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进展报告

中国科学技术信息研究所

2016-12-20

目　录

图清单

表清单

1 引言

对于“大数据”（Big data）研究机构 Gartner 给出了这样的定义。“大数据”是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力来适应海量、高增长率和多样化的信息资产。麦肯锡全球研究所给出的定义是：一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。

事实型科技数据是指长期积累形成的与科技创新全过程相关的各类科技信息，涵盖了客观描述科技创新决策和具体的科技创新活动全过程的各类科技信息。它主要包括两大类数据，第一类是客观的科研产出和技术产出数据：包括科技期刊文献数据、专利数据、学位论文数据，科技报告数据。技术成果和标准，政府和企业的研发投入数据、相关科技档案和国内外各领域的科技进展资料、具体研究案例、科技基础设施，研发机构和研发力量等，这类数据来源较固定，数据格式较规范，呈结构化或半结构化；第二类是互联网络数据：包括各级组织、科研机构，企业发布的科技政策、新闻等网页信息，科研个体的个人学术网站、微博，以及科研论坛等产生的动态、实时和交互式网络数据，这类数据较离散，数据格式规范性差，呈非结构化。在大数据环境下，科技数据来源分布广泛、数据质量良莠不齐，数据内容深度千差万别。事实型科技大数据呈现的特点是：数据量大且增长速度快；数据结构类型多；有价值的数据比例小；数据具有敏感性和积累性，会涉及到国家安全和利益；数据管理和数据分析具有复杂性。因此，事实型科技大数据在提供大量数据的同时，也同时增加了获取有价值科技情报的难度，使已有情报分析方法难以有效挖掘数据的价值，已有情报分析工具的局限性日益显著，特别需要采用更有效的情报分析方法和工具来解决这些问题。本课题今年的研究重点是互联网的网络数据，重点以各级组织、科研机构，企业发布的科技政策为研究对象。分别在以下几个方面取得一定的研究成果：1）网络数据获取软件的研发；2）科技术语获取技术研究与工具实现；3）科技政策术语的特殊性及自动获取技术；4）科技政策内容分析方法研究；5) 深度学习；6）基