

服务于科技大数据情报分析的方法及工具研究

曾文车尧 张运良 徐红姣

【摘要】【目的/意义】计算机技术的发展给情报学,特别是情报分析的方法带来了冲击和变化,传统的情报分析方法面对时代的发展和技术进步的事实,需要新方法和新技术充实到科技情报分析的工作中,情报分析的技术和工具需要更具实用性和应用性。【方法/过程】本文以科技大数据为视角和分析对象,对国内外的科技情报分析方法,工具及存在的问题进行分析,并提出面向科技大数据情报分析服务的方法,阐述方法的设计及工具研发过程。【结果/结论】本文以科技政策数据为实例,介绍科技政策数据分析的方法,设计并构建科技政策数据分析工具实现数据分析过程,验证了本文提出方法的有效性。

【关键词】科技大数据;情报分析;术语抽取;科技数据分析;情报分析工具

【作者简介】曾文(通讯作者)(1973-),女,博士,中国科学技术信息研究所副研究员,主要从事科技情报分析技术、情报理论与方法研究(北京 100038);车尧,《情报学报》编辑部(北京 100045);张运良,徐红姣,中国科学技术信息研究所(北京 100038)。

【原文出处】《情报科学》(长春),2019.4.92~96

【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);中国工程科技知识中心建设项目“知识组织体系建设”的研究成果之一。

1 引言

我们生活在一个充满“数据”的时代,“数据”几乎已经渗透到人类工作和生活的各个方面,人类不仅是“数据”的使用者,同时也是“数据”的生产者。根据IBM的研究,人类每天生产大约2.5亿字节的数据。这些数据源自不同的数据源,例如:社交媒体网站及用户的应用,智能手机的使用活动,线上线下的经济交易活动,数字图片和视频数据的传输等。全球知名的咨询公司麦肯锡在2011年最早提出“大数据”的概念。简言之,大数据就是通过已有的技术和工具,在可接受的时间范围内处理和分析数据集。大数据在发展体量(volume)、多样性(variety)、价值密度(value)及处理速度(velocity)方面都具有突出的特点,被称为大数据的4V特性。由于互联网技术的发展,科技数据处理、商业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,均需要改变传统方法,运用新方法进行采集、存储、操作、管理和分析。当人们认识到数据的价值,分析大数据就成为工作的扩展和延伸,而逐渐上升至情报分析。当数据源是科技数据时,科技大数据的情报分析则应运而生。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,也称为事实型科技大数据。具体地说:科技大数据是指长期积累形成的与科技创新全过程相关的各类非数值型科技信息,它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,最常见的一种科技大数据即科技文献数据。科技大数据主要包括两类数据,一类是客观的科研产出和技术产出数据,例如:科技期刊文献数据、专利数据、学位论文数据和科技

报告、技术标准数据等,这类数据相对较为集中,数据格式较为规范,呈结构化或半结构化的特点;二类是各级组织、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、实时和交互式网络事实型数据,这类数据较为离散,数据格式规范性差,主要呈非结构化的特点。目前无论是从事情报分析的专家学者,还是从事科学技术研究的科技人员,面对日益增长和积累的庞大数据集都会期待运用某种手段或方法以发现有价值的情报信息。和大数据一样,科技大数据多是呈现结构化、半结构化或非结构化的数据结构和状态,处理起来较为烦琐且需要较长时间,传统的数据管理和处理方法已经难以解决这一问题。因此,无论是从科学研究还是从应用的角度看,针对服务于科技大数据情报分析的方法及工具研究已经成为科技信息发展的自然延伸。

2 相关研究现状

2.1 科技情报分析的方法研究

传统的科技情报分析方法有联合分析法、内容分析法、技术路线图法、德尔菲法、情景分析法等。目前,情报研究的理论和实践水平在自动化、智能化、可视化等技术发展趋势的影响下,网络分析、模型/模拟、文献计量分析、专利地图分析、Web文本挖掘、数据挖掘和知识发现、技术前瞻等方法也融入科技情报分析方法中。在国外已经开展基于科技文献和互联网信息,研究和开发文献监测模型、监测算法和工具的研究工作。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。在国内有关科学技术领域的情报分析研究自1978年以来一直占据各领域情报分析研究的首位,内容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,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。

2.2 情报分析的工具研究

国内外针对科技文献数据的分析工具研究占主流,主要采用数据挖掘技术和文献计量学的方法。国外情报分析工具的典型代表是:①专利分析工具:美国Thomson公司Thomson Data Analyzer(TDA)软件,既能对专利数据进行深度挖掘并展开可视化分析,也可用于分析英文论文;②专利数据、专利引文分析平台:Thomson集团Aureka信息平台,它可提供对专利数据和专利引文进行分析,揭示专利信息间的关联;③知识发现和知识管理平台:ISI Web of Knowledge数字研究平台,蕴含知识发现和管理功能,提供知识发现与管理工具,基于内容与引文的跨库交叉浏览、检索结果的信息分析、检索结果的信息管理等;④基于文献计量特征的可视化系统:美国爵士大学开发的AuthorLink、ConceptLink、PNASLink等三种基于引文分析理论的信息检索可视化系统,提供引文分析的可视化。

在国内,情报分析工具的典型代表是:①北京理工大学采用基于文献的数据挖掘方法来进行科技监测,并开发了情报分析软件,提供文献关联分析以及结果的自动生成;②清华大学针对科技文献开发面向计算机领域的英文科技文献监测系统(ArnetMiner系统),它以开放文献数据库DBLP、CiteSeer等爬取的文献数据为基础,集成在Web上抽取研究者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。

2.3 存在的主要问题

已有研究主要问题集中在以下五个方面:①国内的传统情报分析方法研究与事实型科技数据联系不够全面和充分、缺乏对事实型科技大数据综合分析方法的研究工作;②国内研究偏重于结构化的实验性科技文献数据分析研究,缺乏对无规则性、非结构化科技事实型数据与规范性、结构化科技文献数

据的融合分析研究工作;③现有情报分析工具强调定向数据源,非定向数据源无法做分析,例如,TDA只支持WoK-DII、WoK-INSPEC、WoK-WoS、Aureka-Patent、Delphion-DWPI、Delphion-Patent的数据格式,对其他格式数据不支持分析;④分析工具的语言局限性,已有的数据分析工具均支持英文文献,而对中文文献的数据分析功能不够,缺乏中外科技数据的分析比较功能;⑤现有工具对事实型数据源支持不足,例如,使用范围较广的工具均是针对某一类事实性数据源,如TDA支持专利文献或论文文献,清华大学的ArnetMiner系统支持爬取的数据库文献数据。

3 服务于科技大数据情报分析的方法研究

科技大数据呈现的主要特点除了数据量大且增长速度快;数据来源和数据结构类型多;有价值的信息相对比例小之外,科技大数据具有敏感性和积累性,会涉及国家安全和利益。因此,科技大数据的处理和数据分析与其他类型大数据相比,更具有一定的复杂性。如数据从被动采集变成主动采集,数据处理更注重实时计算,通过历史数据进行分析和预测。科技情报分析方法和工具存在的问题导致科技情报分析结果尚有局限性,从而定量和定性分析准确性难以保证。所以,也造成数据分析工具的用户群体有限,难以应用推广,使科技用户受益。作为科技情报分析研究和工作的重要数据源,情报分析人员利用情报分析方法,实现从科技大数据的数据内容中分析出有“价值”的情报信息是科技情报分析的目的之一。科技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。

3.1 科技数据的术语抽取方法

科技术语可以表征科技概念,表达科技数据(非数值的科技数据)的核心内容,是科技数据情报分析的重要内容之一。各种类型的科技数据的术语抽取在基本方法上基本雷同,但在具体细节上会有不同。科技数据术语抽取的基本流程是:①科技数据的预处理。常用的方法是科技数据的格式化处理、数据的分词和词性标注、去噪等;②采用相关方法从

科技数据中获取术语。常用的方法是语言规则的方法、统计的方法、语言规则和统计相结合的方法等^[4-6]。国内外已有的术语获取方法很多且各有优劣,术语的自动抽取是科技大数据处理和分析的基础。为此,本文设计并研发融合多种术语抽取方法的中文科技术语抽取方法。

图1显示了基于规则和统计相结合的中文科技术语抽取方法的设计架构。该方法所需数据集是任一领域的中文文档集、少量该领域现有的中文术语、参照领域中文文档词频统计结果。其关键技术由文档预处理、基于规则的术语抽取和基于统计的术语抽取三部分组成,其中文档预处理部分包括了中文分词及词性标准、去停用词以及将长句切分为短句三个步骤;基于规则的短语获取模块利用领域术语样例,从中提取规则模板,然后按照规则模板从文档集合中获取候选术语短语。本文提出的中文科技术语抽取方法融合七种算法:TF^[7]、TF-IDF^[8]、CValue^[9]、Entropy^[10]、GlossEx^[11]、TermEx^[12]和Weirdness^[14],其中后三种算法需要用到参照领域词频统计结果,采用归一化和简单的投票方法将七种算法的结果进行合并,得到最终的科技术语抽取结果。

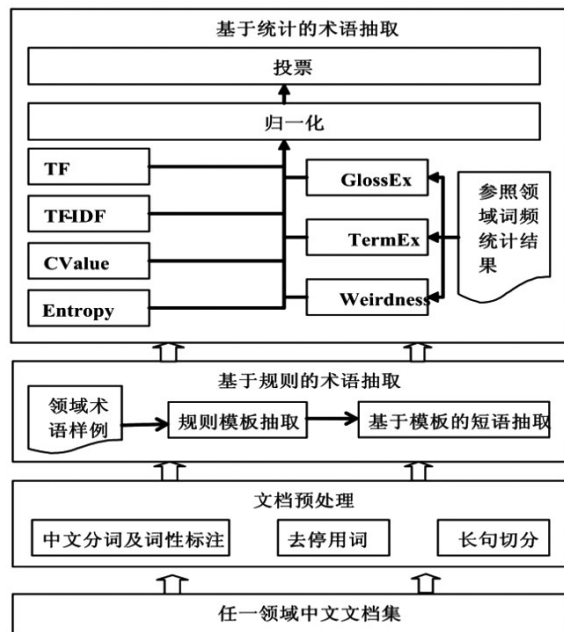


图1 中文科技术语自动抽取方法设计架构图

3.2 基于科技术语的科技数据分析方法

科技数据分析是指利用计算机处理技术自动地

从科技文本中提取简练且有代表性的语句,识别出文本的核心内容或用户感兴趣的重要语句内容。而语句权重的计算是判断重要语句的依据。其计算上,应重点使用特征组合方法。特征组合的方法是将数据中语句的多个特征按一定方式组合,然后根据对每个特征的计算和特征组合后计算值抽取相对重要的语句。常使用的特征包括:词频、与标题的相似度、句子位置、线索词等。本文使用的特征是词频、语句与标题的相似度、语句的技术强度等。

(1)基于词频的科技数据内容权重计算。

词频越大则代表词的重要性越高,那么包含重要性高的词越多的语句的重要性越大,则更有可能称为重要句。该方法是将词频作为词语的权重,再根据词语的权重来计算语句的权重。其假设文本是语句的线性排列,语句是词的线性排列,若一个语句中权重高的词汇越多,那么其包含的信息量就越大,因此这个语句就越重要。基于该假设,本文做如下改进:首先基于词频的科技数据内容权重计算不是衡量语句中的每个词的权重,而是结合科技术语和停用词表处理和衡量语句中每个术语的权重,这是由于语句中不仅包含有实际含义的词语,也包含一些停用词、无意义虚词,而术语是本领域中专业概念的集合,理论上语句中的术语包含了语句的主要思想。因此,本文在计算语句权重时,考虑语句中每个术语的权重,既可以简化计算又能提高计算准确率。计算方法如下:

$$\text{sentence}_i = w(t_j) / \sum_term(\text{sentence}_i) \quad (1)$$

其中, $\sum_term(\text{sentence}_i)$ 表示科技数据的每个语句含有的术语总数; $w(t_j)$ 表示语句中术语的 tf-idf 值。

(2)基于标题相似度的科技数据内容权重计算。

科技数据文本的标题是一个很重要的信息,标题通常与文章的中心内容相关性很大。本文通过语句与标题的相似度计算来进一步衡量语句的权重,语句与标题的相似度的计算采用基于向量空间模型的相似度计算方法。本文利用术语形成的词典对每个语句和标题进行分词,把语句和标题用其含有的术语即特征项的向量表示出来,将每个语句与其对应的标题用特征项的向量表示,计算语句与

其对应的标题的向量间的夹角,向量夹角越小,标题与语句的相似度越高。语句与标题的相似度的计算公式如下:

$$\text{Similarity}(\text{sentence}_i, \text{title}_i) = \frac{\text{sentence}_i \cdot \text{title}_i}{\|\text{sentence}_i\| \cdot \|\text{title}_i\|} \quad (2)$$

其中, sentence_i 表示科技数据的语句向量, title_i 表示科技数据的标题向量。

(3)基于技术强度的科技数据内容权重计算。

科技数据内容不可避免地会涉及技术术语。本文假设:如果在某个科技数据的语句中包含有科技术语,则这个含有科技术语的语句,我们认为它相对其他不包含科技术语的语句是更重要的。依据汉语科技词系统(中国科学技术信息研究所组织编撰)的内容,本文设计如下基于技术强度的科技数据内容权重计算方法。

$$\text{Technical_strength}(\text{sentence}_i) = \text{weight_value}_i \quad (3)$$

如果科技数据的语句中含有核心科技术语,则该语句的技术强度权值为 0.8;如果科技数据的语句中含有非核心科技术语,则该语句的技术强度权值为 0.5;如果科技数据的语句中含有相关科技术语,则该语句的技术强度权值为 0.3。对于其他情况,该语句的技术强度权值为 0.1。

综上,本文提出的科技政策数据内容的重要性计算公式如下:

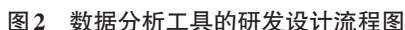
$$(\text{importance_sentence}_i) = (\text{sentence}_i) + \text{Similarity}(\text{sentence}_i, \text{title}_i) + \text{Technical_strength}(\text{sentence}_i) \quad (4)$$

4 服务于科技大数据情报分析的工具研发与实现

基于前文所述,我们研发相应的软件工具进行方法的实现,以科技政策数据为例,介绍数据分析及其实现过程。工具研发的关键技术是对科技政策数据内容进行语义分析,实现每篇科技政策数据内容的语句权重计算。权重计算的主要内容涉及基于词频的科技数据内容权重计算、基于标题相似度的科技数据内容权重计算、基于技术强度的科技数据内容权重计算等三个方面,工具研发的设计流程如图 2 所示。

软件工具的具体数据分析过程可描述为:

(1)用户启动软件工具,进入软件工具操作状态,输入待分析文本路径和分析结果存储路径后,软件



(3)软件工具进入科技政策数据内容分析计算的过程,实现文本内容的三次权重计算。

(5)任务结束后,软件工具自动终止其工作状态。

数据正在成为一种生产资料,成为一种稀有资源和新兴产业。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,面对科技数据来源的日益多元化,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手段^[13]。本文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,服务于科技大数据情报分析的方法及工具在一定程度可以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。

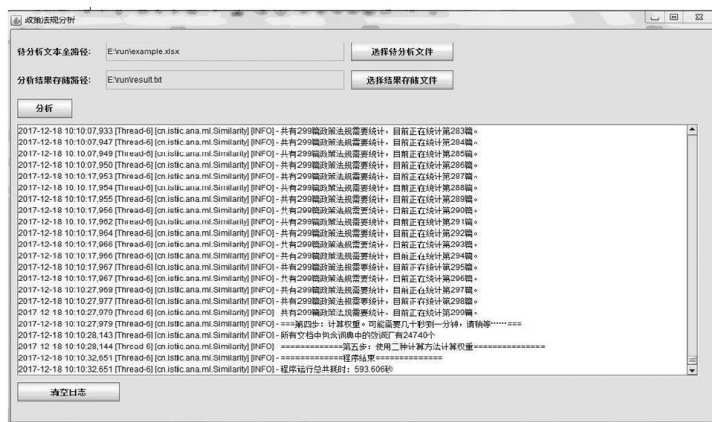


图3 科技政策数据自动导入和分析界面

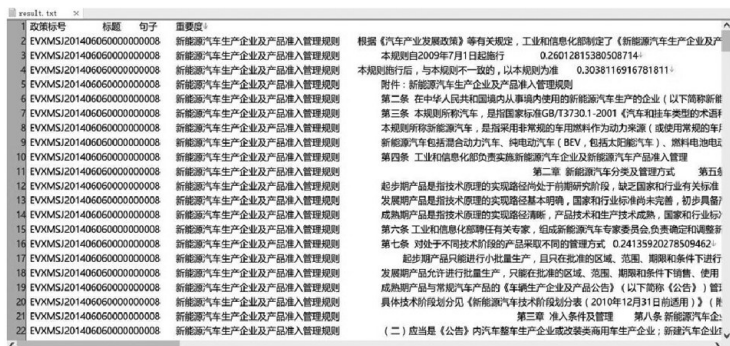


图4 科技政策数据分析完成

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] 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.

[5] 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.

[6] 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.

[7] LOSSIO-VENTURA J A, JONQUET C, ROCHE M, et al. Biomedical term extraction: overview and a new methodology[J]. Information Retrieval Journal, 2016, 19(1): 59-99.

[8] CHURCH K W, GALE W A. Inverse Document Fre-

quency (IDF): A Measure of Deviations from Poisson[M]. Natural Language Processing Using Large Corpora, 1999: 283-295.

[9] FRANTZI K, ANANIADOU S, MIMA H. Automatic recognition of multi-word terms: the C-value/NC-value method [J]. International Journal on Digital Libraries, 2000, 3(2): 115-130.

[10] ASTRAKHANTSEV N A, FEDORENKO D G, TURDAKOV D Y. Methods for Automatic Term Recognition in Domain-Specific Text Collections: A Survey[J]. Programming and Computer Software, 2015, 41(6): 336-349.

[11] KOZAKOV L, PARK Y, FIN T, et al. Glossary extraction and utilization in the information search and delivery system for IBM Technical Support[J]. IBM Systems Journal, 2004, 43(3): 546-563.

[12] SCLANO F, VELARDI P. TermExtractor: a Web Application to Learn the Shared Terminology of Emergent Web Communities[J]. Enterprise Interoperability II, 2007: 287-290.

[13] 孙晓平. 大数据知识计算的挑战[J]. 情报工程, 2015, 1(6): 43-50.

Methods and Tools for Serving Intelligence Analysis of Science and Technology Big Data

Zeng Wen Che Yao Zhang Yunliang Xu Hongjiao

Abstract: [Purpose/significance] The development of computer technology for information science, especially the method of intelligence analysis has brought serious impact and changed the traditional intelligence analysis method. Facing the development of the times and technological progress, it is necessary that new methods and technologies are used to enrich the work of intelligence information. Intelligence analysis techniques and tools need to be more practical and the applicable. [Method/process] From the perspective of science and technology of big data, the paper analyzed the methods, tools of scientific and technological information analysis and their problems at home and abroad. It put forward the methods of information analysis for big data, and expounded the design and development process of tools. [Result/conclusion] The paper took the data of science and technology policy as an example, introduced the method of data analysis, designed and constructed a tool to realize the process of data analysis and verifies the effectiveness of the method proposed in this paper.

Key words: Science and technology big data; Intelligence analysis; Terminology extraction; Science and technology data analysis; Intelligence analysis tools