性。情报分析面对的往往是价值稀疏且存在较多冗余的数据，需要在情报分析之前，对数据进行预处理，去除冗余数据同时，通过特征属性提取等方法，在高维稀疏的数据中，抽取对分析目标最重要的数据特征，从而减少数据挖掘工作量，提高情报分析效率。

（5）高级数据可视化需求日益迫切

大数据时代，情报分析的结果需要最终汇总整合并达到用户可以理解的程度，简单的统计分析表格或关联分析规则等初级可视化工具已不能满足用户需求，这就必须充分考虑情报分析结果整合和结果如何呈现给客户的问题，高级可视化分析能够直观的呈现大数据背景下情报分析特点，同时能够被用户所接受。

2.2 大数据环境下情报分析的机遇

从数据挖掘的视角看，数据规模庞大、数据类型复杂、数据源的多样引起的情报分析难题都可以通过数据挖掘技术手段的发展去解决，而在此过程中，情报分析与数据挖掘结合后将可能产生极好的社会价值，对情报分析发展而言无疑是很好的机遇。

（1）数据的全面社会化对情报分析具有极大促进作用

占有足够多的数据是数据挖掘结果有效性的重要保证，对于情报分析而言，分析人员用以分析的数据越全面，最终得出的分析结果可能越接近真实。大数据时代的一个重要趋势是数据的社会化，从Web、社交网络（微博、微信、博客、Facebook等）、各类论坛等网络活动场所（包括移动互联），借助数据挖掘技术，随处可以收集到用于情报分析的资料，情报分析人员比以往任何时候都便捷的获取分析对象的第一手资讯，极大提升情报分析的真实性与精准性。

（2）专业化情报分析的研究地位将极大提升

随着大数据概念不断深入人心，人们将逐渐改变以往依赖独立的内部信息和对外部世界的简单直觉作为依据的决策方式，将逐渐接受依据事实而获得的可执行的情报。专业化的情报分析研究将有能力整合、分析和开发结构性数据和非结构性数据，帮助各行业的领导者改善决策，极大提升人们对专业化情报分析的认知度，实现从信息时代至分析力时代的转变。在这一过程中，理解业务需求、熟悉相关数据挖掘和情报分析技术方法、了解数据资源的情报人员也将扮演越来越重要的角色[1]。

（3）情报分析实时性要求更高

随着数据存储技术及数据挖掘技术的进步，情报分析的难点将不再是能否分析大数据，更多的是考虑如何提升分析的效率，人们对情报分析的实时性要求将更高。数据挖掘技术的引入使得人们发现隐藏在数据背后的知识的可能性极大提升，同时，有可能将最终提交人类进行决策分析的情报数目控制在可接受的范围内，进一步提升情报分析的实用性。

（4） 政府开放数据共享

政府主导开发的电子政务信息系统，其中的数据具有较高的社会价值，但由于缺乏相应的放开机制，有相当一部分逐渐变成“信息孤岛”。随着政府对大数据的不断重视，不断开放数据库，大量的相关信息以数据的形式生成、处理和存储，情报分析人员用这些数据就能创造价值，可能创造新的服务行业。

（5）数据可视化程度提高

对于大多数人而言，很难掌握深厚的数据挖掘技术，如果没有能帮助人们理解大数据的工具，许多信息背后的知识可能很难为人所知晓，数据可视化工具能够有效的呈现数据之间的内在关联。包括Visual.ly、Tableau、Vizify、D3.js和R语言在内的很多可视化工具可以帮助人们更容易、更快速的从越来越大的各种数据集中发现新的内容。如果能很好的应用数据可视化工具，许多非技术专业人士也可能从大数据角度提出见解。

3 基于数据挖掘的情报分析模型

将数据挖掘方法引入情报分析已有较多研究，但这些研究较多关注于将传统的情报挖掘方法移植到云计算环境。如采用Map Reduce方法将各种已有的文本挖掘算法（聚类、分类、序列分析、关联规则等）应用于Hadoop平台架构[13-14]，或将数据采集方法移植到云计算环境下，再采用传统的情报挖掘算法进行分析。

现实中的情报分析除了要搜集Web信息、微博、微信等半结构化信息，还需要搜集如文献资源、专利资源、查新报告、政府公文等非结构化的文档，在构建情报分析挖掘模型阶段，必须在研究情报分析挖掘算法的同时，统一考虑如何利用云计算的海量存储能力将各类情报分析数据源融合，拓宽情报分析数据来源，这样才能充分适应大数据环境下的情报分析需求，得到的情报分析结果才能更及时、更准确。另外，情报分析的最终目标是满足用户需求，如何将各个节点分析的结果融合并将机器分析的结果可视地展现给用户也是必须要包含在的内容。

在上述研究的基础上，本文提出了适用于大数据环境的采用数据挖掘算法的情报分析模型，如图1所示。该模型主要由以下几部分组成：

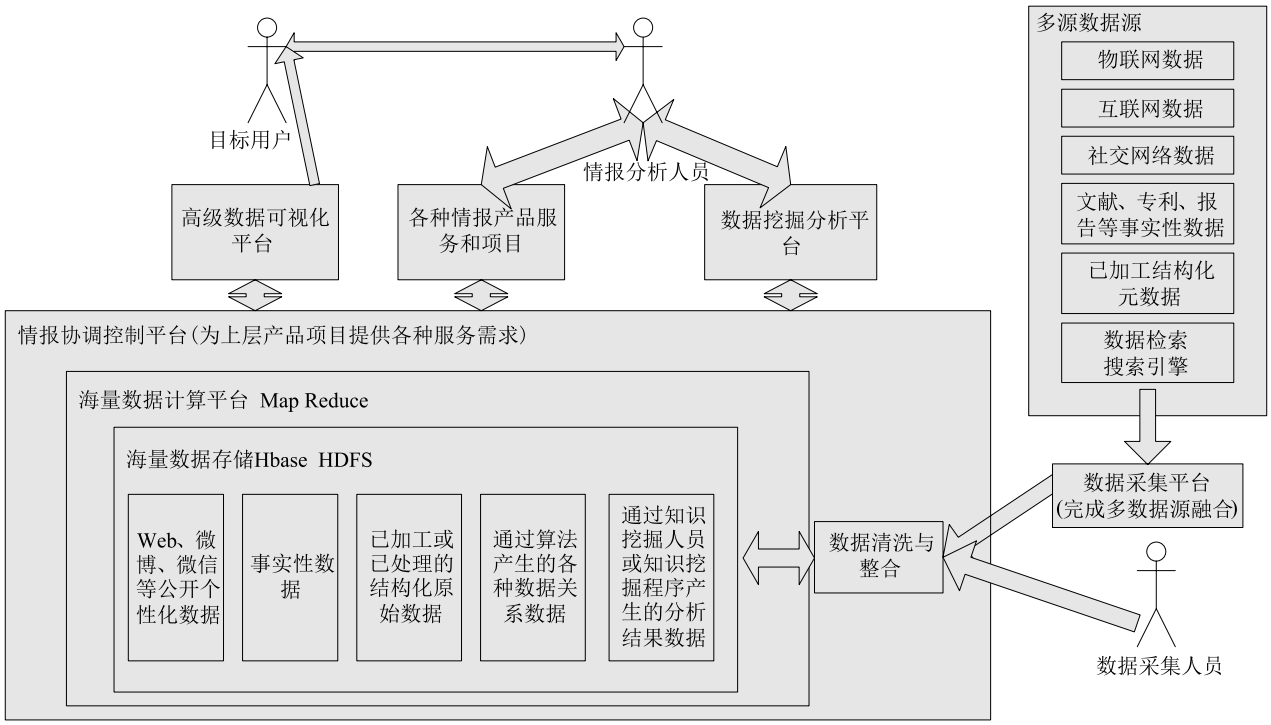

图1 适用于大数据环境的情报分析模型

（1） 基于Hadoop/Map Reduce的大数据存储与计算

考虑到该模型需要处理的资源异常庞大，且为保证一定的资源存储能力，拟采取基于大规模云服务应用的Hadoop底层架构。底层是HDFS，用以存储海量和多类型的数据，使用Hbase统一管理各类数据，借助Map Reduce的计算能力，分解计算任务并重组结果，由数据可视化模块或情报分析模块统一展示给用户。

该底层架构普遍应用于大规模用户群体和大数据处理平台上，该架构是一种能够对大量数据进行分布式处理的云计算软件框架，能解决许多要求极大伸缩性的问题，可以广泛运用在分析处理TB级的数据文件上，大大提高处理效率。

（2）多数据源融合与清洗平台

来自物联网、互联网、SNS社交网络、专利及文献等事实性数据、已加工的结构化元数据以及来自人工搜索整理的数据，通过数据采集平台收集整理进行初步加工（去除重复与冗余数据）后，利用数据清洗平台完成最终的数据入库过程，洗去其中的“脏数据”，必要的时候还需要人工干预，避免出现遗漏重要数据等问题。

（3）数据挖掘分析平台

传统情报分析方法不下100种，能直接用计算机描述并适合云计算环境的算法却寥寥无几，情报分析与挖掘算法的实现是情报挖掘的关键难题。数据挖掘分析平台中的数据挖掘与情报分析算法将计算向存储迁移，在存储节点完成各项任务，既保证每个分治节点能协作完成情报挖掘与分析工作，也能独立完成个性化信息分析工作。

情报分析是人智力加工的产物，所有计算机辅助产物都是在为减低人工工作量及提升决策准确性做帮助，最终由数据挖掘分析平台产出的结果可能是一组关联规则，也可能是分类的辅助信息，情报分析人员对这些决策辅助信息做出解释与判断，最终的决定权还是由情报分析人员做出并提交用户。

（4）高级数据可视化平台

整个模型产出的结果，除了情报分析人员做出的专业判断，还可以借助第三方工具如Visual.ly、R等，将每一个决策数据项作为单个图元元素表示，由全部的决策数据集构成数据图像，同时将数据的各个属性值以多维数据的形式表示，帮助人们更容易、更快速的从数据集中发现新的东西，理解情报分析的结果。

4 结语

大数据是寻求搜集新技术、新思想的一种强大的

发现工具，情报分析是借助这一工具开展知识发现的实践，数据挖掘则是这一实践的有益补充。本文从数据挖掘角度分析大数据环境下情报分析的机遇与挑战，并提出基于数据挖掘的情报分析模型，是拓展大数据战略背景下情报分析研究思路的一次有益尝试，如何将模型付诸情报分析实践，在真实大数据环境中得到实际应用，还有待进一步研究。

参考文献

[1] Hohhof B. Big data and competitive intelligence [J]. Competitive Intelligence Magazine, 2012, 15(3):5-6.

[2] Chen H, Chiang R H L, Storey V C. Business intelligence and analytics: from big data to big impact [J]. MIS Quarterly, 2012, 36(4): 1165-1188.

[3] Minelli M, Chambers M, Dhiraj A. Big data, big analytics: emerging business intelligence and analytic trends for today's businesses [M]. John Wiley & Sons, 2012.

[4] Chung P T, Chung, S H. On data integration and data mining for developing business intelligence[C]. Systems, Applications and Technology Conference (LISAT), 2013 IEEE Long Island. IEEE, 2013: 1-6.

[5] 贺德方. 大数据环境下的情报学[J]. 数字图书馆论坛,2012 (11):2-5.

[6] 黄晓斌,钟辉新. 大数据时代企业竞争情报研究的创新与发展[J]. 图书与情报,2012 (6):9-14.

[7] 吴金红,张飞,鞠秀芳. 大数据：企业竞争情报的机遇、挑战及对策研究[J]. 情报杂志,2013,32 (1):5-9.

[8] 王晓佳,杨善林,陈志强. 大数据时代下的情报分析与挖掘技术研究—电信客户流失情况分析[J]. 情报学报,2013,32(6):564-574.

[9] 黄晓斌,钟辉新. 基于大数据的企业竞争情报系统模型构建[J]. 情报杂志，2013,32 (3):37-43.

[10] 王翠波,吴金红. 大数据环境下技术竞争情报分析的挑战及其应对策略[J]. 情报杂志,2014,33(3):6-10.

[11] 贺德方. 基于大数据、云服务的科技情报工作思考[J]. 数字图书馆论坛,2013 (6):2-9.

[12] 张玉峰,何超. 基于数据挖掘的企业竞争情报分析研究[J]. 情报学报,2012,31(1):65-71.

[13] 李军华. 云计算及若干数据挖掘算法的Map Reduce化研究 [D]. 成都:电子科技大学,2010.

[14] 陈晓美,孙中秋,王秀艳,等. 大数据时代数字资源整合与聚合研究[J]. 数字图书馆论坛,2014 (6):28-34.

作者简介

王翔，1982年生，安徽省科学技术情报研究所文献中心暨NSTL合肥服务站副主任，助理研究员，合肥工业大学博士生，研究方向：数据挖掘、图书文献研究，E-mail：wangxiang@ahinfo.gov.cn。

侯威，1963年生，安徽省科学技术情报研究所文献中心暨NSTL合肥服务站主任，助理研究员。

Research of Information Analysis in Big Data Era from the Perspective of Data Mining

WANG Xiang[1,2] , HOU Wei[1]

(1. Institute of scientific and technical information of Anhui, Hefei 230011, China;
2. School of computer and information, Hefei University of Technology, Hefei 230009, China)

Abstract: Based on the research into the important hot issues of information analysis in the Big Data era, this paper studies the problem of information collection, information integration, data storage, data mining algorithm and information analysis results visualization in the Big Data environment. This paper analyzes the challenges and opportunities in the field of information analysis from the perspective of data mining, and proposes a kind of information analysis model using data mining technology，which is suitable for Big Data environment. Through the establishment of the model ,a new theoretical model for the analysis of information and a better operational method is explored.

Keywords: Data Mining; Information Analysis; Dig Data

（收稿日期：2015-08-10；编辑：雷雪）