DOI:10.13530/j.cnki.jlis.140006

数值信息抽取研究进展综述

吴 超 郑彦宁 化柏林

摘 要 通过对数值信息抽取文献的调研,先从文献类型、学科领域、高频关键词三个方面进行定量分析,从抽取数据源、抽取对象、抽取方法与技术、结果评价和应用等方面对当前数值信息抽取研究进行了梳理和总结。研究发现当前对于数值信息抽取的研究具有五个特点:抽取数据源以新闻语料、Web 网页为主,抽取对象以基数类数值信息和数量类数值信息为主,抽取方法以基于规则的方法为主,抽取结果评价指标比较单一,但应用领域较为广泛。图 4。表 3。参考文献 56。

关键词 数值信息 数值知识元 数值信息抽取 命名实体识别

分类号 G350

Numerical Information Extraction: A Review of Research

Wu Chao, Zheng Yanning & Hua Bolin

ABSTRACT This paper first makes a quantitative analysis on the documents of numerical information extraction from three aspects: document type, subject area and high frequency keywords. Then the research context is summarized from four aspects: data source type, object for extraction, extraction method and technique, result evaluation and corresponding application. Our findings are as follows: news corpus and web pages are the main data sources; cardinal numbers and quantitative phrases are the main objects for extraction; extraction method and technique are mainly rule-based and the result evaluation indicators are relatively simple but have a wide scope for application. 4 figs. 3 tabs. 56 refs.

KEY WORDS Numerical information. Numeric knowledge element. Numerical information extraction. Named entity recognition.

1 引言

信息抽取就是从给定的自然语言文本中抽取预先制定的信息,并将其形成结构化的数据[1]。本文的研究对象是数值信息抽取,指对给定文档集中的数值型信息进行抽取。

数值信息具有较强的领域相关性,在不同的研究领域、应用场景下,数值信息的具体对象各有差异。如在时间短语识别或时间序列挖掘的研究中,时间是数值信息抽取的具体研究对象。在名量短语和动量短语的识别研究中,数量短语则是具体的研究对象。因此,在对数值信息进行抽取前需要对该领域中常见数值信息的类型进行归纳,并定义要抽取的数值信息。

在中文信息处理的研究中,对于数值信息的处理是必不可少的,是基础研究方向之一[2]。它的研究成果将直接影响到文本信息自动化处理的深层次研究,它是信息抽取、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言处理技

通讯作者:化柏林,Email: huabolin@istic.ac.cn

术及应用的重要基础[3]。

在当今海量信息环境下,现代情报工作中的信息采集任务基本都交由计算机自动完成,如商业竞争情报监测、科技动态监测、武器装备监测等,这些工作所关注的一个重点是监测对象的相关数值信息,获取这些信息是情报研究的基础。因此数值信息抽取在其中发挥着重要作用。此外,科技文献数据库、专利数据库是高质量的信息源,目前研究针对的是标题、作者、语种等外部特征元数据。对其内容进行细分描述的研究很少,可以使用数值信息抽取技术深入研究文献中含有丰富数值信息的内容元数据[4]。

虽然关于信息抽取的综述论文已有不少,但是还没有专门针对数值信息抽取的综述论文。图书情报领域中已有很多综述与述评[5-7],尚缺乏结合定量分析的方法进行综述与述评。根据定量分析可找出热点文献与重要文献,分析核心作者与重要研究机构,统计研究热点等,然后通过分类归纳把相关研究划分为相关流派与不同技术路线等进行定性述评[8]。因此,本文综合运用定量与定性的方法,从多个角度对数值信息抽取的研究现状进行综述,其体系框架如图 1 所示。

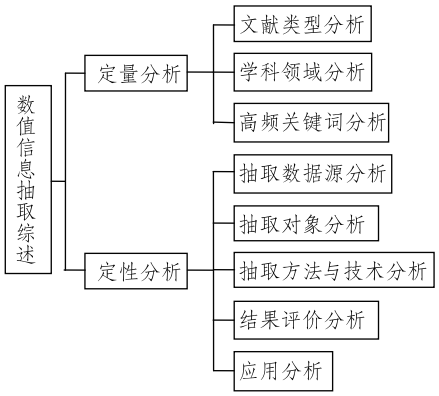

图 1 本文综述体系框架

2 相关文献定量分析

文献[9]认为,有些人做主题研究时往往只使用万方数据或者中国知网数据,难免会导致数据不完备的情况,特别是在计量分析类论文或综述类论文中,这一问题尤为明显。计量分析类论文要尽量使用多源数据,例如,万方数据与中国知网数据的融合。因此,本文同时使用了这两个数据库,利用主题检索功能查找同时含有“信息”和“抽取”的文献,两数据库中检索结果分别为 14,384 篇和 12,864 篇。然后利用在结果中检索这一功能,查找主题中同时含有“数值”的文献,结果两个数据库中分别有 242 篇和 230 篇,经文献筛读后其中只有两篇与本文主题相关,即毋菲的论文《数值信息的抽取方法研究》和《基于决策树的中文事件论元值的抽取》。对于外国文献,在 Web of Science 中使用检索式“information AND extract*”进行标题检索,检索结果有 2,839 条。修改检索式为“numeric* AND information AND extract*”,同样进行标题检索,检索结果大为减少,只有八条,经过筛选后只有两篇和主题相关,并且这两篇文章的第一作者都是 Murata M。从上面的检索结果来看,关于数值信息抽取的研究并没有引起足够的重视。

数值信息具有较强的领域相关性,在不同的研究领域、应用场景下,数值信息的具体对象各有差异。所以,我们可以调整检索策略,结合其他与数值信息相关的更为具体的检索词进行检索,如“数值知识元”、“数词识别”、“数量短语识别”等。另外,在信息抽取领域,命名实体识别是其首要和基础的任务。从狭义上说,命名实体是指现实世界中的具体或抽象的实体,如人、组织、公司、地点等,通常用唯一标识符(专有名称)表示,如人名、组织名、公司名、地名等。从广义上讲,命名实体还可以包含时间、货币以及各种数字等。至于命名实体识别的确切含义,只能根据具体应用来确定。比如,在具体应用中,可能需要把身份证号、电子信箱地址、电话号码、编号、URL 网址等作为命名实体[10]。因此,广义的命名实体中所包含的时间、货币以及各种数字等数值信息,也是本文的研究内容。在以命名实体为研究对象的文献中就很可能包含对数值信息的探讨。所以我们也把“命名实体识别”作为一个检索词。同时分别以抽取对象为检索词扩充检索范围,如“价格抽取”、“时间抽取”等。

但是检索出来的很多文献并没有介绍数值信息抽取内容,因此还需要人工逐篇阅读判断,最后通过对检索过程的调整以及检索结果的人工筛读后,得到 49 篇样本文献,接下来我们对这些样本文献进行了简单的计量分析。由于受文献获取所限,定量分析结论仅供参考。

2.1 文献类型分析

我们把检索得来的文献大致划分为期刊论文、会议论文、学位论文三种类型。据此对样本文献进行类型统计(见图 2、表 1)。从图、表中可见,期刊论文最多,达到 25 篇,所占比例为 51%,会议论文有 13 篇,比例为 27%,学位论文有 11 篇,比例为 22%。说明对于数值信息抽取的研究成果大都发表于期刊和会议当中,比较零散、浅显,缺少系统的研究。

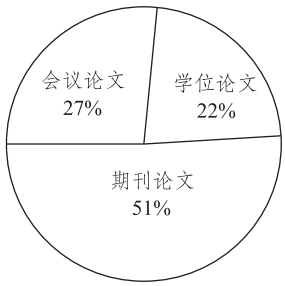

图 2 文献类型分布

表 1 不同类型的文献数量

文献类型	期刊论文	会议论文	学位论文
数量(篇)	25	13	11

2.2 论文学科领域分析

根据论文刊载期刊、中图分类号、作者专业或研究方向、论文标题、会议名称等信息综合判断,人工标引出样本文献所属学科并进行统计,结果如图 3、表 2 所示。不难发现,对于数值信息抽取的研究大多属于计算机科学、情报学、计算语言学三个学科,特别是计算机学科领域,论文数量有 28 篇,在全部样本文献中所占比例接近五分之三。另外情报学和计算语言学也占据不小的比例,这表明对于数值信息抽取的研究需要交叉学科知识。分别对这三个学科领域的样本文献的标题进行归纳分析后发现,计算机学科领域的研究重点为命名实体识别、信息抽取。情报学领域的研究重点为知识元抽取,其中温有奎有三篇相关论文。计算语言学领域的研究重点为数量短语识别和机器翻译。

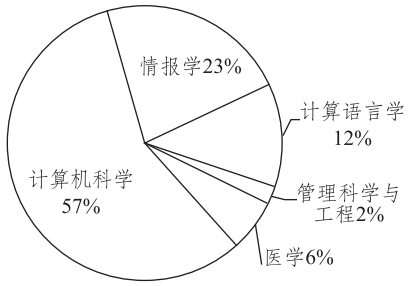

图 3 论文学科领域分布

表 2 各学科领域论文数量

学科类型	计算机科学	情报学	计算语言学	医学	管理科学与工程
数量(篇)	28	11	6	3	1

2.3 高频关键词分析

对样本文献的关键词进行统计分析,其中英文文献关键词翻译为中文再统计。表 3 列出了对这些关键词进行初步归类后的高频关键词(频次≥3)。关键词可以传达论文主题,概括论文核心研究内容。高频关键词可以反映研究热点。通过表 3

表 3 前 9 名高频关键词(频次≥3)

关键词	频次
命名实体识别	13
信息抽取	12
自然语言处理	6
数量短语	5
知识元	4
数据抽取	4
时间短语识别	4
数值信息	4
机器翻译	3

可以发现,在数值信息抽取研究中研究较多的对象有数量短语、知识元、时间短语、数值信息。而且对数值信息抽取的研究多出现在命名实体识别、信息抽取、自然语言处理、数据抽取、机器翻译等研究领域当中。

3 相关研究综述

完成定量分析后,下面从数据源、对象类型、抽取方法与技术、结果评价和应用五个角度梳理和总结数值信息抽取的研究现状。

3.1 抽取数据源分析

数据挖掘是数据库中知识发现的核心步骤[11],主要是对数据库中结构化数据信息实现关联分析、分类、聚类、时序演变分析、预测等挖掘功能,以提取隐含的、未知的、有用的信息和模式的过程[12]。和数据挖掘不同,本文所指的数值信息抽取的数据源是文本,特别是非结构化或半结构化文本,因此难度更大,更有意义。

非结构化文本又称自由文本,因为它几乎没有结构信息,是自然语言自由表达的文本。在数值信息抽取领域,常见的有报纸、以新闻文章为主的测评会议提供的数据集、图书。文献[13,14]利用1998年1月份《人民日报》语料作为实验语料库,完成了包括数值信息在内的多种中文命名实体的识别。文献[15,16]设计了一个半自动化系统,使用1998年和1999年日本《每日新闻》的所有文章做了数值信息的抽取实验。文献[17]使用CWMT新闻语料和IWSL09评测语料。文献[18]则使用了一些新闻报道、两本儿童故事书和一本历史书作为文档数据集。

半结构化文本相较于自由文本处理起来容易些,数据的获取也较容易。Web网页就是一种典型的半结构化文本,而且网络资源非常丰富且易于获取。因此,有不少研究所用数据源为Web网页,如文献[19-23]。除了普通的Web网页外,年鉴作为一种半结构化文本,富含大量社会经济文化发展的数值信息,也是数值信息抽取的重要对象。肖洪和薛德军详细描述了从海量年鉴文本中抽取数值信息的基本流程和各主要环节的算法[24]。温有奎设计了一套抽取软件,对年鉴中的数值信息进行了抽取实验[25-27]。另外,文献[28-30]发现在科研过程中产生的科学数据,很多都保存成文本格式,以一种半结构化方式存储,这种保存方式有其特有的方便、快捷等优势。而在科学数据的处理分析阶段,需要将其转换成结构化数据。因此作者设计了一个系统用于抽取半结构化的科学数据,以利于科学数据的共享与标准化。

3.2 抽取对象类型分析

数值信息抽取的对象是数值信息。在中文信息处理中,所要处理的词除了一般意义上的名词、动词、形容词之外,还有一类常见、重要、特征明显的词,即主要由数字或数词组成的具有特定意义的语言表达形式,我们称之为数值信息。

3.2.1 数值信息的特点

通过对数值信息表达形式的分析,可以发现数值信息具有以下两个特点:

(1)与数值信息直接相关的词不多。简单地对数值信息进行分析即可发现,任何数值信息使用的词主要就三种,数字、量词、符号。数字有0、1、2、…、9等,量词有年、月、日、时、分、美元、元、吨等,符号有\、-、¥等。它们数量有限,在全部词汇中所占比例很低。

(2)数值信息表达形式多种多样。尽管组成数值信息的词不多,但是其搭配和组成方式多样,不同的构成方式带来不同类型的数值信息,如时间、日期、货币、电话号码、产品型号,等等。就连含义相同的数值信息的表达方式都可以不一样,如2013/3/15,2013-03-15和15/03/2013就具有相同的含义。有人通过对网络上大量语料的提取分析,发现数值信息大致有570多种语言表达形式[23]。

3.2.2 数值信息的类型

我们在文献调研和个人研究的基础上,把数值信息分为以下三个大类,这三个大类自下向上层层递进,所含信息量逐渐增多,复杂度也逐步提升,如图4所示。

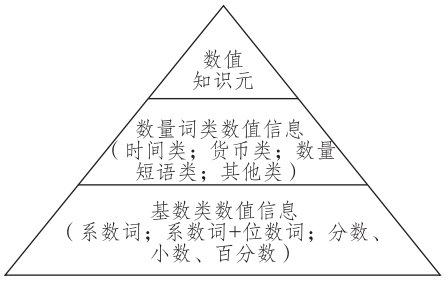

图4 数值信息分类

(1)基数类数值信息

第一大类最简单,是相对“单纯”的数字,没有与量词结合,叫做基数类数值信息。这类还可以分成三个小类,如单一的系数词100、365、23,000等。还可以包含位数词,位数词有十、百、千、万、亿,如10万、13亿等。另外,分数、小数、百分数也算一类,如4/5、3.14、80%等。这类数值信息构成简单,所含信息量较少,数值信息本身几乎没有实际意义。但是它是另外两类数值信息研究的基础,而且在日常生活中也有大量应用。文献[31,32]较早地研究了这类数值信息的构成规则,并给出了特征解析式。文献[17,33]通过对汉语和英语中这类数值信息的细致研究,高质量地实现了对它的机器翻译。文献[3,14]则是通过对不同应用背景下该类数值信息的研究,制定匹配规则,实现了电话号码、银行卡号,甚至是身份证号码的抽取。

(2)数量类数值信息

第二大类为数量类数值信息,是由第一类数值信息加上量词或者特定符号组成的。经过细分,我们把这类数值信息分为四个小类。

①时间类数值信息

时间类数值信息由数字加上时间量词或对应符号构成,如2013年3月15日、2013/3/15、2013-03-15、20点58分、20:58等。时间类数值信息是一种重要的信息,在文本信息中的重要程度仅次于专有名词。通过时间信息我们可以了解一个事件的开始、过程、结束以及事件发生的频率,把握重要的时间节点有助于对事件的全面掌控。美国国家标准技术研究院(NIST)组织了自动内容抽取(Automatic Content Extraction, ACE)评测会议,ACE会议的一项基本任务就是时间表达识别与标准化(Time Expression Recognition and Normalization, TERN),对时间表达式的评测给出了详细的定义与要求。对时间类数值信息进行研究的文献较多,如文献[13,17,18,33-37]。

②货币类数值信息

货币类数值信息由数字加上货币量词或符号构成,如200美元、20万元、￥100、$80等。货币类数值信息构造相对简单,变化不多,特征词明显,因此抽取起来相对容易。文献[3]从语料中获取货币类的部分特征词,然后经过扩展,最终得到含有84个特征词的集合,并总结出12条与货币类数值信息相关的规则。文献[13]对此类数值信息也做过研究,不过没有文献[3]介绍得详细。文献[38,39]对网页中商品价格信息的抽取进行了研究。

③数量词(也称数量短语)类数值信息

数量词类数值信息由数字加上计量量词或对应符号构成,如36吨、1600篇、40米、27℃等。数量词是自然语言文本中出现频率很高的一类短语,对它的处理在中文信息处理中是必不可少的,由于其构成相对简单,一般以数词开头,以量词结尾,是一个较稳定的结构,在语句处理中可以建立独立的识别模块,对数量短语进行优先处理[2]。文献[40]提出了一种不依赖于分词的数量词的识别方法,与采用分词技术的识别过程相比,显著地提高了处理效率。文献[41]根据数词和量词的不同表现形式,对其做了详尽的分析和分类,建立了数词库和量词库,实现了数量词的识别和抽取,最后抽取实验获得的召回率和准确率都达到80%以上。文献[15]设计了一个半自动化的数值信息和命名实体抽取系统,从报纸中抽取日本的城市天气信息,包括气温、气压、湿度、风速等数量词类数值信息,并能够自动绘制多种图形,效果不错。文献[13,42,43]也对数量词类数值信息的识别与抽取进行了研究。

④其他类数值信息

如果一种数量类数值信息不属于任何一类则归为其他类。如京A·80293、歼-10、波音737等。这类数值信息比较杂,一般由汉字、字母、数字、符号等组成,在其应用情景下代表一定含义,因此对

于此类数值信息的抽取,必须结合相应情景进行个性化分析。文献[14]介绍了国内汽车牌照号码编码规则不统一的问题,认为无法找到100%精确且永久适用的模型算法,不过作者经过仔细研究牌照的组成模式,仍然提出了一个比较复杂但暂时适用目前牌照规律的提取算法。文献[19]通过分析总结Web上出现的航空产品型号数据,然后依据领域知识,发现了中、美、俄三国军用飞机的命名规律性和特点,据此制定了匹配模式,实现对此类数值信息的抽取。

(3)数值知识元

第三大类可以看作是在第二大类的基础上加上句子其他组成成分所形成的,称为数值知识元。它是指含有数值信息的较完整描述客观事实的独立单元。如"2001年,中国电子信息产业,全年实现销售收入8,237亿元"。温有奎是国内较早研究数值知识元的学者,他认为实现数值知识元的抽取是建立知识元库的先导和基石。他研究了数值知识元的特征和抽取规则,开发出一套数值知识元抽取软件,用于从年鉴、网页文本中自动抽取数值知识元并将抽取结果自动存入库中[20,25-27]。文献[24]详细描述了从海量年鉴文本中抽取宏观数值知识元的基本流程和各主要环节的算法,首先识别出包含数值的句群,然后依据给定模式从各句群中抽取数值知识元,作者认为在特定领域内将数值知识元抽取提高到实用水平是可行的。

3.3 抽取方法与技术分析

在信息抽取领域,存在两大传统的技术方法,分别为基于规则的方法和基于统计的方法。虽然目前对这两种方法孰优孰劣并无定论,但是数值信息的抽取技术都采用了基于规则的方法。这自然与上文中指出的数值信息的特点有关,由于数值信息的相关词不多,表达形式尽管多样但在特定领域中有限,且构成比较简单,有较为明显的规律性,因此通常利用基于规则的方法对其进行识别和抽取。

使用基于规则的方法,一般需要先建立特征词库和规则库。特证词在某些研究中也被称为关键词、触发词。特征词库具有领域限制性,根据不同类型的数值信息需要建立不同的特征词库。如文献[3]为了抽取货币类数值信息建立了含有84个特征词的词库。文献[41]为了抽取数量词类数值信息分别建立了数词库和量词库。当然,特征词库的建立不是必须的。我们可以不使用特征词库,只需要规则库也可以达到抽取某些种类数值信息的目的。比如对于数字类的数值信息,它们只含有数字,所以可以认为不需要特征词库。文献[3,14]只使用了正则表达式构建的匹配规则就实现了对电话号码、银行卡号的抽取。

既然使用的是基于规则的方法,规则库的构建就是必须的。规则库有语形规则库和语法规则库之别。根据规则库的不同,基于规则的方法可以细分为基于语形规则的方法和基于语法规则的方法,两者的区别在于后者要进行语法分析,需要标注词性。语形规则指的是特定数、词等文字的搭配规则,语法规则指的是词性的搭配规则。在数值信息抽取中,绝大部分的研究单独使用了语形规则[14,17,34,38,41,44]。由于正则表达式就是描述文本规则的代码,很适用于字符串匹配,它功能强大,并且许多程序设计语言都支持正则表达式。因此,对于语形规则,研究人员几乎都是使用正则表达式来表示。不过也有其他的规则表示方法,文献[40]由于使用了谢菲尔德大学研究开发的GATE系统,就使用了它提供的JAPE规则语法来表示规则。文献[23]使用了陈肇雄在1992年提出的SC文法规则[45]来建立规则的表示形式。另外还有一部分研究[3,20]既使用了语形规则,又使用了语法规则,将两种规则结合起来使用,提高抽取的准确性。

使用语法规则需要先进行词法分析,有部分研究因此采用了自然语言处理技术,对文本进行了分词和词性标注。文献[37]的预处理工作包括分词和词性标注,然后利用词语之间的关系建立基于语法的抽取规则,实现时间短语的抽取。文献[46]更进一步,采用了自然语言处理技术中的词法分析、浅层句法分析技术,然后将浅层句法特征、构词特征、上下文特征等多种特征集分别引入条件随机场模型和半监督机器学习方法中,研究从生物医学文献中识别生物医学命名实体的问题。

3.4 抽取结果评价分析

对数值信息抽取的结果进行评价不同于对抽取系统的评价,对系统的评价可以有处理速度指标、能耗指标、可移植性指标、抽取质量指标等一系列指标。对抽取结果的评价是对系统评价中抽取质量评价指标的一部分。对结果进行评价可比较各种抽取方法或规则的优劣,有利于指导、改进现有的方法。

目前对抽取结果评价的指标比较单一。仅仅通过仿效信息检索中的查全率和查准率,引入抽全率(P)和抽准率(R)及其综合指标F值。而使用这些指标并不能全面反映抽取结果的好坏。比如抽取出的信息一部分相关,一部分不相关;或者没抽取出来的信息有一部分是相关的。这些情况就不能简单地用抽全率和抽准率来进行评价[47]。

要对抽取结果进行评价分析,必须利用抽取规则或抽取系统在测试集上进行抽取实验。测试集也是抽取数据源,不同的文章选择了不同的测试集类型和规模。文献[19]用了四个网站共500个相关网页进行了测试。文献[40,42]随机抽取了1998年1月份《人民日报》中一万字的新闻语料进行测试。文献[41]对2M多的真实语料进行了抽取。

同样只是使用P、R、F三个指标,但由于对训练集和测试集的不同处理,还可以优化对抽取结果的评价与分析。值得借鉴的是,文献[48]分别使用50篇、100篇、200篇网页新闻作为训练语料,而测试语料则固定使用100篇。通过三轮测试得到的召回率和准确率,探讨了训练语料的规模对于抽取结果的影响。使用不同规模的训练语料进行对比测试,更容易得出有价值的结论,对结果的评价也更加准确而全面。而文献[37]根据训练集和测试集的不同关系,分别进行了封闭测试和开放测试。封闭测试就是用训练集的一部分作为测试集;而开放测试则是将数据集中的一部分作为测试集,剩下的作为训练集。可以通过对比封闭测试和开放测试的结果来评判规则的适用性、系统的可移植性,因此这种选择不同测试集的实验方法也具有积极意义。

3.5 应用分析

数值信息抽取应用范围非常广泛,包括医学领域、商业领域、军事领域,也可应用于机器翻译系统、自动问答系统、构建数值知识元库等,它在各个细分的应用领域均取得了不错的效果。对数值信息抽取的实际应用情况进行分析有助于研究人员完整了解该领域的研究现状,发掘研究价值,明确研究方向,探索未来的发展前景。

Voorham J和Denig P研究电子病历文本中糖尿病治疗相关的数值临床数据,包括13种测量数值,如血压、身高、体重、糖化血红蛋白等,设计了一个数值触发系统,在两种常用的电子病历系统中进行了实验,结果平均有效率为89.8%,抽取100个病人的相关数据平均需要7.8分钟,而手工收集需要10分钟[49]。Turchin A等人同样针对电子病历,利用正则表达式从医师记录文本中抽取血压值,得到93.2%的准确率[50]。文献[51]从MEDLINE中选取与蛋白质相关的文章摘要,利用规则抽取其中的蛋白质名,在不区分蛋白质名是已知还是新定义的情况下,获得98.84%的召回率和94.70%的准确率。

文献[52]和[53]的研究有望应用于商业领域进行趋势预测,他们从新闻报纸文章或博客中抽取趋势信息,例如时间、价格、销量、市场占有率、内阁支持率等各种数值信息,然后以时间为横轴,其他信息为纵轴自动做出趋势图,文献[53]在新闻报纸文章中得到的准确率和召回率分别为31.3%和6.3%,在博客语料中获得44.8%的准确率和60.3%的召回率。文献[38]和[39]对网页中商品信息的研究则可以应用于价格抽取,其中文献[39]对于商品列表页抽取的准确率达到98.7%,商品详细信息页面抽取的准确率达到92.8%,最后可以形成比价系统。

数值信息抽取同样可以应用于军事领域,文献[54]将基于规则与基于统计的方法结合起来识别防务新闻报道中国防领域命名实体,如武器装备型号等,最后实验的F值达到89%以上。另外有文献针对军事网站中航空产品型号进行了抽取研究,获得了95.6%的召回率和90.8%的准确率[19],温有奎

同样将其数值知识元挖掘系统应用于军事领域[20]。

文献[17]、[23]、[33]研究了机器翻译系统中数值信息的翻译问题,其中[17]实现了汉英时间和数字双向翻译的工具,成功应用于机器翻译系统,在汉语新闻语料中取得93.93%的准确率,在英语语料中得到95.40%的准确率。文献[23]研究了网页实时机器翻译系统中有关数字和数词的语料库,提出了基于动态模版和知识库的识别处理模型,从实践效果看,较好地达到了预期目标。文献[33]研究了汉语数词的翻译问题,其汉英机器翻译系统中数词的识别和翻译达到99.6%的准确率。

在自动问答系统中也可以找到有关数值信息抽取的研究,如文献[55]研究了问答系统中需要给出数值答案的问题,取得91.97%的准确率,该结果对于问答系统的研究具有积极意义。

对于数值知识元的抽取可以构建数值知识元库,进而提供参考咨询与决策支持服务。文献[24]和温有奎的多篇论文著作[25-27]即对年鉴中的数值知识元进行了相关研究,认为在特定领域内将知识元抽取提高到实用水平是可行的。

4 结论与讨论

通过上述分析,我们发现当前对于数值信息抽取的研究具有以下五个特点:

4.1 抽取数据源多以新闻语料、Web网页为主

通过对数据源的分析,我们发现,目前的研究中数据源多以新闻语料、Web网页为主。原因可能有以下几点:首先,新闻基本构成要素较固定、完备和准确,篇幅较短,语言精练;其次,测评会议所用语料多半都是新闻语料,这也促进人们在信息抽取技术研究中对新闻语料的应用;再次,已经标注过的语料库也大多为新闻类型,同样促使人们对该类语料的研究。另外,Web网页具有半结构化特点,降低了研究的难度,而且较为容易获取。

4.2 抽取对象多以基数类数值信息和数量类数值信息为主

抽取对象多以第一类——基数类数值信息和第二类——数量类数值信息为主,以第三类数值信息,也就是数值知识元为抽取对象的研究不多。数值知识元是一个可以表达完整意思的信息单元,因而对其抽取更有意义。对期刊论文中的数值知识元的抽取研究,有利于打破传统的以篇章为单位的组织方式,真正实现文献从知识单元上的组织、管理和利用,是信息服务转向知识服务的基础。虽然对于期刊论文中数值知识元抽取的研究不多,但是可喜的是市场上已经出现了相关应用。中国知网的数字搜索以及科技数值知识元库平台已经可以提供数字知识和科技数值知识元的搜索服务[56]。

4.3 抽取方法以基于规则的方法为主

抽取方法以基于规则的方法为主,少有基于统计的方法,另外自然语言处理技术在数值信息抽取中起到重要作用。原因是数值信息“天生”特征明显,非常适合用基于规则的方法。而基于统计的方法需要大规模标注语料,目前并没有这种语料库。因此在数值信息抽取领域中,基于规则的方法占绝对优势。但并不是所有的自然语言现象都可以用确定性的规则来刻画,而且基于规则的方法可移植性较弱。随着自然语言处理技术的快速发展和大规模语料库的建立、完善,我们也许可以考虑将基于统计的方法引入数值信息的抽取研究中,将两种方法有机结合,取长补短,或许可以进一步改善抽取效果。

4.4 抽取结果评价指标和方式单一

这是信息抽取领域的共性问题。面对这一问题,我们可以考虑从多个角度对抽取结果进行评价,丰富评价指标,构建全面的评价体系。例如对抽取结果不是简单的“是非”判断,而是借鉴调查研究中常用的李克特量表法,采用3分表示正确、1分表示部分正确、0分表示不正确的评分方法,同时对每个结果由两人分别评分,如果两人所给分数

不一致,则由第三个人重新判断。最后将抽取结果所得分数与理想情况下的最高分数相比,换算成百分制形式,80 分以上表示优秀,60~80 分表示一般,60 分以下表示较差。另外,对实验程序进行创新设计,通过得到不同的实验结果来进行评价,也可以改善评价方式单一的问题。比如选择不同规模的测试数据或者分别进行开放测试和封闭测试等。

4.5 应用领域较为广泛

数值信息抽取具有较高的实用价值,而且较为容易应用于实践,取得的实际效果也不错。但是数值信息抽取的应用不限于综述中指出的几类,更多的应用领域与应用系统还有待研究人员进一步探索与发掘。

数值信息抽取的研究具有较高的实用价值,虽然目前已经出现了一些相关的数值库、事实库等应用系统,但是仍然缺少较为系统的研究。由于数值信息抽取属于跨学科领域,对其研究需要用到多种学科知识,因此,最好是由具有不同知识背景的人员进行团队合作研究,可以考虑分别从数据源、抽取对象、抽取技术与方法、结果评价以及应用等方面找到切入点。

参考文献

[1] 郑彦宁,邓擘.信息抽取技术在情报学中的应用分析[J].情报理论与实践,2008,31(5):769-772.(Zheng Yanning, Deng Bo.Analysis of information extraction technology in information science[J]. Information Studies: Theory & Application, 2008, 31(5): 769-772.)

[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.(Cheng Xianyi,Zhu Qian,Wang Jin.The principle and application of Chinese information extraction[M].Beijing: Science Press, 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.(Wu Fei. Research on value extraction from Chinese Text [D].Taiyuan: Shanxi University, 2010.)

[4] 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. (Zheng Yanning,Hua Bolin. Application of the sentence-level knowledge extraction in information science[J]. Information Studies: Theory & Application,2011,34(12):1-4.)

[5] 陆伟,周红霞,张晓娟.查询意图研究综述[J].中国图书馆学报,2013,39(1):100-111.(Lu Wei,Zhou Hongxia,Zhang Xiaojuan.Review of research on query intent[J]. Journal of Library Science in China, 2013,39(1): 100-111.)

[6] 王芳,史海燕.国外 Web Archive 研究与实践进展[J]. 中国图书馆学报,2013,39(2):36-45.(Wang Fang,Shi Haiyan. Progress of foreign research and practice in web archive[J]. Journal of Library Science in China, 2013,39(2): 36-45.)

[7] 吴丹,邱瑾.国外协同信息检索行为研究述评[J].中国图书馆学报,2012,38(6):100-110.(Wu Dan,Qiu Jin.A review on foreign studies of collaborative information seeking behavior[J]. Journal of Library Science in China, 2012,38(6): 100-110.)

[8] 化柏林,武夷山.文献综述,标"新"立"异"[J].情报学报,2013,32(4):337. (Hua Bolin,Wu Yishan. Literature review, seek to be different [J]. Journal of the China Society for Scientific and Technical Information,2013,32(4): 337.)

[9] 化柏林,武夷山.多"源"信息需要多"方"融合[J].情报学报,2013,32(3):225. (Hua Bolin,Wu Yishan.Multi-source information need multi fusion method[J]. Journal of the China Society for Scientific and Technical Information,2013,32(3): 225.)

[10] 李保利,陈玉忠,俞士汶.信息抽取研究综述[J].计算机工程与应用,2003,39(10):1-5,66. (Li Baoli, Chen Yuzhong, Yu Shiwen. Research on information extraction: A survey[J].Computer Engineering and Applications, 2003,

39(10): 1-5,66.)

[11] Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases[J]. AI magazine,1996, 17(3): 37-54.

[12] Hard D J. Principles of data mining[J]. Drug Safety, 2007,30(7): 621-622.

[13] 史海峰.基于 CRF 的中文命名实体识别研究[D].苏州: 苏州大学, 2010.(Shi Haifeng. Study on CRF-based Chinese named entity recognition[D].Suzhou: Soochow University,2010)

[14] 霍焰.基于非结构化文档数据的抽取与分析系统的信息抽取[D].天津: 天津大学, 2011.(Huo Yan. The information extraction of unstructured document extraction and analysis[D].Tianjin: Tianjin University,2011.)

[15] Murata M, Qing Ma, Torisawa K,et al. Extraction and visualization of numerical and named entity information from a large number of documents[C]// IEEE NLPKE-08, 2008: 1-8.

[16] Murata M, Shirado T, Torisawa K, et al. Extraction and visualization of numerical and named entity information from a very large number of documents using natural language processing[J].International Journal of Innovative Computing, Information and Control, 2010,6(3B): 1549-1568.

[17] 翟飞飞,夏睿,周玉,等.汉英双向时间和数字命名实体的识别与翻译系统[C]//第五届全国机器翻译研讨会,南京,2009:172-179.(Zhai Feifei, Xia Rui, Zhou Yu, et al.An approach to recognizing and translating Chinese & English time and number named entities[C]//The 5th China Workshop on Machine Translation, Nanjing, China, 2009: 172-179.)

[18] Küük D, Yazici A. A hybrid named entity recognizer for Turkish[J].Expert Systems with Applications, 2012, 39(3): 2733-2742.

[19] 袁利华.基于本体的 Web 航空产品型号信息抽取技术研究[D].南京:南京航空航天大学, 2009.(Yuan Lihua, Study on web-oriented aviation products model information extraction technology based on ontology[D].Nanjing: Nanjing University of Aeronautics and Astronautics, 2009.)

[20] 温有奎,孙明,温浩,等.基于 Web 的情报知识元挖掘与语义集成地图[J].情报学报, 2008, 27(2): 163-168.(Wen Youkui, Sun Ming, Wen Hao, et al. Information knowledge element mining based on web and semantic integrating map [J].Journal of the China Society for Scientific and Technical Information, 2008, 27(2): 163-168.)

[21] 赵洪.基于 Ontology 的 Web 表格数值知识元抽取研究与实现[D].天津: 南开大学, 2008.(Zhao Hong, Research and implementation of numerical knowledge element extraction over web tables based on ontology[D].Tianjin: Nankai University.)

[22] 赵洪,肖洪,薛德军,等.Web 表格信息抽取研究综述[J].现代图书情报技术, 2008,24(3): 24-31.(Zhao Hong, Xiao Hong, Xue Dejun,et al. A survey of the research on information extraction over web tables[J].New Technology of Library and Information Service, 2008,24(3): 24-31.)

[23] 曹罗生. Internet 浏览翻译系统数字与数词表达形式的识别技术[J].佛山科学技术学院学报:自然科学版, 2004, 22(3): 31-34.(Cao Luosheng. The technology of distinguishing expressions of digits and numerals for browsing translation system[J].Journal of Foshan University(Natural Science Edition), 2004,22(3):31-34.)

[24] 肖洪,薛德军.基于大规模真实文本的数值知识元挖掘研究[J].计算机工程与应用, 2008, 44(30): 150-152,222.(Xiao Hong, Xue Dejun. Numeric knowledge element mining based on large-scale realistic corpora[J].Computer Engineering and Applications, 2008, 44(30): 150-152,222.)

[25] 温有奎,温浩,徐端颐,等.基于知识元的文本知识标引[J].情报学报,2006,25(3):282-288.(Wen Youkui,Wen Hao,Xu Duanyi,et al.Text knowledge indexing based on knowledge element[J].Journal of the China Society for Scientific and Technical Information,2006,25(3):282-288.)

[26] 温有奎.文本知识元标引[C]//第十九届全国计算机信息管理学术研讨会,宜昌, 2005: 59-66.(Wen Youkui. Text knowledge element indexing[C]//The 19th China workshop on computer information management, Yichang, China, 2005: 59-66.)

[27] 温有奎,徐国华,赖伯年,等.知识元挖掘[M].西安:西安电子科技大学出版社,2005:177-183.(Wen Youkui, Xu Guohua,Lai Bonian,et al.Knowledge element mining[M].Xi'an: Xidian University Press, 2005: 177-183.)

[28] 王永璨. 面向复杂科学文本数据抽取转换及装载技术的研究与应用[D].沈阳:东北大学,2009.(Wang Yongcan. Study and implementation of ETL technology for complex scientific text data [D]. Shenyang: Northeastern University,2009.)

[29] 曹顺良,刘杰,王健,等. RE-OEM:一种半结构化生物数据的信息抽取模型[J].计算机应用研究, 2008, 25(9): 2647-2650,2654.(Cao Shunliang,Liu Jie,Wang Jian,et al.RE-OEM: Model for information extraction from semi-structured biological data[J].Application Research of Computers, 2008, 25(9): 2647-2650,2654.)

[30] 邓绪斌.面向复杂数据源的数据抽取模型和算法研究[D].上海: 复旦大学, 2005.(Deng Xubin. Study on data extraction model and algorithm for complex data sources[D].Shanghai: Fudan University,2005.)

[31] 林鸿飞,庄恩贵,姚天顺.中文文本挖掘中数字特征的抽取和表示[C]//1999年青岛-香港国际计算机会议论文集,青岛,1999:1116-1119.(Lin Hongfei,Zhuang Engui,Yao Tianshun.Digital features extraction and representation for Chinese text mining[C]// Proceedings of the Qingdao - Hong Kong International Conference on Computer in 1999, Qingdao,China, 1999: 1116-1119.)

[32] 林鸿飞,杨志豪,赵晶.中文文本的信息自动抽取和相似检索机制[J].小型微型计算机系统, 2007, 28(11): 2074-2079.(Lin Hongfei,Yang Zhihao,Zhao Jing. Mechanism of automatic extraction and similar retrieval for Chinese texts [J]. Journal of Chinese Computer Systems, 2007, 28(11): 2074-2079.)

[33] 陈鄞,赵铁军,吕雅娟,等.汉英机器翻译中数词的识别和翻译[C]//2002年全国机器翻译研讨会,北京, 2002: 86-91.(Chen Yin,Zhao Tiejun,Lv Yajuan,et al. Identification and translation of numerals in Chinese-English machine translation[C]//National Symposium on Machine Translation 2002,Beijing,China, 2002: 86-91.)

[34] Treumuth M. Automatic extraction of time expressions and representation of temporal constraints[EB/OL].[2013-02-25]. http://math.ut.ee/~treumuth/NLP/Treumuth_Term_Paper.pdf.

[35] Negri M, Marseglia L. Recognition and normalization of time expressions: ITC-irst at TERN 2004[R]. Trento: ITC-irst, 2005.

[36] Hacioglu K, Ying Chen, Douglas B. Automatic time expression labeling for English and Chinese text[J]. Lecture Notes in Computer Science,2005, Volume 3406/2005:548-559.

[37] 赵国荣.中文新闻语料中的时间短语识别方法研究[D].太原: 山西大学, 2006.(Zhao Guorong.Research into temporal expressions of Chinese news[D].Taiyuan: Shanxi University, 2006.)

[38] 吴早亮.基于搜索引擎的商品信息抽取与融合的研究及实现[D].上海:上海大学,2007.(Wu Zaoliang. Study and implementation of merchandise information extraction and fusion based on search engine[D].Shanghai: Shanghai University, 2007.)

[39] 杨少华,林海略,韩燕波.针对模版生成网页的一种数据自动抽取方法(英文)[J].软件学报, 2008, 19(2): 209-223.(Yang Shaohua, Ling Hailue, Han Yanbo. Automatic data extraction from template-generated web pages[J].Journal of Software, 2008, 19(2): 209-223.)

[40] 熊文,张玲.一种基于规则不依赖于分词的中文数量短语的识别[C]//第七届中文信息处理国际会议论文集,武汉,2007:36-40.(Xiong Wen,Zhang Ling. A rule based and no word segmentation Chinese quantifier Phrase Chunking [C]//Proceedings of the 7th International Conference on Chinese Language Processing,Wuhan,China, 2007: 36-40.)

[41] 白晓革,李义杰.数量短语的构成模式及其识别[C]//第三届 HNC 与语言学研究学术研讨会论文集,北京, 2005: 171-178.(Bai Xiaoge,Li Yijie.The language model and recognization of quantifier phrases[C]//Proceedings of the 3th HNC and Linguistics Research Symposium, Beijing,China, 2005: 171-178.)

[42] 张玲,熊文,李义杰,等.基于知识库的现代汉语数量短语的识别[C]//第七届中文信息处理国际会议论文集,武汉, 2007: 295-299.(Zhang Ling, Xiong Wen, Li Yijie, et al.Recognize modern Chinese quantifier phrases based on a knowledge-database[C]//Proceedings of the 7th International Conference on Chinese Language Processing,Wuhan,China, 2007: 295-299.)

[43] 方芳,李斌.基于语料库的数量名短语识别[C]//第三届学生计算语言学研讨会论文集,沈阳,2006:331-337.(Fang Fang,Li Bin. Corpus based investigation on MQN phrase[C]//Proceedings of the 3th Student's Workshop of Computational Linguistics,Shenyang,China, 2006: 331-337.)

[44] Tatar S, Clcekli I. Automatic rule learning exploiting morphological features for named entity recognition in Turkish[J]. Journal of Information Science, 2011, 37(2): 137-151.

[45] 陈肇雄.SC 文法功能体系[J].计算机学报,1992(11):801-808.(Chen Zhaoxiong. A new context-sensicive subcategory (SC) grammar for machine translation[J].Chinese Journal of Computers, 1992(11): 801-808.)

[46] 孙承杰.基于判别式模型的生物医学文本挖掘相关问题研究[D].哈尔滨:哈尔滨工业大学,2008.(Sun Chengjie. Research on associated issues in biomedical text mining based on discriminative models[D]. Harbin: Harbin Institute of Technology.)

[47] 刘一宁.学术定义抽取研究综述[J].情报探索,2011(8): 1-4.(Liu Yining. Overview of research on academic definition extraction[J].Information Research,2011(8): 1-4.)

[48] 杨永贵.中文信息抽取关键技术研究与实现[D].北京: 北京邮电大学, 2008.(Yang Yonggui. Research and realization on the key technologies of Chinese information extraction[D].Beijing: Beijing University of Posts and Telecommunications,2008.)

[49] Voorham J, Denig P. Computerized extraction of information on the quality of diabetes care from free text in electronic patient records of general practitioners[J]. Journal of the American Medical Informatics Association: JAMIA, 2007,14 (3): 349-354.

[50] 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.

[51] Fukuda K, Tsunoda T, Tamura A, et al. Towards information extraction: Identifying protein names from biological papers[J]. Pac Symp Biocomput,1998: 707-718.

[52] Murata M, Qing Ma, Kanamaru T, et al. Development of an automatic trend exploration system using the MuST data col-

lection[C]// Proceedings of the Workshop on Information Extraction Beyond The Document, 2006: 1-11.

[53] Nanba H, Okuda N, Okumura M. Extraction and visualization of trend information from newspaper articles and blogs [C]//Proceedings of NTCIR-6 Workshop Meeting, 2007: 243-248.

[54] 高强,游宏梁.基于层叠模型的国防领域命名实体识别研究[J].现代图书情报技术,2012, 28(11):47-52.(Gao Qiang, You Hongliang. Study on named entity recognition based on cascaded model for field of defense[J]. New Technology of Library and Information Service, 2012, 28(11): 47-52.)

[55] Nadi F, Ranaivo-Malancon B. Using frames to infer numerical extracted answers[C]//2nd IEEE International Conference on Computer Science and Information Technology, 2009: 14-18.

[56] 周秀会.知识元搜索引擎:CNKI 知识搜索平台[J].现代情报, 2007, 27(5): 220-222.(Zhou Xiuhui. Knowledge element search engine: CNKI knowledge searching platform[J].Modern Information, 2007, 27(5): 220-222.)

吴 超 中国科学技术信息研究所硕士研究生。通讯地址:北京市复兴路15号。邮编:100038。

郑彦宁 中国科学技术信息研究所研究馆员,博士生导师、博士后合作导师。通讯地址同上。

化柏林 北京大学信息管理系博士后。通讯地址:北京市海淀区颐和园路5号。邮编:100871。

(收稿日期:2013-05-28;修改日期:2013-09-21)