情报科学
第37卷 第2期 2019年2月

数值信息抽取研究概述及应用分析

李春杰[1,2],马建玲[2],主雪梅[3]

(1.中国科学院兰州文献情报中心,甘肃 兰州 730000; 2.中国科学院大学 图书情报与档案管理系,北京 100190;
3. 河北水利电力学院,河北 沧州 061001)

摘　要:【目的/意义】面对海量的信息,人们需要更为高效准确的信息获取方式。数值信息抽取的研究使隐含在无序信息载体中的大量有价值数值信息可以得以利用,从而满足科研工作者数据驱动型研究的信息需求。【方法/过程】本文旨在总结和归纳数值信息抽取研究的相关内容,包括数值信息抽取的内涵、数值信息抽取研究概况、面临的困境和制约因素以及应用等。【结果/结论】数值信息抽取仍然面临着巨大的挑战,且现有的数值信息抽取研究较少,对于数值信息的抽取,基于规则和统计学习的方法各有利弊,总体来说,基于规则的抽取方法仍是主流方法。

关键词: 数值信息;数值信息抽取;数值信息抽取理论

中图分类号: G250.2　　**DOI:** 10.13833/j.issn.1007-7634.2019.02.007

A Overview of Numerical Information Extraction Research and Application Analysis

LI Chun-jie[1,2], *MA Jian-ling*[2], *ZHU Xue-mei*[3]

(1.Lanzhou Library, Chinese Academy of Sciences, Lanzhou 730000*, China; 2.Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing* 100190*, China;* 3. *Hebei University of Water Resources and Electric Engineering, Hebei* 061001 *, China)*

Abstract:【Purpose/significance】Faced with massive amounts of information, people need more efficient and accurate ways to obtain information. There are also a lot of useful disordered information, research on numerical information extraction can help us use this information to meet the information needs of scientific research workers who does data-driven research. 【Method/process】This paper aims to summarize the relevant content of numerical information extraction research, including the connotation of numerical information extraction, the research overview of numerical information extraction, the dilemmas, the main constraints and applications.【Result/conclusion】The extraction of numerical information still faces enormous challenges, and there is less research on the extraction of existing numerical information. For the extraction of numerical information, methods based on rules and statistical learning have advantages and disadvantages. Overall, rules-based extraction methods are still the mainstream method.

Keywords: numerical information; numerical information extraction; numerical information extraction theory

随着大数据时代的来临,海量的信息、数据、知识隐藏于结构化、半结构化或无结构的文本集合中。对于信息用户而言,需要海量数据的发现、抽取、统计和分析以满足数据驱动的科学研究,而基于关键词检索的信息获取方式无法提供细粒度、深层次和有序化的数据信息,特别针对科技文献,关键词检索只能提供摘要、全文信息,无法获取文本内容层的重要概念、实体和数据;对于信息组织者而言,作为信息的加工者,面对海量的原始资料信息,获取、过滤、组织、建库是十分艰巨的任务,基于此,自动化处理海量信息的信息抽取技术得到越来越广泛的应用,对于信息抽取技术的研究也在不断发展。

信息抽取常指抽取特定的事实信息,MUC测评会议将信息抽取定义为"从单个或多个文本中选择性地组织和结合所抽取隐式及显式数据的数据的技术[1]。"数值信息抽取是中文信息处理研究中的基础研究方向,是信息抽取、信息检索、机器翻译、自动问答系统、知识库或事实库等多种自然语

收稿日期: 2018-09-27

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,主要研究信息资源组织与建设; 通讯作者:马建玲.

表1 国内数值信息抽取研究

时间	处理对象	作者	方法
2018	英文科技文献	郭少卿	利用远程机器学习、语义标注得到取值关系模板，识别语义关系，处理相对数值信息，计算实际取值。
2018	网络问答信息	张桂平	通过制定模板模块、模式匹配模块，利用条件随机场补全模块，最后时限时间地点抽取。
2016	中文文本数值关系	吴胜	利用词性、句法树等特征生成候选主体集，用J-W距离打分得到取值主体。
2015	医学数据	谢维佳	基于规则、正则表达式以及相似度匹配的方法，抽取体温数据、脉搏、呼吸、心率等数据。
2015	时间数据	张远鹏	人工标注生成训练数据集，利用HMM的方式产生抽取规则。
2010	新闻法律事件	毋菲	利用规则的方法，通过正则表达式、决策树C4.5算法以及句法分析完成对于特定法律事件中的数值元的抽取。
2008	年鉴	肖洪	基于规则与机器学习相结合的方法，借助各类行业词表，从而建立抽取规则。
2008	网页抽取	杨少华	训练模板，利用模板学习的方式抽取网页中的数值信息。

言处理技巧的基础[2]，本文主要对数值信息抽取技术的研究进行归纳。

1 数值信息抽取概念分析

数值信息抽取是信息抽取重要的研究领域之一。明确数值信息抽取的概念，能够帮助研究者明晰数值信息抽取研究的边界。

数值信息：从表达形式来看，数值信息大多以“数字”表达为构成主体，与量词、符号、字母组合，从而赋予数值具体的信息类型，例如时间类数值信息、货币类数值信息等；同时，数字与其他符号组合，形成具有特殊含义的字符串，如电话号码、邮箱地址、各类编号、车牌号码等，也属于数值信息的范畴；“数字”的存在并不是数值信息的必要特征，在ACE测评会议中，曾将特定情景下的事件名称也作为数值信息处理，如“北京奥运会的赞助单位为搜狐公司”、“小明在百度公司担任总经理”，其中的单位信息、职位信息也被认为是数值信息的一种[3]。

数值信息抽取：数值信息抽取可以简单定义为对给定文本集合中的数值型信息进行抽取。数值是一种有价值的信息，在于它存在于一定的上下文中，因此数值信息抽取，需要对包含有数值信息的客观事实描述语句进行识别，而并非只单纯获得数字表达，其抽取结果一般以元组形式呈现，在抽取结果中呈现出与命名实体之间的逻辑关系，从而产出有用的知识。

2 数值信息抽取相关研究概述

数值信息抽取在数据驱动的文献综合集成研究、各类决策支持系统以及信息服务平台中都逐渐发挥出重要的作用，作为信息抽取重要的研究分支，国内外学者对其进行理论及技术方面的探索和研究，以求解决各领域对于数值信息收集、提取、利用的问题。

2.1 数值信息抽取相关学者研究

数值信息抽取始于2000年12月正式启动的ACE评测会议，在医学领域、商业领域、军事领域等都取得了一定的成效。本文从国内和国外两个维度，对数值信息抽取已有研究进行归纳总结。

国内对于数值信息抽取的研究自2008年至2018年十年间，发表的研究成果如表1所示，本文筛选出主题符合度较高的8篇文献，分析了近年来数值信息抽取研究的内容。这些研究主要围绕网页信息、科技文献、病例与新闻报纸等载体类型，涉及到中英文文献的抽取，囊括了对于事件中知识元、绝对数值信息以及相对数值信息的识别和发现，其中决策树、句法树、机器学习、正则表达式是应用最为广泛的技术手段，基于规则的抽取方法仍为主流方法。

表2归纳了国外学者对于数值信息抽取的研究情况，从表中内容可以看出，早期国外对数值信息抽取的研究主要集中在医学、新闻领域，借助字典和规则的方法为主要的数值信息抽取手段，2009年开始，机器学习的方法开始应用到数值信息抽取的研究中，通过训练模板进行信息抽取，自此，SVM、HMM等机器学习的模型开始广泛应用在数值信息抽取的过程中，并逐步引入神经网络等先进技术。

与国内研究相比，从研究的时间来看，国外数值信息抽取研究开始于1998年，起步更早；从研究内容来看，国外数值信息抽取研究在生物医学领域的研究偏多，新闻语料的处理占较大比例，而国内主要偏重于自由文本、科技文献等载体类型，但在题材的全面性和覆盖性来看，差别不大；从抽取技术的使用上来看，手段上大体相似，基于人工、机器学习的方法都用应用。

从现有的国内外研究成果可以看出，对于数值信息抽取的研究涉及到的领域、技术方法、抽取对象的丰富性不足；从数值信息抽取技术来看，基于规则的数值信息抽取方法，广泛应用决策树算法训练数据集，准确率较高，但需要较多的人为参与和时间投入；基于机器学习的信息抽取方法近年来不断完善，主要分为有监督和无监督或半监督的机器学习，效率较高，但是受到语料库、属性值全面程度的影响，抽取的准确率不尽人意。同时，各类数值信息抽取系统还会受到领域差异的影响，抽取方法的移植性变得很弱，缺乏泛化能力，数值信息抽取技术仍然面临着诸多挑战。

2.2 现有数值信息抽取研究的常用模式

通过对电子病历记录[4]、科技文献、网页、新闻报纸等不同载体类型的数值信息抽取技术进行归纳研究，本文总结得出数值信息抽取研究的常用模式，如图1所示，包括训练样

表2 国外数值信息抽取研究

时间	处理对象	作者	方法
2016	实验数据规模描述	Sarker	将词共现、词相似度、段落结构作为特征加入SVM对句子进行二分类。
2016	解释短语库	Chaganty	利用RNN完成指标取值换算。
2015	自由文本	Maiya	正则表达式抽取带单位的数值
2012	生物领域实验报告	Santos	领域本体构建规则识别生物领域的实验参数。
2010	手写字母数字序列	Simon Thomas	使用HMM进行模式识别,提取特征向量序列。
2009	新闻语料中指标取值	Murata	使用远程监督发现模板,借助人工构建的外部知识库学习语料中模板。
2009	日文新闻报纸	Masaki MURATA	借助半自动的单词分类词典的方法提取数字信息和命名实体集。
2007	电子病历	Voorham J Denig P	文本识别算法、正则表达式抽取糖尿病相关的数值临床数据。
2007	新闻报纸	Nanba H	基于规则的抽取方法抽取时间、价格、销量、市场占有率等信息。
2006	医师记录文本	Turchin A	正则表达式抽取血压值。
1998	文章摘要	Fukuda K	基于规则的抽取方法抽取蛋白质名。

本集的准备工作、机器学习系统的选择,目标语句识别以及最后的抽取。

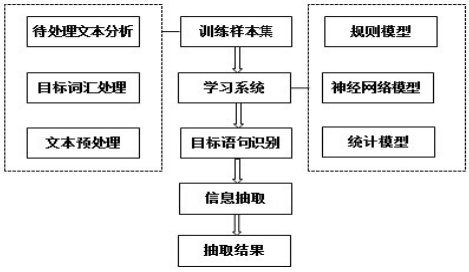

图1 数值信息抽取研究模式

(1)待处理文本分析。

待处理文本分析主要指对数值信息的载体以及需要抽取的数值信息相关词汇进行分析,包括文本结构及重要词汇,该过程可以帮助简化抽取过程,以地学领域为例,该领域论文的位置信息一般都存在于"研究区域"章节;重要词汇分析主要指重要的触发词、停用词、引导词、领域专用词等,以完成文本的特征选择和特征值分类,要借助特征选择,从数据集合里按照指定的规则选择出具有良好的区分特性的特征子集,以此降低计算开销和提高分类性能,减小数据处理量,从而节省处理时间[5]。对这些重要的特征词汇进行汇总,建立特征词库和规则库,从而更高效的识别目标词汇,简化抽取步骤。

(2)目标词汇处理。

目标词汇指最终抽取的数值信息(300mm)以及与之相对的实体特征值(年降水量)、主体(兰州)等,即最后需要呈现的元组元素。对目标词汇的分析包括对目标词汇类型的分析、词性的分析,以及各成分的构成模式等;目标词汇类型分析,如"时间数值信息"、"货币数值信息"、"电话号码"、"邮箱地址"等是按照不同的范畴类别进行分类,或者是"绝对数值信息"、"相对数值信息"广义上的数值信息分类方式,都是影响抽取规则的重要因素,在目标词汇处理阶段,将数值信息的类型、构成分析到位,能够在很大程度上帮助归纳总结抽取规则以及定制抽取模板,从而为抽取工作打下基础。

(3)文本预处理。

通过文本预处理生成计算机能够识别和处理的数据集。该过程需要先将文本信息转化为机器可读的格式,从而进行下一步骤的操作。文本预处理与待处理文本分析作用相似,区别在于,该步骤一般运用非人工的方式进行,包括对语料的分词、去停用词、词性标注、词性过滤等内容,并直接服务于抽取的过程[6]。经过预处理环节生成带标注的数据集,从而使计算机能够更快速准确的了解语料内容,并识别相关的信息。

(4)目标语句识别。

目标语句的识别是数值信息抽取的难点。数值信息识别分为两种类型,一种是目标词含有数字、规律性较强的数值信息,一类是目标词中不包含数字特征的词,规律性不强的数值信息。如图2所示,对于规律性较强的一类,使用正则表达式[7-8]或者根据规则生成匹配树即可解决[6],将目标词处理过程中形成的各类表达规则、模板转化为匹配树,从而形成相应的抽取策略实现数值信息抽取,该方法主要依赖于人工编制规则;对于规则性较弱的数值信息,人工编制规则不足以发现和识别所有的目标语句,于是引入机器学习的方法予以辅助[9],其迭代学习过程会增强识别效果,根据语料内容,人工方式定制出种子模板,得到原始语料中应被识别的句子,分析抽取结果的表达模式,丰富模板,然后将该过程重复;基于隐马尔科夫模型[10]、条件随机场、决策树等理论的诸多机器学习算法已经广泛应用于数值信息抽取系统中,这些技术弥补了人工归纳抽取规则、建立模板的不足,可以更好的识别实体、属性等内容。

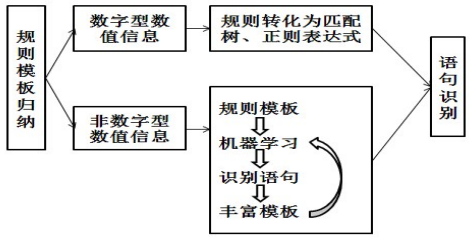

图2 数值信息抽取的模板归纳

(5)数值信息抽取。

现有研究中,正则表达式、匹配树可完成数字特征明显且具有较大规律性的抽取,句法分析可以针对规律性较弱的非数字类数值信息进行抽取,基于统计的句法分析是比较主

表3 信息抽取技术的理论基础

重要理论	功能	特点
决策树理论（C4.5、C5.0、ID3、ACLS、CART-TRELIEF、CART、CHAID）	文本的分类与识别	复杂度较小，速度快且抗噪声能力强，可伸缩性强，既可用于小数据集，也可用于海量数据集，且学习过程中使用者不需要了解很多背景知识，且分类原理简单易懂[11-12]
隐马尔科夫模型	分词及词性标注	不需要大规模词典和规则集、移植性好、投入成本较少等显著优势[13]
条件随机场理论	词性标注	解决了标注偏置问题，可以参考抽取目标文本的上下文信息，且算法的特征设计更为灵活[14-15]
VC维理论	文本分类和语句识别	解决了机器学习过程中训练数据样本的数量的界定，从而保证产生的分类器的效果尽可能最好[16-17]
神经网络（KNN、DNN、SVM、DL、BP、DBN、RBF、CNN、ANN）	灵活模拟数据之间的复杂的关系	有很强的鲁棒性、记忆能力、非线性映射能力以及强大的自学习能力[18]。
结构风险最小理论	使学习机器在样本集上的期望风险得到控制	寻找最小经验风险，考虑经验风险和置信范围，取得实际风险的最小化[19]。
最大熵理论（迭代尺度法(IIS)、梯度下降法、牛顿法、拟牛顿法）	词性标注、短语识别、指代消解、语法分析、文本分类、问题回答	基于最大熵理论构建模型，特征选择灵活，模型可以应用在不同领域，可移植性强[20-21]。

流的抽取方法，主要的句法分析器包括Stanford parser和Berkeler parser，也有基于规则的句法分析算法，包括Chart、CYK、GLR等。句法分析与规则抽取相结合，一般会得到“主体、指标名称、数字、单位”或者“主体、指标名称、数值元”的内容，最后生成需要的元组表达结果。数值信息抽取可能会得到绝对的数值信息，完成抽取；而针对相对数值信息，则需要进一步的数值信息换算，从而得到实际的数值信息，如图3所示。

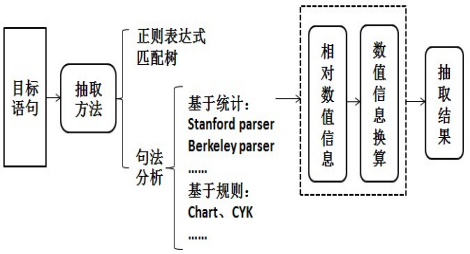

图3 数值信息抽取的目标语句识别

面对不同的数值信息抽取任务，抽取方式各有不同，以上内容属于较为普适性的数值信息抽取模式，针对要解决的难点对抽取方法进行调整，最后找到最适合的数值信息抽取方法。

2.3 数值信息抽取相关理论

基于目前的研究现状来看，数值信息抽取技术主要分为基于规则的数值信息抽取、基于统计的数值信息抽取以及基于规则和统计相结合的信息抽取方法，这些技术方法背后存在相应的理论基础，基于这些理论设计各类算法，实现数值信息抽取过程中规则模板的生成、目标语句的识别。

如表3所示，决策树理论、隐马尔科夫模型、条件随机场、VC维理论、神经网络、结构风险最小化理论以及最大熵理论是目前信息抽取技术最为依赖的科学理论。在这些原理的指导下，逐步解决了文本分类、分词、词性标注、语句识别、句法分析等自然语言处理的任务。

从目前来看，数值信息抽取过程中仍然需要人工的参与，但是通过自动化的机器学习等技术来完成大部分的工作，仍然是我们要追求的目标，因此，对于技术的革新依然是一大难题。

3 数值信息抽取技术的应用

3.1 数值信息抽取技术在综合集成研究中的作用

综合集成研究是一种数据驱动的科学研究模式，对数值信息的需求量极大。以地学领域为例，地球空间数据一般表现为分布式、多尺度的数据，数据量大、获取费用高、共享效益高、涉及专业部门多[22]，把分散的数据集成到统一的数据仓库，是综合集成研究重要的步骤。例如，在亚洲气候变化集成研究中，数据的收集、整编规范工作十分繁杂，需要收集各种时间分辨率的代用记录，最终建立亚洲两千年气候代用资料数据库[23]；同样，全球变化研究也是基于数据集成的研究项目，属于全球尺度的、多学科、多组织、大团队的集成综合研究，对基础数据和衍生数据的收集有重要需求，很多研究项目都需要对专门的数值信息进行发掘、统计和集成分析[24]。基于该类科学研究的需求，数值信息抽取技术成为一项重要的研究课题。

3.2 数值信息抽取技术在决策支持中的作用

决策支持系统是一种计算机应用系统，对事实数据进行加工，从而辅助决策者作出决策。支持决策系统依赖于数据库、模型库、知识库、方法库进行数据收集、数据处理、数据分析，需要对数据进行转换与清理。数值信息抽取技术，是获取数据的手段，同时，也是将数据进行统一、规范化的过程[25]。

对于决策支持系统，分散的数据也是无法支持其使用的，系统必须进行数据分析、处理、融合等操作以构建相应数据模型，进而满足用户需求。例如地质灾害预报预警及决策支持系统需要集中的数据支持，需要分散的相关数据源集成到统一的数据仓库[26]。数值信息的抽取，更好的满足集成研究对多源数值数据获取的需求。

3.3 数值信息抽取技术在信息服务中的作用

在信息服务行业，利用数值信息抽取技术可以提升信息服务的水平。面对信息用户个性化的信息需求，信息服务工作者需要获取信息用户的相关信息对用户的隐性信息需求进行判断，提供主动、友好、准确的信息，例如用户在图书馆主页的访问数据、读者借阅历史等资料，都是帮助用户需求分析的有用信息，从而依据此为用户提供有效的、针对性的信息推送服务。

数值信息抽取在信息服务中的应用，还体现在为知识组织和知识服务提供结构化数据和知识。基于对文献内容的深层挖掘，建立专题知识库或领域本体，如抽取领域命名实体，包括国家、机构、主题、地名、重要数据等，从而建设领域知识库，为知识组织和知识服务提供数据基础。

在面向数值信息的问答系统中，数值信息抽取技术同样得到应用。面向数值信息的问答系统的建立需要考虑数值答案的抽取和生成，并构建答案库，从而为信息用户提供接近人工问答水平的体验以及准确无误的服务质量[27]。

数值信息抽取技术为综合集成研究构建了数据库基础，为支持决策系统提供了必要的数据材料，同时在信息服务过程中，最大限度的解放人力，力求利用已有的信息，提供信息服务。在各项工作中，数值信息抽取技术面对的是多元化、分布式、多尺度的处理文本，需要解决数据的规范化处理以及与各类系统的对接。面对数据时代的来临，数值信息抽取技术将面临更大的挑战。

4 数值信息抽取面临的挑战

数值信息抽取研究很难找到一种普适性的方法，领域差异性、数值信息丰富多样的类型都决定了数值信息表达方式的多样性和复杂性，从而影响到抽取过程中具体的实施策略；同时，中文文献数值信息表达没有词间空格作为明显的分词标志，对分词技术以及词性标注技术要求较高，这些因素都增加了数值信息抽取的难度。

4.1 领域差异性对数值信息抽取的影响

领域差异性决定了相同类别的数值信息具有不同的表达形式[28]。如表4所示，以新闻学、地球科学以及历史学三个领域的时间抽取任务为例，在新闻学领域，新闻事件的发生时间、新闻的报道时间作为有价值的数值信息，常使用日常的时间表达方式，如“xxxx-xx-xx”、“xxxx年xx月xx日”等；而在地球科学和历史学等学科，时间表达的领域特征就极其明显，时间跨度、专有时间名词等都与日常时间表达有很大的差异。

数值信息抽取研究涉及到对数值信息的词汇表达、句法特征的分析，表达方式不同，处理方式就会大相径庭，以时间类数值信息抽取为例，对于不同的表达方式，正则表达式的构造也会有所差异。

表4 不同领域时间数值信息表达差异

	时间数值信息
新闻	2018年3月2日 2018-02-13 2018.03.26 2018/6/11 ……
地球科学	末次盛冰期 早/中/晚/初更新世 2000Ma BP(a BP) ……
历史学	庆历2年 公元前334年 13世纪初 ……

(r"\d{4}[-/]\d{2}[-/]\d{2}") (1)

该正则表达式会匹配到“xxxx-xx-xx”以及“xxxx/xx/xx”等表达方式；

(r"(\d{4}-\d{1,2}-\d{1,2}\s\d{1,2}:\d{1,2})") (2)

该表达式会精确到具体时刻，几时几分；

([一二三四五六七八九十○]{4}年[一二三四五六七八九十○]{1,2}月[一二三四五六七八九十○]{1,2}日") (3)

该正则表达式就可以抽取出“xxxx年xx月xx日”形式的时间。

领域差异会在很大程度上影响到抽取策略的编写，因此需要本领域的特征词表、触发词词表、专业术语词表等作为辅助信息。

4.2 数值信息类型差异性对信息抽取的影响

不同类型的数值信息抽取策略多有不同。表5以地球科学领域的沙漠学数值信息类型为例，包含了地理信息、时间空间信息、气候信息、属性信息、生态信息等。沙漠学文献中涉及到的气候信息包括降水量、蒸发量、干燥度、风速、温度等内容，降水量、蒸发量一般都有标志性的数值和数值单位组成，而气候类型的数值信息抽取目标词汇则一般为“温湿、半湿润、寒冷干旱、半干旱、干冷多风”等；时空信息一类中，经纬度、海拔有特定的表达结构，而研究区域则相对弱规律化，例如“巴丹吉林沙漠位于(阿拉善高平原中西部，集中分布于弱水东岸的古日乃湖以东、宗乃山和雅布赖山以西，拐子湖以南，北大山以北的地区)”，括号内的内容为需要抽取的内容，涉及的词表类型、表达规则丰富多样，这都为数值信息的抽取增加了难度。

表5 沙漠学领域数值信息类别

地理信息	空间时间信息	气候信息	属性信息	生态信息
湖泊个数	经纬度	降水量	长度	植被盖度 归一化植被指数 ……
湖泊面积	海拔	蒸发量	宽度	
蓄水量	年代	干燥度	面积	
地下水位	区域	气候类型	厚度	
……	……	……	……	

4.3 相对数量关系表达增大了数值信息抽取难度

数值信息一般分为绝对数值信息和相对数值信息，绝对数值信息一般是确定的结果，而相对数值信息可能是一种倍数表达或者相对关系[29]。以时间为例，对于“2018年3月26

日”这种确定时间，正则表达式的抽取方法已经完全可以解决，而对于“2018年3月26日……三天后发生了……”，对于事件发生的真实时间“2018年3月26日”三天后，即“2018年3月29日”可能就需要复杂的技术手段来识别和计算。同样，对于其他表达数量关系的语境，例如“2017年河南省粮食产量为……，今年增长了……”，或是“某地区2013年降水量为……，今年的降水量是2013年的……倍”，这些相对的数量关系抽取都是数值信息抽取工作面临的挑战。

4.4 行文表达的复杂性对数值信息抽取的影响

对数值信息抽取而言，对无结构的自由文本的抽取是抽取工作面临的最大挑战。数值信息抽取需要呈现出完整的信息链条，包括主体、指标类型以及数值表达，包含多种类型信息，而在文本表达时这些必要的部分往往并不会有序出现在同一个句子单元，甚至处于不同的段落；再者，当文章中以指代形式对数值信息的主体对象进行表达时，也会涉及到对代词所指对象的识别；在文献中还存在句子成分缺失的情况，以主语缺失最为普遍。在数值信息抽取时常会涉及到主体识别，要分辨出该数值短语所表达的属性属于哪一实体及相互关系是十分困难的。同时，实体名称的判断与数值信息的识别又是相辅相成的，即命名实体的位置很可能会出现相应的数值信息表达。因此，要识别目标语句，并得到完整的信息链，是数值信息抽取需要解决的难题。

面对数值信息抽取的困境，各类技术方法也在不断革新。但现有的数值信息抽取技术仍然不能完全解决这些问题。数值信息抽取技术仍然面临诸多挑战[29]。

5 结 语

本文概述了数值信息抽取的基本概念和研究范畴，总结和对比了近年来中外数值信息抽取的研究成果，理清数值信息抽取的研究模式，对数值信息抽取技术的应用做了简单的归纳，并分析了中英文表达差异、领域差异性、数值类型多样、行文不规范等因素对数值信息抽取的影响。

数值信息抽取仍然面临着诸多难题，训练模板的全面程度限制了目标语句的识别效果，分句之间数值信息与实体的关系识别仍然很难解决，中文分词的准确度不够，以及抽取方法的可移植性不够强等。数值信息抽取对于科研具有重要的意义，在数据驱动研究范式的大趋势下，数值信息抽取技术是提供结构化、易获取的数值信息的有效手段；对于信息组织者而言，数值信息抽取技术是自动处理海量无序信息，支持数据库建设的重要工具，因此，虽然实现数值信息抽取的全自动化仍然十分困难，但是将其提高到可以解决部分现实问题的水平仍然是十分必要的工作。

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 程显毅, 朱倩, 王进.中文信息抽取原理及应用[M].北京: 科学出版社, 2010: 181 -182.
3 毋 菲. 数值信息的抽取方法研究[D].太原：山西大学, 2010.
4 Voorham J,Denig P. Computerized extraction of information on the quality of diabetes care from free text in electronic pa-tient records of general practitioners[J]. Journal of the American Medical Informatics Association:JAMIA, 2007, 14(3) :349 -354.
5 李云静. 基于石油领域本体的Web信息抽取技术研究[D].大庆：东北石油大学,2015.
6 吴 胜, 刘茂福, 胡慧君, 等. 中文文本中实体数值型关系无监督抽取方法[J]. 武汉大学学报:理学版, 2016, 62(6): 552-560.
7 Maiya A S, Visser D, Wan A.Mining Measured Information from Text[C]//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile. New York, USA: ACM, 2015.
8 Turchin A, Kolatkar N S, Grant R W, et al. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes[J]. Journal of the American Medical Informatics Association: JAMIA,2006,13(6):691-695.
9 Madaan A, Mittal A, Ramakrishnan G, et al.Numerical Relation Extraction with Minimal Supervision[C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence.USA: AAAI Press,2016: 2764-2771.
10 张远鹏,董建成,周慧玲,王 理,吴辉群,耿兴云. 基于HMM的H7N9事件中时间信息的抽取[J]. 中国数字医学,2015,10(10):23-26.
11 卢东标. 基于决策树的数据挖掘算法研究与应用[D].武汉:武汉理工大学,2008
12 周海波. 基于决策树的分类算法研究[D].兰州:兰州大学, 2009.
13 路明懿. 基于深度置信网络算法的作者信息抽取研究[D]. 长春：东北师范大学,2016.
14 万 静,涂 喆,冯 晓. 基于条件随机场的医药领域症状信息抽取[J]. 北京化工大学学报(自然科学版),2016,43(01):98-103.
15 Mooney R J, Bunescu R. Mining knowledge from text using information extraction[J]. ACM SIGKDD explorations newsletter, 2005, 7(1): 3-10.
16 侯伟涛,姬东鸿. 基于Bi-LSTM的医疗事件识别研究[J]. 计算机应用研究,2018,(7):1-2.
17 李 琴,伊晓玲,曹根牛.关于支持向量基中二次函数集VC维的研究[J].科技信息(学术研究),2007,(33):78-79.
18 郑红军,周 旭,毕笃彦.统计学习理论及支持向量机概述[J].现代电子技术,2003,(4):59-61.
（下转第124页）

的,可以激活共同的语境假设,并推理出交际意图;经过语言转码后,目的文化民众无法通过明示信息激活源文化的语境假设,不能获得正确的交际意图。这就要求信息传递者在忠实性、求效性原则下,对源文化的语境假设进行重构,完成正确交际。

如:但是,茶和酒并不是不可兼容的,既可以酒逢知己千杯少,也可以品茶品味品人生。——《在亚欧大陆架起一座友谊和合作之桥》

(Но это не означает, что чай и пиво «не уживаются» на одном столе: близкие друзья, встретившись после долгой разлуки, могут пить вино, и сколько бы они ни выпили, этого будет недостаточно, чтобы выразить связывающую их глубокую дружбу, — также они могут пить и чай, ведя долгие задушевные беседы о жизни.)[11]

在汉语文化的信息辖域中,酒代表的含义除了失意时可以"消愁"外,还可以在喜庆环境下助兴,与友人相聚时的畅快。而俄罗斯的酒文化中,是否饮酒、饮多少是个人的事情,和朋友、感情并没有多少关系。因此酒在这里代表的文化内涵已发生变化,这时就需要将没有明示的文化内涵补充进来。书中转码后的信息彰显的是友谊,强调的是"尽兴"而不是"多"。这里信息传递者放弃了源文化中"千杯"这个"多"的概念,灵活的转换为"尽兴",达到了激活共同语境假设的目的,遵循了适切性原则。

4 结 语

本文从信息生态的视角考察跨文化语言信息传播,是在跨学科研究上进行的探索,是从新的角度审视跨文化语言信息的本质和过程。在信息生态这个更加强调动态和谐、更加广阔和更加强调整体关联的背景当中考察,就是希望能够揭示出更全面、更深层的跨文化语言信息传播的活动机理与规律。本文在论证二者在理论上接缘的可行性基础上,参照信息生态位的三个优化原则,对信息生态指导跨文化语言信息传播进行了实证分析,为跨文化语言信息传播的研究范围和领域开辟了新的视角、提供了新的思路。

参考文献

1 董丽梅,宋 微,戴 磊. 宏观信息生态系统的概念、构成与功能研究[J].情报科学,2014,(8):27-31.
2 肖希明.信息生态理论与公共数字文化资源整合[J].图书馆建设,2014,(3):1-16.
3 谢 刚,冯 缨,田红云,李文鹏. 信息生态视角下移动网络隐私问题及防治措施[J].情报理论与实践,2015,(8):21-26.
4 张向先,郑 絜,靖继鹏,我国信息生态学研究现状综述[J].情报科学,2008, (10):1589-1600.
5 陈海涛,马艳丽,陈 博.信息生态系统演化研究回顾[J].情报科学,2014, (2):157-161.
6 富金鑫,孙笑宇,季忠洋.信息生态研究的整体演化视角[J].情报科学,2015,(11):9-13.
7 张 蕊.再论认知语言学与批评话语分析的融合——以"侧重"识解操作为例[J].外语研究, 2015, (6):34-41.
8 高 远,李福印. 罗纳德·兰艾克认知语法十讲[M].北京:外语教学与研究出版社,2007:17.
9 韩子静.信息生态系统初探[J].图书情报工作,2008年增刊(2):230-234.
10 周承聪, 娄策群, 杨小溪. 信息服务机构信息生态位优化原则与方法[J]. 情报科学,2011, (4):596-599.
11 习近平.习近平谈治国理政(俄文版)[M].北京:外文出版社,2014:383-384,388.
12 Ronald W.Langacker,王义娜,李亚培.认知域的种类[J].外国语言文学研究,2008, (3):11-15.

(责任编辑:徐 波)

(上接第45页)

19 哈明虎,田景峰,张植明.基于复随机样本的结构风险最小化原则[J].计算机研究与发展,2009,46(11):1907-1916.
20 张 妍. 基于隐马尔可夫模型的中文信息抽取算法研究[D].鞍山:辽宁科技大学,2014.
21 徐延勇,郭忠伟,周献中.基于最大熵方法的统计语言模型[J].计算机工程与应用,2002(05):53-55,121.
22 李 军,费川云.地球空间数据集成研究概况[J].地理科学进展,2000,(3):203-211.
23 葛全胜,郑景云,郝志新.过去2000年亚洲气候变化(PAGES-Asia2k)集成研究进展及展望[J].地理学报,2015,70(03):355-363.
24 葛全胜,陈泮勤,张雪芹.全球变化的集成研究[J].地球科学进展,2000,(4):461-466.
25 叶 明,谷晨霞."大数据"时代决策支持系统新发展[J].信息安全与技术,2013,4(8):6-8.
26 张文江. 地质灾害数据集成关键技术研究[D].成都:成都理工大学,2013.
27 张 宁. 面向数值问题的答案抽取与生成[D].沈阳:沈阳航空航天大学,2018.
28 余 丽,陆 锋,张恒才. 网络文本蕴涵地理信息抽取:研究进展与展望[J]. 地球信息科学学报,2015,17(2):127-134.
29 郭少卿,乐小虬.科技论文中数值指标实际取值识别[J].数据分析与知识发现,2018,2(1):21-28.

(责任编辑:毛秀梅)