【资源·共享】

科技文献信息抽取方法浅析*

●敖　龙　谢海先　（深圳职业技术学院　广东深圳　518055）

[摘　要]文章在Web of Science等影响力较大的国际数据库中检索内容与“科技文献”和“信息抽取”相关的文献，经设定条件筛选后获得63篇相关文献。回顾相关文献，从抽取的信息与抽取的方法两个角度进行分类与分析，总结该领域已有的研究成果和存在的不足。从科技文献中抽取的信息主要为结构化信息、显式信息和隐式信息，最新最先进的抽取方法主要集中在机器学习、自然语言处理和统计学中。语义信息的抽取有一定的进步空间及挑战性，灵活结合机器学习和自然语言处理方法是处理此领域问题的未来趋势。

[关键词]信息抽取　科技文献　语义信息　机器学习　自然语言处理

[中图法分类号]G253　　[文献标识码]A　　[文章编号]1003－7845(2022)02－0024－04

[引用本文格式]敖　龙，谢海先.科技文献信息抽取方法浅析[J].高校图书馆工作，2022(2)：24－27

1　引言

科技文献记录了最新的科学研究进展和成果，在科研工作中发挥着至关重要的作用。随着自然科学、技术、管理、人文、社会科学等学科研究的迅速发展，越来越多的科技文献形成了庞大的研究信息群体，提供了丰富的原始研究工作信息，是研究人员交流最新知识的载体。学者们需要捕捉其中的重点，对它们进行有效的检索，找到相似或相关的文献并迅速获得它们的具体内容。因此，有必要开展信息抽取工作。

信息抽取是指从特定领域的结构化和非结构化数据中提取具有特定目标、有意义的知识，它为获取科技文献中的信息内容提供了新的途径。从科技文献中抽取信息的主要任务是对文献内容进行结构化处理，将其转化为满足用户需求并能被用户有效利用的信息。简而言之，这是一个将各种原始科技文献转化为特定格式的、独特的信息的过程。

由于对科技文献处理的需求多种多样，信息抽取对于读者检索、分析和总结科技文献具有重要意义。总结这一领域的现状和进展，有助于读者对信息抽取如何辅助科技文献分析有一个清晰的概念，并且了解最先进的抽取方法以及这一研究领域的发展方向。

2　研究目的和方法

本文的目的是通过系统地收集不同维度和层面的知识，调查这一领域的研究进展并提供这一领域的概况，同时归纳信息抽取的方法，以帮助读者深入了解这一领域。

本文所研究的科技文献，主要是指发表在学术期刊、会议论文、毕业论文、书籍、技术报告和专利论文中的文章。所综述的文献范围有：在Web of Science、Elsevier Science Direct、PubMed、ACM Digital Library和IEEE Xplore digital library等国际数据库中检索主题、标题或关键词中包含“information extraction”（信息抽取）、“literature”（文献）、“scientific literature”（科技文献）、“paper”（论文）、“article”（文章）、“publication”（出版物）和“extract”（抽取）等的文献；发表时间为2013年1月1日至2019年12月31日。综合考虑期刊或会议质量和影响力、文献本身质量和影响力、文献与本研究相关性等因素进行筛选，最终获得的文献数量为63篇。

通过对文献的阅读和分析，本文设定了两个方面的研究内容，具体表述如下。

（1）在现有的研究中，从科技文献中抽取的主要信息是什么？关于此项问题，本文提出了一个三层信息架构，即将科技文献中的信息分为不同的类别：结构化信息、显式信息和隐式信息。

（2）从科技文献中抽取信息的最新最先进的方法是什么？本文对近年来有关抽取方法的文献进行研究，发现统计学、自然语言处理和机器学习方法是

*　本文系深圳市哲学社会科学“十三五”规划课题“深圳智慧图书馆联盟设计研究”（SZ2018B030）研究成果之一。

应用比较广泛的方法,其中机器学习方法是应用最广泛的方法。

3　抽取信息的内容

本文定义的科技文献中的第一层信息是指它们中的结构化数据,也称为结构化信息。科技文献中的每一章节都明确地展示了特定的信息。在科技文献中的每一部分,文本、图像、图形和表格是基本的信息符号,在特定的布局中展现了文章中所有的内容。它们不包含任何语义信息,提取过程只需要识别它们的标签即可。

在第一层信息的基础上,抽取的是科技文献的显式信息,如标题、摘要、引言、前人研究、方法、数据、实验、结果、讨论和结论等。这些都是文献不同部分的具体信息,在一定程度上表明了不同的含义,可以被用来对文献进行索引和分析。此外,它们还为更高级别的信息抽取提供原始数据源。最高层的是概念信息,也称为隐式信息,这种信息的抽取相当于文献中语义信息的抽取和整合。不同层次的信息类别包含不同的内容,如表 1 所示。

表 1　科技文献中的信息类别

类别	定义	具体信息内容
结构化信息	信息符号:在文章结构中嵌入的可见内容。	文本,图表,图像。
显式信息	从结构化信息中抽取,通常使用不同形式的符号描述特定的内容。	标题,作者,关键词,元数据,引文,参考文献,文本内容,图表内容,图像内容。
隐式信息	概念层面的信息,科技文献中内容的深层语义表达。	主题,新颖发现,实体,实体关系,论述论证,知识,情感分析。

3.1　结构化信息

结构化信息可以是文本、图像、图形和表格等。结构化信息的抽取只需要识别和区分文献不同部分的具体信息,不涉及信息的含义。可扩展标记语言(XML)是存储和呈现科技文献的常见格式,XML 使得文献可以在万维网上直接被阅读。XML 中的结构识别也属于结构化信息抽取。与 PDF 相比,XML 更清晰地描述了文献的逻辑结构,并且能够呈现 PDF 的几何排版标记、字体和布局等。因此,XML 中结构化信息的抽取,本质上是标记和标签的识别[1]。

在大多数情况下,结构化信息的定位和识别是显式信息抽取的基础。以往许多研究将结构化信息的抽取和显式信息的抽取结合在一起。

3.2　显式信息

显式信息的抽取包括标题、作者、关键词、参考文献的抽取等。科技文献的标题是文章的观点、发现和贡献的浓缩,一些标题甚至包含了整篇文献的结论。标题也是科技文献的一种代表性概念,标题的抽取为文献综述奠定了基础。作者信息是抽取工作的另一个焦点。同一作者的文献在某些方面往往是相互联系的。作者信息抽取主要指抽取姓名、机构、国家、资助机构和项目。科技文献作者的隶属关系是重要的元数据之一,它可以帮助自动处理和分析出版物记录[2]。此外,作者的隶属关系有助于作者的识别和姓名消歧。关键词抽取要能够自动识别重要的、具有代表性的主题术语或概念,描述并总结文献内容。关键词抽取有助于科技文献的处理和检索,被证明是辅助数据挖掘的有效方法。它也是信息检索和自然语言处理的关键要素,例如科技文献分类、总结、推荐和聚类[3]。参考文献和引文抽取通常对科技文献中的内容和书目部分进行抽取、分段和解析,从而获得一系列的组成部分,如作者、标题、年份、期刊名称、会场类型、会议地点、地点、卷、页以及引文主题和内容等。从科技文献中抽取参考文献信息有三个步骤:第一步是参考文献段的检测;第二步是参考文献的分割;最后是对每个信息字符串(如作者和标题)的注释[4-5]。

3.3　隐式信息

在科技文献内容层次信息的基础上,还可以抽取更高层次的概念信息。科技文献中包含科学陈述、新颖发现和科学知识。其中,科学知识包括事实、概念、假设、猜测、观点和预测。科技文献中的关键概念往往涉及主体思想、技术和应用等,它们有助于将科技文献的贡献描述得更加清晰。在本研究中,隐式信息包括主题、新颖发现、知识、论证、情感等诸多方面,代表了科技文献中深层的语义信息。

科技文献中的知识来源于概念、内容词、实体和实体关系。知识抽取往往需要探索语义信息。以生物医学文献为例,从中抽取的知识信息有两类:既有知识和新兴知识,新兴知识往往与某一特定领域的新发现或新观点有关[6]。新的科学假说在解决研

究问题方面发挥着重要作用,它们也可以从科技文献的结构内容中抽取和生成。推测信息通常出现在包含实验性质的科技文献中,它是基于实验证据的假设表达,也为未来的研究提供了发展空间。论证是形成知识的关键过程,是科技文献中的必要内容。论证由论点和论证关系组成,而每个论点又由几个关键部分组成。论证信息抽取需要自动识别和鉴定前提、结论和论点之间的关系[7]。实体和实体关系的抽取主要是针对生物医学、化学等领域的文献。过去的研究探讨了生物医学实体与实体之间的关系,如基因表达关系、疾病—突变关系、药物—疾病关系等[8]。化学文献中的信息抽取也包含命名实体识别和关系提取,其中化学药物与疾病的关系是典型的被抽取信息。在隐式信息方面,虽然已经在实体和实体关系抽取上获取了大量信息,但长短句和符号、部分、整体和琐碎实体是目前实体识别中的一些挑战。此外,描述某一特定领域概念的术语,也是另一类需要抽取的高层次隐式信息。除上述内容外,科技文献的隐式信息抽取还包括事件抽取、情感抽取等。

4 抽取信息的方法

从科技文献中抽取信息最常见的挑战是准确性、覆盖率和可扩展性。根据信息类型和抽取需求,支持科技文献信息抽取的方法可以分为三个不同的大类:统计学、自然语言处理和机器学习。

(1)统计学:统计学方法是科技文献中信息抽取的最基本方法。通常以词为最小单位进行抽取处理,如词频计算、词频—逆向文献频次计算等。近年来,统计学方法在这一领域已很少单独使用,通常是与自然语言处理方法和机器学习方法结合使用。以下方法可以归类到统计学方法中:网页排名、单词/短语频次、条件随机场、词频—逆文本频率指数。

(2)自然语言处理:自然语言处理是一种基于统计方法与人工智能相结合的方法。一般来说,在所综述的文献中,从科技文献中抽取信息的自然语言处理方法包括斯坦福自然语言处理解析器、Python自然语言处理工具集(NLTK)、词性、词嵌入、本体论与词汇模式、命名实体识别、语义演算、文档向量、基于本体、分词、词干提取、词形还原、基于词汇、依存关系等。

(3)机器学习:在先前的研究中,机器学习已被用来做文本挖掘、文本分类和数据挖掘,以识别包含不同类型信息的文本模块。机器学习方法可以分为监督学习方法、半监督学习方法和无监督学习方法。换句话说,机器学习方法可以分为用标记数据处理、非标记数据处理以及两者的整合。在科技文献相关领域的信息抽取中,常用的机器学习方法包括基于网络图、分类器、支持向量机、逻辑回归、K均值、逻辑模型树、多元逻辑回归、重复增量剪枝以减少误差、线性逻辑回归、随机森林、决策树、缩减误差修减树、决策表、随机树、朴素贝叶斯、决策树桩、神经网络、向量空间模型、文档主题生成模型等[9-10]。

在63篇文献中,有28篇文献至少应用了三大类方法中的一种。而在三大类方法中,机器学习所占比例最大(占比约47%),其次是自然语言处理(占比约39%),最后是统计学(占比约14%)。

5 结语

本文首先从抽取内容和抽取方法两个不同的维度,介绍了当前科技文献信息抽取领域的研究内容。相关领域的研究人员进行了大量的研究和探索,取得了一定的成果,成果体现了一定的价值。相关研究从结构化信息、显式信息和隐式信息三个层面对科技文献进行信息抽取,取得了很多突破。但是,以往的研究并没有涉及到处理XML格式中具有相同标签的不同内容,这会造成文献内部段落歧义的问题;也没有涉及PDF文件中跨页表格的抽取等等。这个需要在未来的研究中加以重视。另外,关键词和引文相关信息的抽取在这一领域也已经有丰富的研究成果,但目前对于方法和算法的抽取研究仅仅停留在"内容层面",即对显式信息的抽取。未来有必要利用语义相关的方法来抽取方法和算法,通过识别和整合来挖掘其中的隐式信息。

其次,本文从方法层面对科技文献的信息抽取方法进行了总结。研究发现,统计学、自然语言处理和机器学习三大类方法被广泛应用在相关研究中。其中,机器学习方法在研究工作中占据了最大的比例。

最后,本文对科技文献信息的抽取内容和抽取方法进行了整合和总结,认为科技文献信息抽取面临着新的挑战。例如,从PDF和XML中抽取结构化信息的研究仍存在一定的进步空间。同时,语义信息的抽取具有挑战性。除了已抽取的信息外,还有很多重要信息的抽取需求,例如,关键发现、前人研究工作、术语等。此外,应用机器学习和自然语言

处理是处理这一问题的趋势。如何将机器学习和自然语言处理结合起来,并在科技文献的信息抽取中获得良好的准确性、覆盖率和可扩展性,仍然是一个挑战。

参 考 文 献

[1] Swaraj K P, Manjula D. Fast extraction of article titles from XML based large bibliographic datasets [J]. Procedia Technology, 2016, 24: 1263 – 1267.

[2] Do H H N, Chandrasekaran M K, Cho P S, et al. Extracting and matching authors and affiliations in scholarly documents[C]//Proceedings of the 13th ACM/IEEE – CS joint conference on Digital libraries. ACM, 2013: 219 – 228.

[3] Gollapalli S D, Caragea C. Extracting Keyphrases from research papers using citation networks[C]//Palo Alto: AAAI Press, 2014: 1629 – 1635.

[4] Bertin M, Atanassova I. Extraction and characterization of citations in scientific papers[C]//Semantic Web Evaluation Challenge. Springer, 2014: 120 – 126.

[5] Kern R, Klampfl S. Extraction of references using layout and formatting information from scientific articles[J]. D – Lib Magazine, 2013, 19(9/10): 15 – 25.

[6] Malhotra A, Younesi E, Gurulingappa H, et al. 'HypothesisFinder:' a strategy for the detection of speculative statements in scientific text[J]. PLoS Computational Biology, 2013, 9(7): 1 – 10.

[7] Green N. Towards creation of a corpus for argumentation mining the biomedical genetics research literature[C]//Proceedings of the first workshop on argumentation mining. 2014: 11 – 18.

[8] Singhal A, Simmons M, Lu Z. Text mining for precision medicine: automating disease – mutation relationship extraction from biomedical literature[J]. Journal of the American Medical Informatics Association, Oxford University Press, 2016, 23 (4): 766 – 772.

[9] Hahn A, Mohanty S D, Manda P. What's Hot and What's Not? - exploring trends in Bioinformatics literature using topic modeling and keyword analysis[C]//International Symposium on Bioinformatics Research and Applications. Springer, 2017: 279 – 290.

[10] Ganguly S, Pudi V. Paper2vec: Combining graph and text information for scientific paper representation[C]//European Conference on Information Retrieval. Springer, 2017: 383 – 395.

[作者简介]敖　龙,深圳职业技术学院图书馆副研究馆员;谢海先,深圳职业技术学院图书馆馆员。

[收稿日期]2021 – 12 – 08　　(刘　平　编发)

An Analysis on Methods of Information Extraction from Foreign Scientific Literature

Ao Long　Xie Haixian

(Shenzhen Polytechnic, Shenzhen, Guangdong 518055, China)

Abstract　Using international databases with great influence such as Web of Science, studies relevant to scientific literature and information extraction were searched and 63 studies were included in this research. By reviewing relevant literature, this research classifies and analyzes the extracted information and the extraction methods and summarizes the contributions of existing research as well as their limitations. The information extracted from scientific literature mainly includes structured information, explicit information and implicit information. The latest and most advanced extraction methods mainly focus on the fields of machine learning, natural language processing and statistics. There is room for improvement as well as challenges in the extraction of semantic information. A flexible combination of methods in machine learning and natural language processing is a future trend for solving the problems in this area.

Keywords　Information extraction. Scientific literature. Semantic information. Machine learning. Natural language processing.