国内外图情领域信息抽取研究文献计量分析

李春杰[1,2],马建玲[2]

(1.中国科学院大学 图书情报与档案管理系,北京 100190;2.中国科学院兰州文献情报中心,甘肃 兰州 730000)

摘　要:【目的/意义】图情领域在数字资源发现、组织与应用中越来越多使用到信息抽取技术,本文将对该领域在信息抽取技术方面的研究进展及应用情况等进行分析,为本领域相关人员提供参考。【方法/过程】以国内图书馆学、情报学领域中国核心期刊和国外33种图情期刊为信息源对其中刊载的信息抽取相关研究成果进行计量分析,检索过程不设置时间限制,并利用CNKI、EndNote、Excel、Python分析工具对主题相关的165篇中文文献和35篇外文文献进行年度趋势分析、期刊发文量分析、机构分析、作者分析以及论文主题分布研究。【结果/结论】得出图情领域信息抽取研究的发展趋势、重要的信息源、重要作者、主题研究以及信息抽取技术在图情领域的具体应用。

关键词: 信息抽取;文献计量;图情领域

中图分类号:G254　　DOI:10.13833/j.issn.1007-7634.2019.04.025

A Statistical Analysis of Literature on Information Extractin of Library and Information Science

LI Chun-jie[1,2], *MA Jian-ling*[2]

(1.Department of Library,Information and Archives Management,School of Economics and Management, University of Chinese Academy of Sciences,Beijing 100190,*China;* 2.*Lanzhou Library,Chinese Academy of Sciences,Lanzhou 730000, China)*

Abstract:【Purpose/significance】 With the rapid development of networking, digitization, and big data, more and more information extraction technologies have been applied to digital resource discovery, organization, and applications. This article will review and analysis the research progress in the field of information extraction technology in this field to provide reference for the relevant personnel in the field.【Method/process】 The data of 18 types of Chinese Core Journals and 32 foreign graphic journals in the field of domestic Library Science and Information Science were collected. Without limiting the years, the retrieved information was selected as the subject of research. In this article, CNKI, EndNote, Excel, and Python analysis tools were used to analyze the annual trends, the volume of journal articles, organizations, the authors and the distributive topics of 165 Chinese articles and 35 foreign articles related to topics.【Result/conclusion】 It draws the development trend of information extraction research, important information sources, major authors, subject research, and the specific application of information extraction technology in the field of Library and Information Science.

Keywords: information extraction;statistical analysis of literatures ;Library and Information Science

信息抽取(IE)是自然语言处理的重要研究内容,在MUC中,IE被定义为"从单个或多个文本中选择性地组织和结合所抽取的数据的技术,这些数据包括隐式及显式数据[1]";Proteus工程的创建者Grishman描述信息抽取的概念:"信息抽取涉及到为从文本中选择出的信息创建一个结构化的表示形式[2]";微软亚洲研究院2005年信息抽取技术暑期研讨班将信息抽取的概念描述为:"信息抽取是抽取和链接基于用户详细说明的相关信息的过程[3]"。信息抽取作为文本信息处理的重要手段,已经成为图书馆学、情报学领域重要的研究内容。笔者以图情领域的重要期刊为检索途径,对国内外图情领域信息抽取研究成果进行统计和分析。

本文选取图情领域作为分析样本的意义在于,从研究对象来看,图书馆学、情报学是关于信息组织、知识服务的学科,信息抽取研究从技术角度解决了信息处理中的诸多难

收稿日期:2018-07-12

基金项目:国家自然科学基金项目"气候变化科学成果集成研究范式及其实现平台研究"(41671535)

作者简介:李春杰(1993-),女,硕士研究生;通讯作者:马建玲.

题，属于图情学科重要的研究内容；从图情实际工作的实践应用角度看，近20年来，随着网络化、数字化、大数据的快速发展，信息抽取作为一种重要的技术方法应用于各个领域，在图情领域为数字资源发现、组织与应用解决了诸多问题。基于信息抽取与图情领域的紧密联系，本文将对该领域信息抽取方面的研究进展，应用情况等进行分析，为本领域相关人员提供参考。

1 国内图情领域信息抽取研究数据收集及检索式分析

1.1 数据收集

如表1所示，本文选定我国情报学、图书馆学18种核心期刊作为检索要素，对国内图情领域的信息抽取研究成果进行收集，以保证研究成果的质量。以CNKI作为主要的检索平台，以“期刊名 AND 主题词”构建检索式限定数据范围。

1.2 检索式主题词分析

本文以选取“信息抽取、知识抽取、数据抽取、实体关系抽取、命名实体识别、抽取规则、机器学习、神经网络、规则抽取”为最终的主题词，“命名实体识别”和“实体关系抽取”是信息抽取研究较为核心的研究方向，“规则抽取”、“机器学习”是较为常用的信息抽取技术，“神经网络”是近年来比较热门的机器学习方法；“抽取规则”是论文标题和主题中常见的表达方式。

笔者将“知识抽取”视为信息抽取的研究范畴，并作为主要的检索词使用。学者张柏林《从知识抽取相关概念辨析看知识抽取的特点和发展趋势》一文中阐述，知识抽取与信息抽取是不同的研究范畴，知识抽取在文献处理的过程中，可以精细到句子级别的抽取，实现文献在知识单元上的组织、管理和利用，主要针对规则性知识、经验性知识，抽取的结果以句子级别的复杂文本为主，且不再是以篇章文献为处理对象[4]；信息抽取的处理对象是事实性知识，重在获取命名实体、实体间关系等，两者的处理对象不同，该学者认为，信息抽取以词的计算为主，对句法分析有要求，但要求不高，知识抽取对句法分析要求很高[5]。笔者认同张柏林学者对于这两个概念的分析，单倾向于将知识抽取作为信息抽取的一个研究子类。知识抽取从实施流程来看，是根据给定的本体，通过语义标注、领域本体、RDF三元组等技术，从无语义标注的信息中识别并抽取与本体匹配的事实的过程；而从技术处理流程角度理解信息抽取，主要分为三个方面：基于用户明确的信息需求、文本信息的抽取、抽取信息的结构化重组，两者在抽取流程上虽然存在一定差异，但是大体可以看作是一种基于元数据的信息抽取[6]；从文本处理深度上来分析，信息抽取已经越来越细化，其分支数值信息抽取的研究也已经在探讨知识元的抽取技术，依赖于句法分析、模型训练，其研究内容与知识抽取的内涵也有交叉，因此，笔者将“知识抽取”作为检索式的重要组成部分。

“数据抽取”同“知识抽取”，在本文中被视为信息抽取的研究范畴。很多人将其理解为数值信息抽取，数值信息是比数据信息更为细粒度的表达，是一种数字的数据，再者，也存在一种现象，数据抽取可能是信息抽取一种不规范的表达。在信息学研究中，数据、信息、知识是依次递进的关系，信息是隐藏在数据背后的规律，知识是更高级的信息的抽象，抽取数据并结构化后得到信息，而知识抽取是信息抽取技术的进一步深层次的研究。因此，“数据抽取”也被作为重要的检索词出现在检索式中。

2 国内图情领域信息抽取研究计量分析

本次数据收集、筛选后得到信息抽取文献165篇，笔者通过EndNote的Subject Bibliography分析功能，以及CNKI本身的数据分析功能，对图情领域的信息抽取研究进行了年度趋势分析、期刊发文量分析、机构发文量分析、主题分析。

2.1 发文年度趋势分析

年度趋势分析揭示了该研究的各个发展阶段，图1展示了各期刊信息抽取研究论文数量。从文献发表时间来看，图情领域的信息抽取的研究大约始于20世纪九十年代，最早发表的文献为1994年发表于《情报学报》的《从文本中提取信息》，该文介绍了基于结构化关键词的文本分析方法，是比较初期的信息抽取研究，主要针对结构化的信息进行抽取。

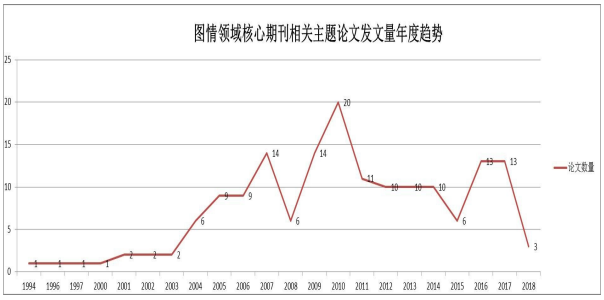

图1 图情领域核心期刊发文量年度趋势图

图情领域的信息抽取研究于2004年开始进入上升期，并于2010年达到高峰，有20篇相关文献发表，9篇文献研究网页信息抽取技术，包括商业信息、网页新闻等内容的抽取，这一时期，领域本体、元数据、机器学习等已经开始成为信息抽取研究的关键词，神经网络也开始成为技术应用的热点内容。最新发表的文献，即2018年发表的三篇文献，主要集中于深度学习方法在信息抽取研究中的应用。深度学习是机器学习最新研究内容之一，主旨在于建立模拟人脑进行分析学习的神经网络。2011至2017年，论文发表的数量开始有所下降，但仍然保持在年均十篇的发文量，可见，信息抽取研究已经成为图情领域一个重要且稳定的研究方向。

2.2 期刊发文量分析

图2揭示了18种期刊的信息抽取的发文数量，反映了研

究成果在期刊中的分布情况，发现165篇论文在期刊中分布并不均衡。

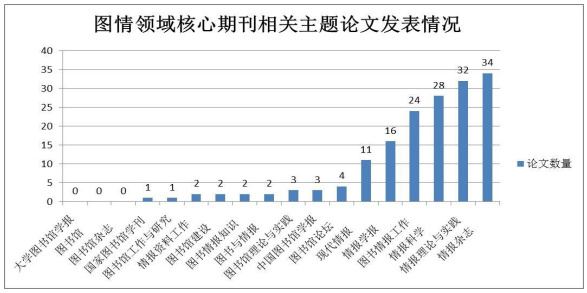

图2 国内图情领域核心期刊发文情况

18种期刊中有15种期刊刊载有相关主题论文，发文量超过10篇的期刊包括《现代情报》、《情报学报》、《图书情报工作》、《情报科学》、《情报理论与实践》、《情报杂志》，其中《情报杂志》和《情报理论与实践》发文量最高，分别为34篇和32篇，与此相对，《大学图书馆学报》、《图书馆》、《图书馆杂志》的发文量为0。期刊的选题侧重点是重要的影响因素，后者更偏向于图书馆学研究成果的收录，信息抽取研究更偏向于情报学研究的范畴。

表1 图情期刊信息抽取研究论文被引情况

刊名	论文篇数	被引次数	篇均被引数
中国图书馆学报	3	45	15
现代情报	11	141	12.8
图书情报工作	24	270	11.25
情报科学	28	313	11.18
情报学报	16	167	10.44
图书情报知识	2	14	7
图书馆工作与研究	1	7	7
图书与情报	2	13	6.5
情报理论与实践	32	190	5.94
情报杂志	34	201	5.91
图书馆论坛	4	15	3.75
图书馆建设	2	6	3
情报资料工作	2	3	1.5
大学图书馆学报	0	0	0
图书馆杂志	0	0	0
图书馆	0	0	0
图书馆理论与实践	3	0	0
国家图书馆学刊	1	0	0
总计	165	1385	8.39

值得注意的是，《国家图书馆学刊》只有一篇相关文章发表，但从发表时间看，该论文发表于2018年，这在一定程度上预示着，信息抽取研究在图书馆学、情报学领域的重视程度在加深。信息抽取技术已经逐步应用到了数字图书馆方向，逐步渗透到图书馆学研究的内容中，并成为图情各期刊重视的内容。

表1以篇均被引频次为依据降序排序，统计了各期刊的发文数量、被引频次、以及篇均被引频次。经计算，15种期刊平均刊载量为11篇，发文量超过平均刊载量的期刊为《情报杂志》、《情报理论与实践》、《情报科学》、《图书情报工作》、《情报学报》；期刊被引频次排名前五位的期刊分别为《情报科学》、《图书情报工作》、《情报杂志》、《情报理论与实践》、《情报学报》；165篇相关文献的篇均被引频次为8.39次，其中，超过篇均被引次数的期刊分别为《中国图书馆学报》、《现代情报》、《图书情报工作》、《情报科学》以及《情报学报》。结合以上指标，《情报科学》、《图书情报工作》、《情报杂志》及《情报理论与实践》是综合排名较高的期刊，可以作为信息抽取领域的重要的情报源。

2.3 机构分析

信息抽取研究机构的计量分析揭示了该研究方向的机构分布。如图3所示，发文量占比超过10%的机构依次为中国科学技术研究所、武汉大学和中国科学院国家科学图书馆（中国科学院文献情报中心），发文量分别为19篇、13篇和12篇，不排除合著的情况。其次为南京大学和中国科学院研究生院，分别为9篇和8篇，五个机构发文量占全部文章总数的38%。同时，结合图情领域核刊发表情况分析，发现这三个重要机构还分别是《情报学报》、《图书情报知识》、《图书情报工作》的主办单位。

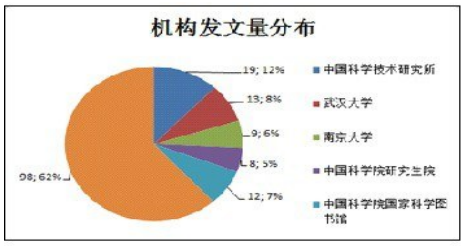

图3 机构发文量分布图

中国科学技术研究所还办有《数字图书馆论坛》、《情报工程》等相关期刊，且该机构拟定于2018年10月举办的第三届知识服务与情报工程学术交流会上，自然语言处理依然是十分重要的研讨内容。该机构对于自身情报研究方向的定位也集中在竞争情报、知识管理、知识抽取、知识发现、智能检索等。中国科学院国家科学图书馆即中国科学院文献情报中心，是国内图书馆学、情报学重要的研究机构，除《图书情报工作》外，还办有《数据分析与知识发现》以及《数据与情报科学学报》等相关刊物。武汉大学图书馆学综合排名一直与北京大学不分伯仲，也是图情领域重要的研究机构，从近年来的研究成果看，武汉大学在信息抽取领域底层应用与工具集方面，开发了细粒度Web数据采集与检索平台WHUREAPER、NLP工具集(词法分析与语法分析)，取得了重大的研究成果。

2.4 作者分析

表2统计了论文产出率较高的重要作者，包括中国科学技术信息研究所郑彦宁、化柏林，武汉大学陆伟，中国科学院文献情报中心张智雄以及华东理工大学李楠。

中国科学技术信息研究所科技报告服务与产业情报研究中心主任郑彦宁教授在2007至2014年间共发表信息抽取相关论文10篇，其中6篇文献是与同机构单位学者化柏林合

表2 图情领域期刊信息抽取相关论文重要作者

作者	机构	论文数量(篇)	论文题目(按年份排序)
郑彦宁化柏林	中国科学技术信息研究所;	10	(2007)信息检索与信息抽取差异性探析 (2008)信息抽取技术在情报学中的应用分析 (2009)汉语信息抽取中事件的定位与分类 (2009)信息抽取中实体关系模式的可信度评估 (2011)国内外属性抽取研究综述 (2011)基于规则的学术概念属性抽取 (2011)针对学术定义的抽取规则构建方法研究 (2011)学术定义抽取系统实现及实验分析 (2011)句子级知识抽取在情报学中的应用分析 (2014)数值信息抽取研究进展综述
陆伟	武汉大学	4	(2005)Web环境下的内容抽取及RSS发布 (2012)产品命名实体特征选择与识别研究 (2012)跨语言信息检索中的命名实体识别与翻译 (2017)学术文本的结构功能识别——在关键词自动抽取中的应用
张智雄	中国科学院文献情报中心	3	(2009)基于Ontology的信息抽取技术方法分析 (2013)开放信息抽取技术的现状研究 (2014)基于对象网格的网络科技信息重要对象识别方法研究

著,研究内容主要集中在时间抽取、学术定义抽取以及信息抽取技术在情报学领域的应用,并分别对信息抽取技术和数值信息抽取作了研究综述;武汉大学情报学专业,信息管理学副院长院,信息检索与知识挖掘研究所所长陆伟在2005至2017年在核心刊物共发表信息抽取相关论文4篇,主要集中在命名实体识别的研究。目前主要研究信息检索、知识挖掘与可视化、竞争情报方法与技术、知识管理;现中国科学院文献情报中心研究馆员、信息系统部主任张智雄在2009至2014年间共发表文章3篇,该学者的研究主要集中在知识技术、信息抽取等方向。

通过计量分析计算得出该研究方向的重要作者,帮助我们更好地追踪信息抽取研究的情报源,帮助该领域的研究人员更好的掌握学术发展的动态。

2.5 主题分析

主题分析反映了该领域的研究水平和总体状况,揭示了该领域的研究现状、热点及发展趋势。表3中图情领域信息抽取论文被归为三类:理论、综述类信息抽取研究,信息化抽取技术实现方法,信息抽取技术在各学科的应用。

2.5.1 理论、综述类信息抽取研究

如图3所示,165篇论文中有25篇属于理论与综述类,包括信息抽取研究综述、计量分析、相关概念辨析、信息抽取方法评估四类。

综述类论文中,包括以"信息抽取"为主题的中外研究综述以及信息抽取子研究的相关综述,例如数值信息抽取研究综述、事件抽取技术研究现状、命名实体识别方法综述、知识抽取项目研究综述等;针对具体问题,现有属性值抽取研究现状、术语识别抽取综述、关键词自动抽取技术研究、信息抽取分类方法研究等综述;从研究对象的呈现类型看,包括对自由文本信息的抽取、开放信息的抽取、web信息的抽取研究综述,针对科技文献,有对于科技文献元数据抽取的研究、基于本体的信息抽取研究综述等,同时,还包括对国内外信息抽取研究的文献计量分析。

评估类论文占比较小,现有研究成果包括对中文命名实体识别算法适用性的评估、实体关系模式的可信度评估;概念类文章主要分析了信息抽取、知识抽取、数据挖掘、文本挖掘、信息检索等相似概念的辨析,以及早期对信息抽取这一概念的介绍性文章,包括对概念、功能、实现方法等基本内容的介绍。

理论、综述类信息抽取文献揭示了信息抽取技术的发展历程、研究趋势、重大成果以及研究的不足,帮助研究者辨明相关概念,对信息抽取保持清晰、客观的认识。

2.5.2 信息抽取技术研究

信息抽取技术按照信息抽取工具的实现原理来分,包括基于自然语言处理方式的信息抽取、包装器归纳方式的信息抽取、基于Ontology方式的信息抽取、基于HTML结构的信息抽取和基于Web查询的信息抽取技术[7]。对于自然语言处理方式的信息抽取可以分为基于知识工程的信息抽取和自动训练的抽取方法;按照抽取方法分类,可以分为基于统计、规则,有监督的机器学习、无监督的机器学习,神经网络以及最新的深度学习的信息抽取技术。

图情领域信息抽取技术方法的相关文献共有118篇,主要围绕术语抽取、网页信息抽取、预测指标信息、主题词、关键词抽取、实体关系抽取、命名实体识别、数值信息抽取、期刊理论抽取等课题的研究。表4统计了每类信息抽取课题研究所用到的技术方法关键词,并在"其他"一栏中罗列了没有归到上述类别所用到的技术方法。

表4统计了165篇文献中出现频率较高的关键词,"神经网络"、"规则抽取"、"条件随机场"、"支持向量机"、"深度学习"、"隐马尔可夫模型"等技术方法出现频率都在四次以上,其中"神经网络"出现11次,规则抽取、条件随机场出现7次,是近年来应用较多的信息化抽取技术方法。图4中列出了信息抽取常用技术在图情领域信息抽取研究中出现的时间和频次,其中深度学习信息抽取技术主要出现在近两年,是比较新的技术研究方向,神经网络从2002年开始出现,在2017年达到应用的高峰,隐马尔可夫模型是一种基于统计的信息抽取模型,自2011年后,便不再是主流的应用技术。

除上述主流技术方法外,信息抽取混合方法的应用也极

表3 主题分类表

信息抽取主题	论文数量（篇）	分类	关键词
理论、综述类	25	信息抽取综述	数值信息抽取、实体关系抽取、命名实体抽取、知识抽取系统、开放信息抽取、关键词抽取、文本信息抽取、网页信息抽取
		信息抽取技术方法效果评估	信息抽取、关系模式、模式匹配、可信度
		信息抽取研究计量分析	年代分布、期刊分布、著者及合著情况、主题分布
		信息抽取相关概念及辨析	信息抽取、知识抽取、数据挖掘、文本抽取、信息检索
信息抽取技术研究	118	术语抽取	关联规则、词形规则模板、PATTree术语抽取模型
		网页信息抽取	双向传播算法、句法分析、统计、文档对象模型、行块分布算法、Jtree、Xpath、Rocchio算法、Widrow Hoff算法、DOM树、Heritrix
		预测指标信息	L-M算法、遗传算法、BP神经网络
		主题词、关键词抽取	TextRank算法
		实体关系抽取	深度学习、规则和本体、支持向量机
		命名实体识别	深度神经网络、规则和统计、支持向量机
		数值信息抽取	语义标注技术、规则抽取技术、正则表达式、C4.5
		期刊理论抽取	条件随机场
		其他	左右信息熵、隐马尔可夫聚类、GATE语义标注、二阶HMM、Viterbi算法、粗集理论、均值聚类算法、传播算法、逆向词性规则、贝叶斯算法、HTTS算法、最大熵模型、Bootstrap
信息抽取技术应用	64	数字图书馆建设	学术概念抽取、摘要生成、信息推送服务、知识组织信息抽取
		情报学研究	专利术语抽取、科学评价数据抽取、产品情报获取、竞争情报系统
		化学领域	化学物质名称识别
		生物医学领域	多文档摘要提取、药物研发信息抽取、语义关系抽取
		商业领域	产品特征提取及情感分类、上市公司风险识别、股票交易、产品命名实体特征选择、电子商务、石油勘探
		新闻媒体	命名实体识别、社会化媒体的本体概念抽取
		管理学	管理问题主体识别

为普遍，包括基于粗集理论和神经网络结合的抽取方法、基于网页分块和统计相结合的网页信息抽取方法，结合语义标注技术、规则抽取技术以及正则表达式的信息抽取、基于本体和DOM树、规则和统计、规则和本体、规则与机器学习的信息抽取方法等也多有应用。

表4 关键词统计表

关键词	出现次数
本体（本体构建、本体学习）	14
Web	13
语义标注、语义信息	13
神经网络	11
术语抽取	11
知识抽取（系统）	11
XML	10
文本挖掘	8
关系抽取（关系提取）	7
规则（规则抽取、规则构建）	7
条件随机场	7
数字图书馆	6
支持向量机	6
半结构化文本	5
抽取规则（抽取技术）	5
情报研究	5
事件抽取	5
元数据（元数据抽取）	5
分类	4
关键词提取、抽取	4
机器学习	4
竞争情报系统	4
深度学习	4
识别（识别效果、识别效率）	4
信息检索	4
学术文本	4
隐马尔可夫模型	4

图4 信息抽取技术高频关键词发展态势

2.5.3 信息抽取技术的应用领域

如表3所示，图情领域的信息抽取主要应用在数字图书馆建设、情报学研究、化学领域、生物医学领域、商业领域、新闻媒体以及管理学领域，共有相关文献58篇，信息技术应用类文献大部分涉及到具体的抽取方法，因此与信息抽取技术类别有所重合。

信息抽取技术在图情领域的应用主要体现在学术概念抽取、摘要生成、信息推送、知识组织信息抽取、专利术语抽取、科学评价数据抽取、产品情报获取、竞争情报系统等内容，以支持数字图书馆的服务与建设、情报系统的建设等工作。

在化学、生物医学领域，现有的信息抽取研究主要集中在相关信息的命名实体识别，如化学名称、药物名称等，同时也有部分语义关系抽取研究，商业领域的研究较为多样化，

表5 国外图情领域学术期刊相关领域发文情况

期刊名称	JCR分区	期刊影响因子	信息抽取文献数量（篇）	信息抽取文献被引次数	篇均被引频次
Information Processing & Management	Q1	2.391	23	277	12.04
Scientometrics	Q1	2.147	2	18	9.0
Library & Information Science Research	Q2	1.185	1	0	0
Information Technology And Libraries	Q3	1.029	1	2	2.0
Journal Of Documentation	Q3	0.853	3	69	23.0
Library Hi Tech	Q3	0.759	1	0	0
Electronic Library	Q3	0.484	2	2	1.0
Library Trends	Q4	0.259	2	86	43.0

包含产品特征提取及情感分类、公司风险识别、股票交易、产品命名实体特征选择等各类信息的抽取任务。同时，还有部分新闻媒体领域的研究，这部分研究主要集中在网页信息抽取技术的范畴，包括命名实体识别的研究，同时，还有关于社会化媒体的本体概念抽取的研究。

3 国外图情领域的信息抽取技术研究

3.1 数据来源

国外图情领域信息抽取研究成果的样本筛选采取了如下检索策略：首先，在WOS的Journal Citation Reports数据库中勾选“Information Science and Library Science”限定学科范围，检索Q1顶级期刊、Q2高水平期刊、Q3普通期刊、Q4一般期刊四个等级的相关学术期刊得到相关刊物85种，包括信息科学、计算机科学、信息管理、医学、图书馆学情报学等诸多信息研究领域；对85种期刊进行筛选后得到相关度较高的33种期刊作为检索源；最后，结合主题词构造检索式如下：

“TI= (“discovery”or"extract*" or "identify*") AND TI= (“information structure”or"information" or "method*" or "technique*" or "knowledge") AND SO=(”Scientometrics” or”Information Processing & management” or”Journal of Information” or”Library & Information Science research” or”Reference & user services quarterly” or”The Serials Librarian” or”The Electronic Library : the international journal for minicomputer” or” Library acquisitions: practice and theory” or”Library Association” or”Journal of Librarianship and Information Science” or” American Libraries” or”Information outlook” or”American Society for Information Science Journal” or”Journal of Documentation” or”College & Research Libraries” or”Library Quarterly” or”Library Resource and Technical Service” or”Cataloging and Classification Quarterly” or”Information Technology and Libraries” or”Library Journal” or”Journal of Library Administration” or”IFLA Journal” or”International Library Review” or”Library Trends” or”Aslib Proceedings” or”Library and Information Science Abstract, London: Library Association” or”Information Science Abstracts”or"library and information science"or"library resources*"or"electronic library"or"library hi tech"or"health information and library*"or"information and organ*")”

3.2 国外图情领域期刊发文数量及被引情况分析

在检索得到的82篇文献中，筛选后得到目标文献35篇，期刊分布和期刊影响力如表5所示。检索式出现的33种图情期刊中，仅有四分之一的图情领域期刊发表过信息抽取研究相关的论文，其中图情领域国际顶级期刊《Information Processing & Management》在35篇文献中发文数量占比为三分之二；《Scientometrics》是由匈牙利出版的一份权威性的国际期刊，主要刊载科学计量学领域的研究论文、短讯和评论，受到了图书馆和情报学专家的特殊重视，在信息抽取领域仅发表过两篇相关论文。

从刊载信息抽取相关研究的期刊数量和期刊的发文数量来看，图情领域期刊在信息抽取研究方向的收文情况并不乐观，在国际上，信息抽取研究并不是图情领域主流的研究内容。纵向来看，与国内图情领域的研究相比，国外信息抽取研究开始时间较早，最早的文献出现在1980年，研究科技文献重要名词短语的抽取技术[8]；被引次数最多的文献为1999年发表于《Library Trends》的一篇文章，被引频次为80次，主要介绍了信息抽取过程中的分类技术[9]。

3.3 国外图情领域信息抽取相关研究主题分析

国外图情领域对信息抽取的讨论主要集中在技术研究和应用方面。

在35篇相关论文中，对语义技术的利用，是图情领域信息抽取研究的重要研究方向。语义分析技术处理文本作为知识抽取的重要手段，囊括了分词技术、构建语言模型、词向量计算、关键词提取等方法，借助机器学习帮助人们识别和获取知识、信息。相关研究包括，应用语义角色标注技术处理自由文本图像的描述信息，对图像描述中基于事件的知识进行自动化提取，包括来自图像描述的主题、动词、宾语、位置和时间信息[10]；除图像描述信息的处理，还包括对科技文献的处理、科技政策的处理，通过对关键词、专业术语的抽取，帮助用户快速有效地获取有价值的信息[11]。

国外图情领域解决信息抽取问题的主要技术方法集中在机器学习算法的研究上，机器学习也是研究成果中出现最多、应用最广泛的信息抽取技术，主要涵盖了支持向量机、维特比算法、马尔科夫模型、隐马尔可夫模型、Patricia Tree、神

经网络、条件随机场等主流方法。这些技术被广泛应用于命名实体识别、命名实体间关系的抽取、语义关系的抽取、元数据抽取以及知识抽取等领域。

国外图情领域学者利用信息抽取技术针对多种类型的文本进行了抽取技术的研究。以科技文献为抽取对象的研究，主要涉及到关键词抽取、概念抽取、专业术语的抽取；对专利文献、科技报告的处理，致力于对文本信息和文本中隐含知识的识别和抽取；web抽取研究中POSIE信息提取系统以及n-gram模型被广泛应用，包括对网页信息、网络公告板信息的处理等；同时，学者们致力于对深度网络文本的抽取研究，深度网络文本集合的特殊之处在于，对于信息用户而言，该类文本内容是不可抓取的，只能通过查询获得。

除上述几种类型外，还有3篇非技术性文献，介绍了信息抽取的发展历程、重要技术和面临的挑战以及信息抽取技术在各领域的应用等。

与国内图情领域信息抽取研究相比，国外此领域论文成果较少，涉及范围也较窄，但是具有一定的影响力。在作者和机构分析中发现，发文最多的学者为Choi.K.S，仅两篇文章，机构分布也不集中，因此在本文中不做具体论述。

4 信息抽取技术在图情领域的应用

图书馆学、情报学本质上是关于信息、知识的科学，随着信息时代的来临，各类载体形式的信息正处于指数型增长，海量的数据以有序的、无序的、结构化、半结构化或者非结构化的形式呈现在图情领域信息工作者面前。面对人力难以处理的海量信息，信息抽取技术使文本信息自动结构化处理成为可能，为图情领域的知识组织与建设、信息服务、情报研究等工作带来了技术性的变革。

4.1 基于信息抽取技术的知识库的构建

知识库构建是图情领域重要的工作内容，依赖于专家系统设计的规则集合的知识库，需要大量基于规则联系的事实及数据。面对庞大的数据和信息原料需求，知识库的构建已经离不开信息抽取技术的支持。信息抽取技术的应用，使海量数据信息的快速处理成为可能，为知识库构建节省了大量的人力资源，同时提高了工作效率。

目前，针对图情领域知识库构建问题已经做了大量的研究。对于科技文献的处理方面包括，基于人工构建规则的方法，完成对科技文献的属性抽取任务，从而建立起属性描述类型的知识库[12]，并通过信息抽取技术对文献内容中的关键词、主题词、高频词进行抽取和分析处理，从而自动生成知识库文献的文摘内容[13]；针对科技文献，信息抽取技术实现了文本内容中重要理论名称的抽取，句子级知识抽取等，在信息检索、参考文献自动标注系统、文献自动综述系统中也有所应用，帮助信息资源的有效利用[14]，支持知识库系统建设；信息抽取技术除作用在科技文献外，还致力于专利信息的抽取，已经有相关技术研究实现了中文专利术语的抽取。目前，基于信息抽取技术，知识库建设已经转向以构建知识元数据库为主[15]，例如CNKI的概念知识元库，通过知识元的抽取与标引，构建知识库。

4.2 基于信息抽取技术的信息服务

图情领域利用信息资源为广大信息用户提供信息服务，信息服务的质量在很大程度上决定了外界对图情领域工作成果的认可度。面对信息用户个性化的信息需求，信息服务工作者利用信息抽取技术可以为信息用户主动、友好、准确的提供所需信息。通过对用户的信息行为进行信息抽取，例如用户在图书馆主页的访问数据、读者借阅历史等资料，从而对用户的兴趣和信息需求进行分析，以此为依据提供有效的、具有针对性的信息推送服务[16]；面向信息服务的信息抽取技术还包括专利术语的抽取，主要应用到专利信息服务中，以支持信息分析、机器翻译等功能。

4.3 利用信息抽取技术构建数字图书馆

图书馆的发展经历了传统图书馆向数字图书馆的变革，数字图书馆建设涉及信息资源的加工、存储、检索、传输和利用的各个环节，与计算机技术、数字技术紧密相关。

信息抽取技术在信息资源的加工、存储环节发挥了很大的作用。首先，在信息的数字化过程中，信息抽取技术实现了特殊文本的识别，例如对于手写汉子、古典名作、地方志等内容的识别；在对纸本学术期刊数字化的过程中，信息抽取技术取代了人工作业，完成了数字图书馆建设对期刊中大量元数据的抽取任务[17]。元数据抽取技术同时还作用于会议文献的处理，用于构建针对会议文献集的处理模板，从而构建和完善数字图书馆的服务系统。同时，利用海量的、瞬息万变的web网页信息，也成为完善数字图书馆建设的重要步骤，信息抽取为此提供了大规模数据及信息采集的思路[18]。为了更好地组织和使用信息资源，信息抽取技术基于“元数据、领域本体、本体解析体系”建立语义模型，实现数字资源语义关系的形式化描述[19]；目前，已有学者利用开源软件构建数字图书馆信息抽取与知识发现平台[20]。

4.4 信息抽取技术在竞争情报中的应用研究

竞争情报工作致力于情报系统的建设以支持决策。决策情报是对组织具有深远意义的情报，信息抽取能够将所需信息、知识、数据从大量的文本中提取并结构化存储，它完成了在网络、文献内容中的数据、数值、信息和知识的抽取和集成，从而为竞争情报分析、支持决策服务提供支撑。

信息抽取研究是情报系统建设的重要环节，目前有很多面向竞争情报领域的信息抽取研究。目前应用于竞争情报的信息抽取研究，以事件抽取为主，将情报分析功能深入到更细粒度的内容挖掘，围绕情报事件的构成元素自动获取较为精确的情报信息，并提供结构化的事件呈现[21]，包括基于本体信息抽取的竞争情报预处理分析，利用本体来深层次抽取竞争情报[22]；基于机器学习的信息抽取技术，例如隐马尔

可夫模型,也被引入到文本信息的抽取中,帮助构建竞争情报系统模型[23]。

信息抽取技术的研究自20世纪50年代至今不断走向成熟。已经成为与信息检索、自然语言理解、文档分类并重的语言工程项目。随着信息技术的发展,图书馆学与情报学研究也不断涌现出新的课题与信息抽取技术紧密结合。面对数据时代的来临,图情领域研究面临着巨大的挑战,信息抽取技术以及更多的信息技术、计算机技术必将成为图情领域迎接挑战、解决问题的利刃。

5 结 语

通过对期刊的发文情况、被引次数和篇均被引频次的统计,分析得出:四种关注信息抽取领域的重要期刊,外文期刊以《Information Processing & Management》为最佳,它们可以作为该研究方向的重要信息源,这一分析结果为科研人员的投稿选择提供借鉴。

在机构分析中,笔者结合发文量、各研究院所的研究方向、办刊情况,得到中国科学技术研究所、武汉大学和中国科学院文献情报中心三个主要的研究机构;重要作者为中国科学技术信息研究所郑彦宁、化柏林,武汉大学陆伟,中国科学院文献情报中心张智雄以及华东理工大学李楠,可以作为重要的学术跟踪信息源,而国外图情领域研究由于数量较少,作者分布和机构分布都较为分散。

笔者重点对165篇中文文献和35篇外文文献进行了主题分析和技术应用领域分析,统计了各领域研究涉及到的信息抽取技术,图情领域的信息抽取技术研究经历了基于规则的知识工程的抽取方法,有监督、无监督、半监督的机器学习方法,以及近些年热门的神经网络的机器学习方法、深度学习的信息抽取方法,广泛应用于生物医学、商业、新闻、图情领域。

本文仍有很多不足,例如,不能保证文献收集的全面性和准确性,在计量过程中,或许会存在小误差,关键词年出现次数排序为通过算法编写抽取,而被引频次为借助EndNote、CNKI的分析功能人工逐篇累加计算所得。研究希望能够尽可能准确地反映图情领域在信息抽取研究领域的现状,敬请广大专家学者批评指正。

参考文献

1 吴 超,郑彦宁,化柏林. 数值信息抽取研究进展综述[J]. 中国图书馆学报,2014,40(2):107-119.
2 Ralph Grishman. Information extraction: Techniques and Challenges.In Maria Teresa Pazienza,editor,Information Extraction[C]. Springer Verlag, Lecture Nots in Artificial Intelligence,Room,1997
3 邓尚民,孙玉伟. 信息抽取系统的研究现状[J].现代图书情报技术,2006,(3):55-58,81.
4 马费成.情报学的进展与深化[J].情报学报,1996,15(5):338-344.
5 化柏林,张新民.从知识抽取相关概念辨析看知识抽取的特点和发展趋势[J].情报科学,2010,28(2):311-315.
6 谌志群,张国煊. 文本挖掘研究进展[J].模式识别与人工智能,2005,18(1):65-74.
7 陆科进,李新颖.基于Ontology的文本信息抽取[J].计算机应用研究,2003,(7):46-48.
8 Maeda, T., Y. Momouchi and H. Sawamura. An automatic method for extracting singificant phrases in scientific or technical documents[J]. Information Processing & Management,1980,16(3): 119-127.
9 Kwasnik, B. H. The role of classification in knowledge representation and discovery[J].Library Trends , 1999,48(1): 22-47.
10 Lin, C.-H., C.-W. Yen, J.-S. Hong and S. Cruz-Lara . Event-based knowledge extraction from free-text descriptions for art images by using semantic role labeling approaches[J].Electronic Library,2008,26(2): 215-225.
11 Zeng, W., C. Yao and H. Li.The exploration of information extraction and analysis about science and technology policy in China[J].Electronic Library,2017,35(4): 709-723.
12 丁君军,郑彦宁,化柏林.基于规则的学术概念属性抽取[J].情报理论与实践,2011,34(12):10-14,33.
13 刘 挺,吴 岩,王开铸.基于信息抽取和文本生成的自动文摘系统设计[J].情报学报,1997,(S1):31-36.
14 胡昌平,林 鑫,陈 果.科技文献副主题词抽取及其在分面检索中的应用[J].情报学报,2014,33(8):837-845.
15 郑彦宁,化柏林.句子级知识抽取在情报学中的应用分析[J].情报理论与实践,2011,34(12):1-4.
16 邱亚娜.信息抽取在图书馆信息推送服务中的应用研究[J].图书馆工作与研究,2011,(1):46-47,55.
17 陈淑平,梁东魁.基于特征分析的数字化期刊元数据自动抽取算法[J].情报杂志,2010,29(3):143-146.
18 刘鲁红,刘力强,胡亚军.信息抽取技术在数字图书馆中的应用研究[J].情报理论与实践,2005,(3):321-324.
19 牟冬梅,陈 倩,王丽伟.基于语义模型的数字图书馆知识组织信息抽取策略[J].图书情报工作,2009,53(15):21-25.
20 奉国和.基于GATE的数字图书馆信息抽取技术概述[J].情报杂志,2009,28(5):171-174.
21 李 楠,吉久明,孙济庆,郑荣廷.基于事件抽取的竞争情报系统[J].情报理论与实践,2014,37(5):77-82.
22 余 丰,朱东华.信息抽取技术在竞争情报研究中的应用[J].情报杂志,2006,(3):25-26,29.
23 徐 萍,邵 波.基于本体信息抽取的竞争情报预处理分析[J].情报杂志,2008,(9):33-35,38.

(责任编辑:赵红颖)