面向网络舆情文本的语义分析技术比较研究

郑 威[a],王荣璞[a],李 志[b]
中国人民警察大学 a. 研究生院; b. 智慧警务学院,河北 廊坊 065000

摘 要: 在归纳面向网络舆情文本的基本处理技术的基础上,详细总结与分析了网络舆情文本语义分析的关键技术。列举在处理网络舆情文本数据时常见的语义分析任务,根据所提出的任务对网络舆情信息语义分析关键技术加以比较,分析优势及不足,挖掘其未来的发展方向,为网络舆情监测和管理等工作提供参考。

关键词: 网络舆情; 语义分析; 文本挖掘
中图分类号: G203 文献标志码: A 文章编号: 1008 - 2077(2020)08 - 0070 - 06

一、引言

随着大数据、移动互联网等信息技术的兴起和发展,人们的生活生产方式发生了巨大变革。根据中国互联网络信息中心(CNNIC)发布的第 44 次《中国互联网络发展状况统计报告》,截至 2019 年 6 月,我国网民规模达 8.54 亿,较 2018 年底增长 2 598 万,互联网普及率达 61.2%,较 2018 年底提升 1.6 个百分点。大量社交媒体平台的广泛使用正逐渐改变着机构与机构之间、机构与人之间以及人与人之间的社会关系,改变着我国舆情的发生与传播模式。在“全民麦克风”时代下,网络舆情作为人们对于公共事件或社会问题的一种意见和态度,是网络社会的“体温计”与“晴雨表”,是社情民意的真实反映,是评价社会和谐稳定与否的重要依据[1]。因此,网络舆情研究作为网络社会治理与网络安全的题中应有之义,得到了学术界的高度关注。当前网络舆情分析研究中,多以舆情预警、舆情获取、舆情演化、舆情分析、舆情引导为主,舆情数据识别技术贯穿于整个网络舆情分析研究。

语义通常指信息中所包含的概念、含义,是对事物的描述和逻辑表示,它表达事物的本质、因果、上下位等各种逻辑关系[2]。语义分析是自然语言处理的重要步骤,其识别信息所包含的语义,并建立计算模型。在网络舆情研究领域,语义分析在信息检索与挖掘、舆情采集与提取、话题发现与追踪、文本倾向性分析等方面发挥着重要的作用,目前面向网络舆情的语义分析研究主要集中于对现有技术的优化与引入新技术两个方向。本文首先对语义识别的基础共性技术进行梳理,然后从理论层面研究当前主流的网络舆情文本语义识别关键技术,通过比较分析各技术应用在具体网络舆情分析任务时的优缺点,更好地了解网络舆情文本语义分析的现状及应用前景。

收稿日期: 2019 - 12 - 12
基金项目: 公安理论及软科学研究计划重点项目“面向大数据的涉恐涉稳网络舆情风险预测与情报感知研究”(2018LLYJWJXY006); 河北省高等教育教学改革研究与实践项目“公安情报可视化教学系统构建研究与实践”(2018GJJG445); 武警学院国家社会科学基金培育课题“网络生态视域下高校网络舆情引导及主流媒体应对研究”(SKJJPY201714)

作者简介: 郑威(1996—),男,福建南安人,在读硕士研究生; 王荣璞(1995—),男,云南腾冲人,在读硕士研究生; 李志(1977—),女,河北衡水人,副教授。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

二、网络舆情文本的基本处理技术

自然语言文本处理技术主要有三个方面，按顺序可分为词法分析、句法分析和语义分析，其中语义分析是以词法、句法分析为基础的，更深层次的知识表示系统[3]。因此，在对网络舆情文本进行语义分析前，首先应该在词语层面对文本进行基本处理。一个文本串，在对其进行分词等基本处理后，就可以开始更高层的语义分析任务。文本的基本处理主要有分词、词性标注、语义消歧等技术。

(一) 分词(word segmentation)

分词技术是进行语义分析预处理的首要步骤，针对不同语言应采取不同策略。汉语文本词与词之间是互相连接的，并没有较为明显的可识别特征，具有与英语等文本不同的特点。目前，汉语分词技术主要有词典分词方法与统计分词方法两类。词典分词方法的基本思想是按照一定策略将待分析字符串与词典中字符串进行匹配，它有三个要素，分别是词典、扫描方向与匹配原则。词典分词方法常用算法有最大匹配法、逆向最大匹配法、最少切分、逐词遍历匹配法、双向扫描法等[4]。尽管词典分词方法具有快速、算法简洁等特点，然而匹配词典中的词总是有限的，在实际算法运用中常常出现词典中不存在的词(未登录词)，而采用共现分析的统计分词方法则不会存在这一问题。基于统计的分词方法的基本思想是: 因为词语是由汉字组合而成的，所以在上下文中，汉字与汉字相邻共现的概率可以反映出这一组合是词语的可信度。根据这一策略，分析文本中相邻共现的汉字，进行频率统计，再基于分析结果进行分词。基于统计的分词方法准确性很高，需要语料库的支持。但是如果单纯使用统计分词方法也具有局限性，例如分词速度相对较慢，且会出现抽取出共现频率很高但并不常用的词。因此，在实际的分词系统中一般将两种方法结合，发挥词典分词方法简洁快速以及统计分词可识别未登录词、高准确度的优点。

(二) 词性标注(part - of - speech tagging)

词性标注是处理自然语言文本的重要基础。根据语法，为句子中的每个词语分配一个正确的词性标注，例如形容词、名词、动词，以此类推。词性标注是自然语言研究中的一个难点，也是近年来的研究热点之一，研究者提出了许多有效的方法，当前的词性标注算法可以大致分为基于规则的方法和基于统计的方法。早期词性标注系统使用的基本是基于规则的算法，随着统计学在语言学中的运用，基于统计的方法成为目前最常见的方法。其中，隐马尔可夫模型(hidden Markov model，HMM)是基于统计的方法中效果较好、运用较为广泛的模型之一。国内外学者对隐马尔可夫模型进行了大量研究，在它的基础上提出了完全二阶隐马尔可夫模型、基于词法信息的隐马尔可夫模型、双状态隐马尔可夫模型等。近年来，诸如决策树和条件随机场等方法也逐渐开始用于词性标注中，并取得了良好的效果。

(三) 语义消歧(word sense disambiguation)

歧义是指一个词语或字符串存在不止一种含义，词义、句义以及篇章含义层次都会根据不同的上下文环境产生不同的意义，语义消歧即在词语具有歧义时，识别出词语在特定文本中的具体含义[5]。语义消歧这一概念在20世纪40年代就已经提出，虽然经过了长时间的研究与发展，至今仍然是一个亟须解决的难题。语义消歧主要包含在知识库中描述词语的含义与在语料库中进行词义消歧两种方法，其中基于语料库的方法根据是否有人工干预又可以细分为有监督的消歧和无监督的消歧。

基于知识库的消歧是指遵循语法关系进行的语义选择限制，例如动词对宾语的语义选择限制等。在基于语料库的方法中，有监督与无监督消歧的区别在于训练数据是否已知，即每个词语是否被标注。训练数据在有监督的消歧方法中是被标注的，在无监督的消歧方法中是未经标注的。因此，有监督的语义消歧常被称为分类任务，通过建立分类器根据上下文和标注结果完成分类任务，并使用划分多义词上下文类别的方法来区分多义词不同的词义。无监督的语义消歧通常被称为聚类任务，使用聚类算法对同一个多义词的所有上下文进行等价类划分，在词义识别的时候，将该词的上下文与各个词义对应上下文的等价类进行比较，通过上下文对应的等价类来确定词的词义。

三、网络舆情文本语义分析关键技术

(一) 传统分析技术

传统的文本语义分析技术包括向量空间模型(VSM)、TF - IDF(term frequency - inverse document frequency)、BM25(best match25)等算法，这些方法主要从词汇的层面进行浅层的语义分析。如TF - IDF算法通过评估字词对于一个文件集或一个语料库中

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

的其中一份文件的重要程度,实现浅层语义分析。基于词汇重合度进行分析的算法通常存在较大的局限性,仅仅在词汇层面进行解析,在实际的语义分析中会出现结构局限与知识局限(对断句、病句、不合逻辑文本处理结果不理想)问题。目前各类网络平台的信息发布量和事件的传播速度都是惊人的,传统分析算法较难满足网络舆情海量数据处理的需求,主要应用于关键词信息检索等领域。

(二)主题模型

主题模型(topic model)是以非监督学习的方式对文集的隐含语义结构进行聚类的统计模型,在自然语言处理与机器学习等领域用于在文本中发现抽象主题。其中代表性的方法有潜在语义分析(latent semantic analysis,LSA)、概率潜在语义分析(probabilistic latent semantic analysis,PLSA)、潜在狄利克雷分布(latent Dirichlet allocation,LDA),以及最新的、基于深度学习的 LDA2vec 等等。1988 年贝尔通信实验室的 Dumais 等人提出了 LSA,它是以向量空间模型方法为基础的一种信息检索代数模型,主要用于知识获取和展示。在 LSA 的基础上,布朗大学的 Hoffmann 于 1999 年提出 PLSA 模型,对 LSA 模型进行了改进,但是 PLSA 模型仍有存在过多参数等问题。针对 PLSA 存在的问题,Blei 等人于 2003 年提出 LDA 话题模型,该模型具有明显的优点: 内在结构清晰,具有较高算法效率; LDA 训练方法是无监督的,与训练样本的数量无关; LDA 具有 PLSA 模型的所有优点,且克服了 PLSA 的缺点[6]。由于互联网数据规模的不断增加,在网络舆情数据分析中,大规模并行 PLSA、LDA 训练将是主旋律。自 PLSA、LDA 后,又提出了很多变体,如 HDP、LDA2vec 等等。

当文本具有特定主题时,部分特定的,与主题相关的词语会频繁出现,因此典型的主题模型将文本视为大量词语形成的集合。这些方法因忽略了句法信息而被叫作"词袋"(bag of words)方法,通过使用单词的统计信息,将文本表示为词向量的集合或者词语与文本间的概率关系。基于这些信息,再进行文本集合的语义分析,如潜在主题、词语的潜在语义结构等等。此类模型对文档的生成过程进行了描述,因此叫作主题模型[2]。在传统的聚类算法中,当文本数据量过大、特征过多,尤其是面对海量网络舆情数据时,距离度量公式相对失去意义,导致聚类算法结果较差,而主题模型相对较好地解决了这一问题,在衡量文本间语义相似性、排除文档中的噪音与次要主题方面都有较好表现。目前,对于主题模型仍然存在争议,认为主题模型适合用于机器学习任务的数据结构化,而不适合作为解决特定问题的独立方法,因为当主题过多时,几乎不可避免地会出现主题重复的问题,这也是主题模型的局限性[7]。

(三)词向量

词向量(word embedding)是语义识别分析的重要工具,它的核心思想是把词映射到一个向量空间,并且这个向量空间很大程度上保留了原本的语义。词向量模型的目标是将词语映射到低维的实数型向量空间当中,词义越相近的词语在空间上的距离越近。引入空间中"距离"这一度量,就可以根据距离来判断相应词语之间的句法和语义相似性。词向量既可以作为对语料进行数据挖掘的基础,也可以作为更复杂模型的输入,是现在自然语言处理的一种主流工具[8]。1986 年,Hinton 等人首先提出了词向量的概念。2003 年 Bengio 等人提出了"神经概率语言模型",通过"学习单词的分布式表示"来减少语境中单词表示的高维度。在此阶段词向量的主流来源仍然是作为主题模型的副产品,真正运用独立模型训练和使用词向量,并使词向量方法作为处理任务的有效工具是 2013 年,由 Mikolov 领导的 Google 团队提出并开源了 Word2vec,它是嵌入式的工具包,用于训练和使用词向量。Word2vec 可比之前的方法更快地训练向量空间模型,由于其简单、高效,迅速得到了学界和工业界的广泛关注。

词向量方法的优势在于它不需要人工标注语料,可以直接输入未标注的文本训练集,理论上文档数据规模越大,学习到的词向量效果越好。可在词向量的基础上通过累加的方式构建短语向量、句向量。除了以累加构建句向量外,Mikolov 还基于 Word2vec 模型提出 Doc2vec 模型用于训练句向量。通过词向量可以解决特定含义有多种表达词汇的情况,但是无法解决一词多义情况,当前是运用多个词向量来表示多义词。

(四)卷积神经网络

近年来,深度学习(deep learning)算法的应用使文本语义分析取得了突破性进展。深度学习是机器学习方法的一种,它起源于人工神经网络,通过多层处理,逐渐将初始的"低层"特征表示转化为"高层"特征表示后,用"简单模型"即可完成复杂的分类等学习任务。深度学习的这些特性使它有非常强的处理大规模数据能力,现在已经有很多深度学习的框

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

架，如卷积神经网络（CNN）、深度置信网络（DBN）、递归神经网络（RNN）、生成对抗网络（GAN）等。其中卷积神经网络是目前应用最广泛，较为成熟的深度学习框架。

卷积神经网络是一类包含卷积计算且具有深度结构的前馈神经网络，该模型最初的设计是用于解决图像识别问题，相比于图像的像素，自然语言的信息表达方式属于更抽象的认知，模型难以识别。20世纪80和90年代的时间延迟网络和LeNet－5是最早的对卷积神经网络的研究。21世纪以来，随着深度学习理论的发展和数值计算技术的改进，卷积神经网络得到了迅速发展，被广泛应用于计算机视觉、图像识别、自然语言处理等领域。在网络舆情文本分析领域，卷积神经网络可以进行文本分类、情感分析、本体分类等数据分析工作[9]。传统的词袋模型和向量空间模型采用统计词频的方法对文本进行建模，将文本视为无序，且词语间没有语义关系。基于深度学习的神经网络模型在文本分类任务中也取得了长足的进步，它摆脱了传统的人工特征提取的缺点，具有自学习功能，能够精确地解决分类、函数拟合和预测等问题。

卷积神经网络在本质上是一种输入到输出的映射，能够学习大量的输入与输出之间的映射关系。它的特殊性体现在两点：一是它的同一特征映射面上的神经元权值相同，所以网络可以并行学习；二是权重共享，即一层中同一个卷积滤波器的权重相同。卷积神经网络以其局部连接与权值共享的特殊结构有着独特的优越性，其结构更类似于生物神经网络，降低了网络模型的复杂性，避免了特征提取和分类过程中数据重建的复杂度。训练时间成本较大是卷积神经网络存在的局限性之一，而且神经网络算法经常由于样本、参数或训练时间等问题导致结果欠拟合或过拟合。尽管卷积神经网络减少了网络中的参数，相对较好地减少模型过拟合，但仍然存在这一问题。

四、网络舆情文本分析任务的语义分析技术运用比较

语义分析技术正在不断蓬勃发展，在分析网络舆情数据时它们具有不同的优势，也存在一些局限性。以下简要列举在处理网络舆情文本数据时常见的语义分析任务，然后对比分析各种语义识别技术在具体分析任务时的运用，比较其优势以及不足。

（一）常见的网络舆情文本语义分析任务

1. 文本分类

文本分类是最常见的文本语义分析任务，是机器学习领域和自然语言处理领域的核心问题之一。文本分类是一种有监督的机器学习方法，其分类原理为：当分类目标已知时，根据已经被标注的训练文本样本集合对分类规则进行归纳，总结出分类规律，得到文本属性与文本类别之间的关系模型[10]，然后根据此关系模型对需要分类的新文本进行正确分类。文本分类的对象可以是短语、句子或者包含文本段落的整篇文档等等，可以从语料库、互联网等渠道获得。文本分类的流程为预处理、文本表示及特征选择、构造分类器、分类[11]。

2. 情感分析

情感分析是自然语言处理领域中广泛使用的技术，主要通过文本数据来量化文本的情感。现有的情感分析方法可以归类为情感词典分析法和机器学习分析法。情感词典分析方法通过剪切文本和删除停用词，从关键词中提取积极与消极关键词，然后计算情感得分。传统的情感词典分析法有很多缺点，首先，情感词典分析法的准确性与情感词典有很大关系；其次，情感词典分析法只是简单的关键字匹配，通过数学化的计算得出消极和积极词语的数量来判断情感具有局限性。因此，现有的情感词典分析方法都针对传统算法的问题进行了改进。机器学习方法在面对海量舆情数据时具有一定优势。使用机器学习算法研究文本情感分析主要是把情感分析看作对情感态度进行分类的过程，提取文本中的关键语言特征信息，输入到机器学习分类模型，通过样本数据训练的模型来预测其他文本的情感类别[12]。

3. 自动文摘

自动文摘的研究始于1958年Luhn发表的论文，自动文摘技术希望通过计算机将冗长的文本压缩到规定长度内，同时保持原始文本主要信息不丢失。自动文摘技术提供了一种快速了解相关话题的方式，该技术可以快速地对文档进行总结，用户只需阅读短短十几句或几十句的总结便可以了解相关信息。用户的需求驱动使得很多类型的文摘方法应运而生，根据摘要的生成方式，自动文摘方法可以分为抽取式摘要和生成式摘要。抽取式摘要直接从原始文本中抽取具有显著性的句子构成摘要，这类方法虽然相对简单，但是直接复用原文中的句子能够准确地捕获原文意图；另一方面，因为是句子级别的操

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

作,生成的摘要连贯性较低,可读性会差一些。生成式摘要采用更加复杂的自然语言生成技术,在对原文理解的基础上,生成新的描述形成摘要,因此可能出现原始文本中不存在的单词句子,这类方法更加贴合人类手写摘要的过程,生成的摘要可读性更强,但对内容的复现上可能会有偏差。

(二)网络舆情文本识别算法运用比较分析

1. 在文本分类任务中的比较分析

传统的文本分类方法大多基于浅层机器学习模型,以概率统计的思想为基础,需要人工构建繁杂低效的特征工程和对模型参数的调优,整个过程耗时耗力,造成模型性能不稳定且鲁棒性差。Word2vec词向量模型的开源是文本分类研究从传统的浅层机器学习模型转向深度学习模型的标志,对词向量进行训练是网络舆情文本分类任务的基石。通过词向量模型,文本中的词语可以映射成连续和稠密的向量形式,从语义层面上描述词与词之间的关系并提取文本抽象特征,使得深度神经网络模型得以应用到文本分类领域,大大推动了相关研究进程。词向量方法不能很好地解决汉语中常见的一词多义的问题,而主题模型可以挖掘到文本中隐含的全局主题信息。在训练词向量的基础上,结合主题模型可以对网络舆情文本语义进行更好的挖掘[13]。

深度学习模型相比传统的机器学习模型能取得更出色的效果,这是因为它由多层神经网络结构组成,经过多次非线性特征空间的变换和海量数据的训练,能学习到对分类更为重要的语义特征。CNN模型独有的卷积和池化操作使其可以很好地捕获网络舆情文本深层次语义信息,是网络舆情文本分类任务的非常有效的模型。

2. 在情感分析任务中的比较分析

由于文本表达的稀疏性问题,传统情感词典分析方法在处理短文本时常常不能得到满意的结果,然而在网络舆情中存在着大量的短文本内容,如评论、微博等等。主题模型方法能有效利用大量未标注数据,同时还能发掘文本的潜在主题信息。通过主题模型对舆情文本进行主题聚类,从语料库中发现潜在主题信息并将它们追加到文本特征中,以丰富文本的特征表示,因此分析效果有着显著提升。而使用词向量可以大幅度提高分析效果,主要是利用词向量模型,将词转化成数组,通过计算词间的数据距离,来衡量词之间的相似度,这样在模型有监督的学习了积极和消极词向量后,就可以得到结果了。通过词向量模型构建的情感词典相较于传统方法在情感分析的准确率上有所提高,因为词向量提供了词汇更加丰富的上下文信息,而传统的情感词典中仅包含单词的情感值[12]。

随着网络舆情文本数据量的不断增加,机器学习方法逐渐体现出特有的优势。其中,卷积神经网络在特征学习、提取、并行计算等方面有着强大优势,它可以自动学习文本的向量表示,同时不需要使用情感词典等额外资源,避免了情感词典覆盖率低的问题。卷积神经网络通过卷积核的运算对输入的词向量进行特征的提取,提取文本信息建立特征表示,进而完成文本建模分析等任务。在处理输入的词汇序列时,通过卷积的方式,捕捉句子中与情感有关联的词语模式,与情感无关的词语模式,在特征映像中取值较小,因此可以最大限度地保留与情感有关联的词语模式,有更好的分析效果,为情感分析中的很多问题提供了新的研究解决思路。

3. 在自动文摘任务中的比较分析

传统的抽取式摘要方法对于网络舆情文本的表示具有很大的局限性,因为汉语文本中存在很多一词多义和多词一义的现象,而仅用各自独立的特征来表示文本,忽略了文本内容中前后文语境的影响,因此得到的网络舆情文摘效果不佳[14]。基于主题模型的方法克服了传统方法的缺点,通过对大规模语料库的统计计算和分析,可以提取和推断出语句篇章中词语的预期语境使用关系,因此获取的文摘质量也有一定的提升。但是这些方法都属于抽取式摘要方法,本质上还是采用了“词袋”模型,因此在文本的语义表达上还存在一定的欠缺,在获取文摘的过程中需要人工选择大量特征,当面对大规模的数据集时,这些工作费时费力,人工提取的特征噪音也会变多。基于深度学习模型的自动文摘方法属于生成式摘要方法,面对舆情文本时实现的文摘效果高于抽取式摘要方法。卷积神经网络模型在对句子进行建模时充分考虑了句子的结构信息,在多层卷积和池化的过程中,始终会将句子的语义信息和特征进行保留,因此,基于卷积神经网络的方法在自动文摘任务中有一定的优势。同时,采用词向量来学习文档向量也是卷积神经网络方法运用的前提,当不同的词向量算法学习文档向量的表示时,会影响文摘获取效果。

将以上比较分析进行汇总,结果见表1。总体来说,卷积神经网络在网络舆情语义分析的各类任务

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 1　网络舆情文本分析任务的语义分析技术比较

文本分析任务	词向量	主题模型	卷积神经网络
文本分类	词向量是卷积神经网络等深度学习方法运用的前提。将词转化成数组，计算词间的数据距离，来衡量词之间的相似度	主题模型可以挖掘到文本中隐含的全局主题信息，解决词向量不易解决的一词多义问题	解决了以往传统方法在复杂函数的表示方面能力受限的问题
情感分析	词向量提供了词汇更加丰富的上下文信息，而传统的情感词典中仅包含单词的情感值，因此在准确率上有所提高	能有效利用大量未标注数据，发掘文本的潜在主题信息，分析效果显著提升；在提取特征时缺乏对词语关联及相关词性的理解	可以自动学习文本的向量表示，同时不需要使用情感词等额外资源，避免了情感词典覆盖率低的问题
自动文摘	采用词向量学习文档向量的表示时，相对于传统简单特征统计方法的文档表示，文摘获取效果有较大提升。词向量的词典数更大时，效果更好	克服传统方法的缺点，通过对大规模语料库的统计计算和分析，可以提取和推断出语句篇章中词语的预期语境使用关系，获取的文摘质量有一定的提升	在多层卷积和池化的过程中，始终将句子的语义信息和特征进行保留。在同样采用词向量进行文档表示的前提下，相比于其他方法具有更好的效果

中都有良好表现，能较好地完成网络舆情语义分析的各项要求。词向量与主题模型可以对传统方法进行有效优化，同时词向量也常常是卷积神经网络运用的前提。当前许多学者针对这三种算法的优缺点，提出了基于这三种算法的变体，这些方法在针对特定任务时会有更好的效果，因此在设计舆情分析系统时应兼顾各种技术的优缺点，合理设计模型。

五、结语

随着网络上各种类型社交平台的发展以及移动平台的崛起，人们在网络上的互动交流能力不断提升，对舆情信息的语义分析能力也提出新的要求。当前在针对各类平台的网络舆情数据进行分析的过程中，对于语义分析技术的使用仍然存在一些局限。

首先，当前将深度学习应用于语义分析任务的方法受到很多学者的关注，可以发现近年来在针对一些特定任务时，基于深度学习的方法逐渐体现出特有的优势，但是这并不意味着不需要关注其他方法的运用。深度学习目前仍具有忽视语言结构的局限性，存在“语句只是单词序列”的逻辑。到目前为止，深度学习的探索才刚刚开始，并且成熟的方法数量有限。未来网络舆情文本的语义分析技术研究趋势应当是以应用深度学习为主，其他方法为辅。

其次，在文本语义分析领域对于不同类别的文本进行分析均有比较优秀的算法，但是缺乏针对网络舆情文本特点的识别技术。网络舆情文本具有一定的特点，应当进行针对性的拓展和改进，包括建立专门的分词词典，完善对分类特征的选择等。近年来网络舆情领域的研究正不断细化，今后的研究方向在不断发掘深度学习方法等前沿技术外，还应当进一步细化分析算法，明确分析对象，还有大量的基础性工作需要完善。

参考文献：

[1] 孙宁，陈雅. 基于信息计量学的我国网络舆情研究综述[J]. 情报杂志，2014，33(5)：136－142.

[2] 李生. 自然语言处理的研究与发展[J]. 燕山大学学报，2013，37(5)：377－384.

[3] 王璐璐，袁毓林. 走向深度学习和多种技术融合的中文信息处理[J]. 苏州大学学报(哲学社会科学版)，2016，37(4)：160－167.

[4] 马张华. 信息组织[M]. 北京：清华大学出版社，2003：63－89.

[5] 吴云芳. 词义消歧相关术语简介[J]. 术语标准化与信息技术，2010(3)：18－20.

[6] 单斌，李芳. 基于 LDA 话题演化研究方法综述[J]. 中文信息学报，2010，24(6)：43－49.

[7] 徐戈，王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报，2011，34(8)：1423－1436.

[8] 江大鹏. 基于词向量的短文本分类方法研究[D]. 杭州：浙江大学，2015：7－8.

[9] 刘磊，李壮，张鑫，等. 中文网络文本的语义信息处理研究综述[J]. 计算机应用研究，2015，32(1)：6－10.

[10] 黄微，张耀之，李瑞. 网络舆情信息语义识别关键技术分析[J]. 图书情报工作，2015，59(21)：33－37.

[11] 张耀之. 网络舆情语义识别的技术分析及识别流程构建[D]. 长春：吉林大学，2016：25－27.

[12] 徐小龙. 中文文本情感分析方法研究[J]. 电脑知识与技术，2018，14(2)：149－151.

[13] 黄微，刘熠，孙悦. 多媒体网络舆情语义识别的关键技术分析[J]. 情报理论与实践，2019，42(1)：134－140.

[14] 秦春秀，祝婷，赵捧未，等. 自然语言语义分析研究进展[J]. 图书情报工作，2014，58(22)：130－137.

(下转第 82 页)

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Enlightenment of “Double First-class” University Project Experience on First-class Police University in China

DU Yuanbin

Academic Affairs Office, China People's Police University, Langfang, Hebei Province 065000, *China*

Abstract: Since the implementation of “double first-class” construction, the universities and colleges in the selected list have carried out practical and efficient reforms, and gained remarkable achievement of construction and valuable experience of reforms. Police universities and colleges also take the “double first-class” construction as the standards and objectives of education and teaching reform. In order to further clarify the construction ideas of China's first-class police academies in the new era, this paper analyses the reform measures of education, teaching and personnel training modes of domestic universities in “double first-class” construction and well-known foreign military and civilian universities in recent years. To construct a world-class police university with Chinese characteristics in the new era, this paper suggests the priority for comprehensive education reform should be given to such tasks as ascertaining school-running orientation, optimizing development arrangement, promoting the reform of recruitment and cultivation mechanism, building a first-class discipline, developing an incentive innovation mechanism and improving the level of international education so as to elevate the level of education and the quality of talent cultivation.

Key words “double first-class” construction; first-class police academy; talent development

(责任编辑 李 蕾)

(上接第 75 页)

Analysis of Semantic Analysis Technology for the Texts of Internet Public Opinion

ZHENG Wei[a], WANG Rongpu[a], LI Zhi[b]

a. Graduate School; b. School of Intelligent Policing, China People's Police University, Langfang, Hebei Province 065000, *China*

Abstract: Based on the summary of the basic processing technology of internet public opinion texts, this paper summarizes and analyzes the key techniques of semantic analysis of internet public opinion texts in detail. A comparison is made on the key techniques used in the common semantic analysis tasks when processing internet public opinion text data to analyze their advantages and disadvantages so as to dig their future development direction, and provide reference for the monitoring and management of internet public opinion.

Key words: internet public opinion; semantic analysis; text mining

(责任编辑 李 蕾)

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net