情报科学
第37卷 第3期 2019年3月

科技大数据的情报分析技术研究

曾　文[1],车　尧[1,2]

(1.中国科学技术信息研究所,北京 100038; 2.《情报学报》编辑部,北京 100045)

摘　要:【目的/意义】现代科学技术的进步和发展给情报研究工作,特别是给情报分析技术带来了变化。传统的情报分析技术面对海量数据的快速增长和技术进步的事实,采用新方法和新技术充实到科技情报分析过程中已是必然趋势。【方法/过程】本文以科技大数据为研究和分析对象,论述国内外的相关研究现状,介绍科技大数据的建模和分析流程,阐述科技大数据分析平台的设计和研发工作。【结果/结论】论文通过实例介绍科技数据分析平台的数据分析过程,为实现科技大数据情报分析平台的实用化奠定了研究基础和方法。

关键词: 科技大数据;情报分析;分析平台

中图分类号:G206　　DOI:10.13833/j.issn.1007-7634.2019.03.016

Research on Information Analysis Technology on Science and Technology Big Data

ZENG Wen[1], *CHE Yao*[1,2]

(1. *Institute of Scientific and Technical Information of China, Beijing* 100038,*China;*
2.*Editorial Department of Journal of the China Society for Scientific and Technical Information* 100038*, Beijing*)

Abstract:【Purpose/significance】The progress and development of modern science and technology has brought changes to the information research work, especially to the information analysis technology. The traditional information analysis technology is facing the fact that the rapid growth of the massive data and the technological progress. It is an inevitable trend to use new methods and technologies to enrich the analysis process of scientific and technological information.【Method/process】In this paper, from the perspective of science and technology of big data, described the research status, introduced data modeling and analysis process.【Result/conclusion】It introduced design and development work of analysis platform and its data analysis process. The work of the paper laid the research foundation and method for practicability of data analysis platform.

Keywords: science and technology big data; information; analysis platform

1 引　言

人类的情报活动可以追溯至远古,而作为一项专门化的现代情报研究活动则是科学研究和信息技术发展到一定阶段的产物。情报分析技术的目的是支持情报研究的方法,数据则是情报分析技术的研究对象。科技大数据是一种特殊类型的大数据,是指与科技信息相关的非数值型数据,最常见的一种科技大数据即科技文献数据。科技大数据情报分析的基本工作流程是根据特定需要进行的情报搜集和信息整序工作,透过现象,揭示具体领域科技大数据所蕴含的数据特征、规律和关联等信息,实现"源于科技大数据,高于科技大数据"的分析结果。因此,科技大数据与科技情报工作生活密切相关,由于互联网技术的发展,科技数据处理、智能数据分析等具有海量需求的应用变得越来越普遍[1-3],面对日益巨大的数据量,无论从形式还是内容上,均已无法用传统的方式进行采集、存储、操作、管理和分析。所以,无论是从事科技情报分析的专家学者,还是从事科学技术研究的科技人员,面对庞大科技数据集都会运用某种技术或方法去简化数据处理的过程,变革传统的数据管理和处理技术或方法

收稿日期:2018-05-01

基金项目:国家社科基金项目"基于事实型科技大数据的情报分析方法及集成分析平台研究"(14BTQ038)

作者简介:曾　文(1973-),女,博士,副研究员,研究生导师,主要从事科技情报分析技术、情报理论与方法研究.

亟待解决。因此,无论是从科学研究还是从应用的角度看,科技大数据的情报技术研究已经成为科技信息发展的自然延伸。

2 相关研究现状

情报分析研究历经科技文献的研究(文献检索和编译报道)、专业研究(分析和综合)、综合研究等三个主要阶段,间接或直接地帮助解决科学技术问题。情报分析方法包括逻辑学方法、数学方法、系统分析方法、社会学方法、情报学方法和经济学方法等。其中,情报学方法是基于信息的序化和转化。最初数据库技术,信息检索技术、信息计量学技术成为情报研究的主要技术,随着计算机等学科技术的发展,文献调查、内容分析、引文分析、专利分析、文本挖掘、数据挖掘等逐步形成情报分析技术的重要内容。科技大数据与大数据一样,同样呈现出"4V1O"的特征,即数据量大、多样化、数据价值密度化、速度快、时效高和数据在线的数据特征。此外,科技大数据具有敏感性和积累性,会涉及到国家安全和利益。因此,科技大数据的处理和数据分析与其它类型大数据相比,更具有一定的复杂性。

目前国内外对科技大数据研究以领域性和专业性科技数据的监测和研究较多,例如:美国德克赛尔大学采用基于突发词的科技动态监测模型和相关算法,辨识和探测学科知识领域的研究热点,预测知识领域发展的前沿趋势。美国科技信息研究所研制的数据分析工具Result Analysis,使用户能够对检索获得的文献数据进行分析,得到文章的年代分布、引文分布、引文频次等信息,帮助用户掌握某一领域科技研究的热点。Baumann以地球和环境科学数据为例,提出地球大数据分析系统——earthserver,它提供覆盖该类型数据集的一个解决方案,并建立一个高性能阵列数据库技术,并通过标准的服务交互,提供全面的查询语言和可扩展到移动访问和可视化分析技术[4]。在国内,以通过情报分析方法进行科技信息监测和知识获取为主。例如:清华大学针对科技文献开发的面向计算机领域的英文科技文献监测系统(ArnetMiner系统),以开放文献数据库DBLP、Citeseer等爬取的科技文献数据为基础,集成在Web上抽取的研究者Profile信息,构建学术社会网络,挖掘提供权威会议/专家/期刊发现、话题检测和关系路径发现等服务。北京理工大学提出以科技信息和数据分析为基础,进行科技信息的技术监测研究工作,利用信息处理技术,并集成专家智慧,对科技活动进行动态监测、分析和评估。中科院图书文献情报中心开展基于文献的知识发现相关理论、方法与应用研究,将基于相关文献的共词和共引理论、基于非相关文献的Swanson理论和基于全文文献的文本挖掘理论整合为知识发现的理论框架等。此外,国内学者构建科技人才科研综合能力评价模型,设计基于科技信息大数据的评价指标体系,对科技人才科研综合能力进行分析[5]。也有国内学者提出科技情报大数据业务平台的设计方案,但距离实用尚有距离[6]。

3 科技大数据的建模和分析

科技大数据的建模是一个为了解决科技数据分析流程问题的过程,科技数据分析目标是数据建模的核心。在科技大数据建模的过程中,合适的数据源是重点,对数据源进行预处理则是难点之一。科技大数据建模时,在数据预处理阶段需要找到适合数据预处理的处理方法。科学数据预处理是把待处理的科技数据转化为格式化的数据,使得数据分析技术更容易利用。数据预处理的方法基本有两种:一种是将科学数据转化为可以分析的数据格式,即无论何种分析方法都需要数据具有适合分析的数据形式;第二种是使科技数据含有与数据分析的问题有关的更多信息。数据分析者应针对具体数据使用合适的分析方法,而科技数据的分析和挖掘过程通常需要通过分析方法与领域知识的结合,来揭示科技数据隐藏的规则(模型)。事实上,科技大数据的建模和分析流程与传统的数据建模和分析流程并无太大差异,最大的不同是科技大数据需要更多地处理半结构化和非结构化数据。整个建模流程可以概括为:定义问题、数据理解、数据准备、模型建立、模型评估、模型更新与结果部署等。具体如图1所示。

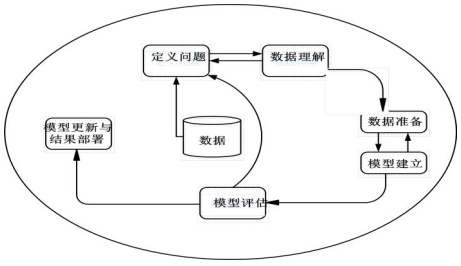

图1 科技大数据建模于分析流程图

4 科技大数据情报分析技术研究

4.1 需求分析

为了解我国从事情报分析、战略分析、行业咨询领域的工作人员,在工作过程中,对于科技数据、数据处理和分析工具在工作中的使用情况,了解相关工作人员的分析工作流程。因此,本文特别面向情报分析、战略分析、行业咨询领域的工作人员展开针对性的问卷调查。调查问卷显示,被调查者目前所从事的研究或服务的领域或行业分布较为分散。在一定意义上,可以视为能够代表研究总体,具有统计学意义。调查结果显示:①现有数据分析工具很难满足专业信息数据处理用户对分析工具的一般要求。期待分析工具具有按照信息数据分析的过程提供可视化的分析结果的功能的被调查者人数最多(56.0%);其余依次为关注分析工具具有处理大规模数据量的能力(49.3%);关注分析工具具有一定的数据加工功能和具有数据的统计功能(46.7%);关注分析工具分析结果的可信度(41.3%);关注分析工具具有数据采

集的功能(40.0%);关注分析工具具有多种数据分析方法的功能模块(36.0%);②用户对分析工具应具备的基本功能更多的倾向于按照信息数据分析的过程,提供可视化的分析结果的功能(41.3%),具有一定的数据加工功能(38.7%),以及具有数据的统计功能(38.7%)。用户对于提供研究报告、提供不同分析方法的分析结果、提供分析的结论或建议、提供分析图表等四种数据分析的结果形式选择频数基本一致。需要提供不同分析方法的分析结果的被调查者(62.7%)、需要提供研究报告的被调查者(56.0%)、需要提供分析图表的被调查者(54.7%)、需要提供分析的结论或建议的被调查者(50.7%)均过半数。需要技术方向或趋势分析和预测的被调查者人数最多,为60.0%;需要从信息数据中发现知识的被调查者,为42.7%;需要信息数据价值评估的被调查者,为41.3%;需要真假信息数据的识别的被调查者,为20.0%;需要分析方法的选择的被调查者人数最少,为16.0%。③用户希望数据分析工具帮助解决的问题较多,从数据格式的转换或归一、数据的分类、数据的抽取、数据的统计计算、数据的分析及从数据中获取有价值的信息等各个方面均有涉及。用户在工作中信息数据分析的基本工作流程相对一致,超过六成的被调查者在工作中信息数据分析的基本工作均包括需求分析、信息数据的采集、数据处理、数据的计算、数据的分析、得出结果或结论、分析结果的反馈等流程。排名前三的选项依次为数据的分析、数据的采集以及数据处理。

可见数据分析工具是影响用户信息数据处理的重要因素。但是,目前数据分析处理工具能力有限,很难满足专业信息数据处理用户对分析工具的一般要求。由于情报分析专业性相对较强,因此,在情报分析的过程中存在较多困难。此外,认为数据获取困难的被调查者人数最多。多数用户期待分析工具按照数据分析的过程,提供可视化的分析结果功能是分析工具应当具备的基本功能。所以,更加高效、便捷的情报分析技术和工具具有广阔的应用前景。

4.2 科技大数据情报分析平台的基本功能

科技情报分析的工作基本流程包括信源发现、信息获取、数据清洗、数据仓库建设和数据分析等五个方面。依据这个基本工作流程,我们可以研发科技大数据情报分析平台,平台的体系架构如图2所示。目前,该平台针对科技文献数据和科技政策两类科技大数据提供基本的分析功能,它可以辅助情报分析人员实现数据的导入、导出,数据分析。及分析结果展示等。该平台实现了作者提出的科技情报分析方法功能[7-9],并将相应的工具软件功能进行模块化的集成,以形成可用于科技数据情报分析的辅助工具,如图3所示。

科技大数据情报分析平台主要包括科技数据获取模块、科技数据管理模块、领域管理模块、科技数据分析模块和科技专利知识管理模块。科技数据获取模块主要实现网络科技数据获取;领域管理模块主要实现科技数据的分类管理,词表的导入导出功能,目的是实现科技数据的组织,科技术语的管理和存储;科技数据分析模块主要实现科技文献数据的关联分析、内容分析、重要性分析和主题演化路径分析等功能;科技专利知识管理模块主要实现科技专利知识的抽取功能。在科技数据的自动分析过程中,数据导入和分析均在后台运行,系统平台提供可视化的操作界面,用户选择上传数据后,系统可自动进行文件上传和启动数据分析过程,科技数据导入完成后,可进行科技数据全文的浏览和查看重要性的数据内容。

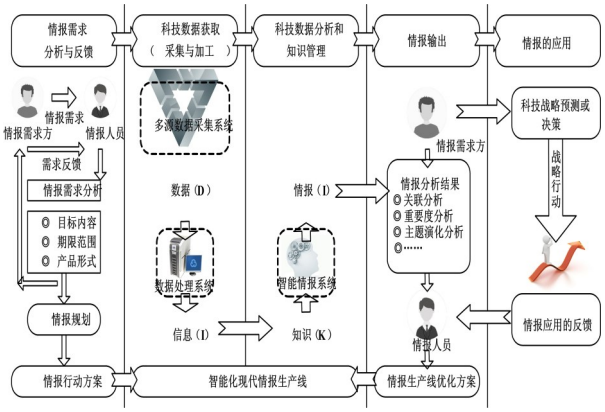

图 2 科技大数据情报分析平台的体系架构

图 3 科技大数据情报分析平台界面

5 结 语

大数据时代,数据正在成为一种生产资料,成为一种稀有资产和新兴产业。任何一个行业和领域都会产生有价值的数据,而对这些数据的统计、分析、挖掘和处理则会创造意想不到的价值和财富。事实上,这些价值和财富的载体是从大数据中获取的“知识”。对于自然人、企业和科研机构而言,如何获取蕴含在数据中的行业、领域“知识”是取得成功的关键之一。面对科技数据来源的日益多源化,科技数据规模的日益庞大,使用适用的科技数据与分析工具是必要的。科技大数据情报分析平台的研究工作仍处于起步阶段,虽然已具备基本的数据获取,处理,分析和结果展示的基本功能,但是距离实用化仍有距离,但是目前的研发成果可以说明,科技大数据情报分析平台可以一定程度辅助情报分析人员

进行数据的处理和分析,智能化的计算机处理和分析技术引入情报分析的研究过程是必要且实际的。我们相信:随着技术的不断进步,情报分析平台的功能会得到不断完善和优化,实现科技大数据情报分析平台的实用化并不是梦想。

参考文献

1 Boris Lublinsky,Kevin T.Smith,Alexey Yakubovich. Hadoop.高级编程:构建与实现大数据解决方案[M]. 穆玉伟,靳晓辉,译.北京:清华大学出版社,2014.
2 李 雯. 大数据时代[J]. 出版广角,2014,(17):39-41.
3 陈 明. 大数据可视化分析[J]. 计算机教育,2015,(5):94-97.
4 Baumann, Peter,Mazzetti, Paolo et al.Big Data Analytics for Earth Sciences: the EarthServer approach[J]. International journal of digital Earth, 2016, 9(1/3):3-29.
5 王运红, 潘云涛,赵筱媛.基于科技信息大数据的科技人才科研综合能力评价及应用研究[J].医学信息学杂志, 2017, 38 (12) :7-13.
6 吴素研,吕志坚, 吴江瑞, 李文波.科技情报大数据业务平台设计[J].现代情报,2018,38(1):131-135.
7 曾 文.科技文献术语的自动抽取技术研究与分析[J]. 现代图书情报技术,2014,(1):51-55.
8 Wen Zeng.The exploration of information extraction and analysis about science and technology policy in China[J].The Electronic Library,2017,35(4):709-723.
9 Wen Zeng. Term extraction and correlation analysis based on massive scientific and technical literature[J]. Int. J. Computational Science and Engineering,2017, 15(4):248-255.

(责任编辑:张连峰)

(上接第92页)

10 陈巧灵,廖祥文,魏晶晶,等.基于DOM树层次特征的多记录网页抽取[J].模式识别与人工智能,2015,28(2):125-131.
11 张儒清,郭 岩,刘 悦,俞晓明,程学旗.任意网页的主题信息抽取研究[J].中文信息学报,2017,31(5):127-137.
12 陈 雪,梁永全,赵相彬.改进的基于本体的Web信息抽取[J].计算机应用与软件,2013,30(7):14-16,42.
13 双 哲,孙 蕾.基于改进的隐马尔可夫模型在网页信息抽取中的研究与应用[J].计算机应用与软件,2017,34(2):42-47.
14 王 辉,郁 波,洪 宇,肖仰华.基于知识图谱的Web信息抽取系统[J].计算机工程,2017,43(6):118-124.
15 王海艳,曹 攀.基于节点属性与正文内容的海量Web信息抽取方法[J].通信学报,2016,37(10):9-17.
16 孙 璐,陈军华,廉德胜.一种基于视觉特征的Deep Web信息抽取方法[J].计算机与数字工程,2016,44(6):1107-1111,1126.
17 袁鸿雁.基于本体的Web表格信息抽取技术的研究[J].青岛大学学报(自然科学版),2010,23(2):47-51.
18 师雪霖,程文涛.Web信息抽取与语义检索框架[J].郑州大学学报(理学版),2010,42(1):29-32.
19 Deng Cai, Shipeng Yu, Ji-Rong Wen and Wei-Ying Ma. VIPS: a Vision-based Page Segmentation Algorithm. Microsoft Technical Report, MSR-TR-2003-79[R].S.L.:s.n ,2003.
20 Liu Bing,GROSSMAN R,ZHAI Yan-hong. Mining data records in Web pages[C]//Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining:New York:ACM Press,2003:601-606.
21 Reis D C, Golgher P B, Silva A S, et al. Automatic Web News Extraction Using Tree Edit Distance[C]//Proceedings of the 13th International Conference on WorldWideWeb:NewYork:ACM,2004:502-511.
22 Crescenzi V, Mecca G, Merialdo P. RoadRunner: Towards Auto-matic Data Extraction from LargeWeb Sites[C]//Proceedings of the27th International Conference on VeryLargeData Base:Sa Francisco:Morgan Kaufman Publishers Inc, 2001: 109-118.
23 Sun F, Song D, Liao l. DOM based content extraction via text density[C]//Proceedings of the 34th Annual ACM SIGIR Conference:Beijing:ACM Press,2011:245-254.
24 Wang J B, Wang L Z, Gao W L, et al. Chinese Web content extraction based on native bayes model[C]//Proceedings of International Federation for Information Processing: Trondheim:IFIP Press,2014:404-413.
25 Kristin S S,Dattatraya J S. Schema inference and data extraction from templatized Web pages[C]//Proceedings of ACM-ICPC:St. Petersburg:ACM Press,2015:1-6.
26 聂 卉,黄贵鹏.树编辑距离在Web信息抽取中的应用实现[J].现代图书情报技术, 2010,(5):29-34.
27 刘守群,朱 明,谭晓彬.一种基于树匹配的网页语义块挖掘算法[J].小型微型计算机系统, 2009, 30(8): 1541-1545.
28 李志义,沈之锐.基于自然标注的网页信息抽取研究[J].情报学报,2013,32(8):853-859.

(责任编辑:张连峰)