Journal of Computer Applications ISSN 1001-9081 2013-06-01
计算机应用,2013,33(6):1574-1578,1607 CODEN JYIIDU http://www.joca.cn

文章编号:1001-9081(2013)06-01574-05 doi:10.3724/SP.J.1087.2013.01574

文本情感分析综述

杨立公[1*],朱 俭[2],汤世平[1]

(1.北京理工大学 计算机学院,北京 100081; 2.中国青年政治学院 计算机教学及应用中心,北京 100089)

(*通信作者电子邮箱 yyllgg@gmail.com)

摘 要: 以文本颗粒度为视角,从情感词抽取、语料库和情感词典构建、评价对象与意见持有者分析、篇章级情感分析、实际应用五个方面对文本情感分析文献进行了梳理,并做出必要评述。指出当前情感分析系统的准确率普遍不高,进一步研究的重点在于:自然语言处理的研究成果在文本情感倾向分析中更广泛和贴切的应用;选取文本情感倾向分类的特征和方法;利用现有语言工具和相关资源,规范、快速地构造语言工具和相关资源并应用。

关键词: 文本情感分析;情感词;语料库;情感词典;意见持有者

中图分类号: TP391.1 **文献标志码:** A

Survey of text sentiment analysis

YANG Ligong[1*],ZHU Jian[2],TANG Shiping[1]

(1. *College of Computer Science*, *Beijing Institute of Technology*, *Beijing* 100081, *China*;
2. *Computer Science and Application Center*, *China Youth University for Political Sciences*, *Beijing* 100089, *China*)

Abstract: This survey summarized the studies on text sentiment analysis in the view of granularity from the following five aspects: sentiment word extraction, sentiment corpus and dictionary construction, entity and opinion holders analysis, document level sentiment analysis, and text sentiment analysis applications. It pointed out that the current sentiment analysis system cannot gain high precision. Further research should focus on: widely and appropriately applying study achievement of natural language processing to text sentiment analysis; finding and choosing suitable features and algorithms in text sentiment classifications; utilizing the existing language tools and relevant resources in fast building standard language tools and resources and applying them.

Key words: text sentiment analysis; sentiment word; corpus; sentiment word dictionary; opinion holder

0 引言

互联网的快速发展促进了其自身由“阅读式互联网”向“交互式互联网”转变。网络不仅成为人们获取信息的重要来源,也成为人们发表自己的观点和分享自己的体验,直接表达喜、怒、哀、乐等各种情感的重要平台。

当前,针对网络文本内容的处理技术主要着眼于客观性信息,相应信息检索一般为关键词检索,信息抽取技术主要抽取文本描述的特定事件发生的时间、地点、相关人物、过程、属性等,信息抽取结果大多是客观性事实,并按照客观主题进行分类。面对大量来自微博、论坛、博客的非结构化或半结构化评论文本,迫切需要通过计算机快速、有效地完成意见性文本信息分类和情感信息的抽取,然后通过挖掘和分析文本中的立场、观点、看法、情绪、好恶等主观信息,对文本的情感倾向做出判断。

1 文本情感分析概述

文本情感分析又称意见挖掘,是指通过计算技术对文本的主客观性、观点、情绪、极性的挖掘和分析,对文本的情感倾向做出分类判断。随着计算机和网络技术的发展,人类开始研究如何让计算机能理解和运用人类社会的自然语言,这一研究取得了丰硕的成果,这些成果为文本情感分析奠定了基础。文本情感分析是自然语言理解领域的重要研究分支,涉及统计学、语言学、心理学、人工智能等领域的理论与方法。

文本情感分析首先需要对文本来源进行处理,对网络文本进行主客观分类。网络文本信息可以广义地分成两种类型:客观性文本和主观性文本。客观性文本就是我们对于实体、事件以及它们属性的客观性陈述;主观性文本通常是我们对于实体、事件以及它们属性的主观性评价,包含着丰富的主观性的意见、情感、观点和态度等。主客观分类从主客观混合的文本中将描述事实的客观性文本与表达意见的主观性文本区分开来,将主观语言的文本抽取出来,过滤掉不带情感色彩的文本。这一阶段研究的主要目的是为文本情感极性分析提供主观性文本。

文本情感分析的下一步是对主观性文本的分析,主要包括文本情感极性分析和文本情感极性强度分析。情感极性分析的任务就是识别主观文本的情感极性。情感极性分为两极,即正面(Positive)的赞赏和肯定、负面(Negative)的批评与否定,也有一些学者在正面和负面之间加入了中性(Neutral)。情感极性强度分析就是判定主观文本情感极性强度,比如强烈贬抑、一般贬抑、客观、一般褒扬、强烈褒扬五个类别。

按照文本的颗粒度,文本情感分析可以划分为针对文本中的词、句子、篇章三个级别的识别与分析。词的情感分析是文本情感分析的基础,它既是判定文本情感的基础,又是句子和篇章情感分析的前提。基于词的情感分析研究主要有情感

收稿日期:2012-11-22;修回日期:2013-01-28。 基金项目:北京市自然科学基金资助项目(4123094)。

作者简介:杨立公(1966-),男,北京人,工程师,博士,主要研究方向:自然语言理解; 朱俭(1976-),男,江苏徐州人,讲师,博士,主要研究方向:自然语言理解; 汤世平(1975-),男,江西吉安人,讲师,博士,主要研究方向:自然语言理解。

词抽取,情感词判定,语料库与情感词典的研究等。句子的情感分析是文本情感分析的核心: 一方面,它综合了情感词的分析结果,给出全句的情感分析的完整结果; 另一方面,句子可以视为短篇章,句子的情感分析的结果在很大程度上决定了篇章的情感分析结果。篇章的情感分析是最不确定性的研究,因为需要综合篇章的各个粒度下的情感分析结果,结合上下文和领域知识库做出判断。本文将以文本颗粒度为视角,分别从情感词抽取、语料库和情感词典构建、评价对象与意见持有者分析、篇章级情感分析及情感分析的实际应用五个方面对文本情感分析领域国外的相关文献进行梳理、评述。

2　情感词抽取

情感词又称极性词、评价词语,特指带有情感倾向性的词。通常情况下,情感词有褒义和贬义两类倾向性。情感词抽取和判别即是词汇级情感分析的基础工作,更是句子级和篇章级情感分析的基础,因此引起了学者的广泛关注和研究。情感词抽取目前主要分为基于语料库和基于词典的两种研究方法。基于语料库的情感词抽取和判别主要是利用大语料库的统计特性,优点在于简单易行。

1997 年,Hatzivassiloglou 等[11] 发表了在词汇级进行情感倾向性分析的研究。他们指出,在例句"this car is beautiful and spacious"中,如果"beautiful"是褒义的,那么因为两个形容词之间使用连词"and",可以推测"spacious"也是褒义的。同样,连词"or"、"but"等都可以设定处理规则,通过大规模语料库提取关联的形容词,并使用通过线性回归模型分析这些形容词的情感极性,最后通过聚类算法将形容词分组,达到 82% 的准确率。Wiebe 等[12-7] 对主观性文本进行系统的分析研究,他们使用了一种相似度分布的词聚类方法在大语料库上完成了形容词性的评价词语的获取。同时他们发现词的主观性信息跟词义是相关联的,主观性标注可以应用在词义消歧。针对有监督的机器学习方法,Wiebe 等提出一个词汇和短语级的标注方法并标注了大量的语料,该语料大量地运用在后续的相关研究中。Wiebe 等还研究了无监督的机器学习方法,提出了基于规则的自动学习方法。Pang 等[8] 使用电影评论作为实验语料,借鉴了传统自然语言处理中的文本分类技术,使用了三种机器学习的分类方法: 朴素贝叶斯、最大熵模型和支持向量机(Support Vector Machine,SVM)模型。Riloff 等[9] 用手工方法制定模板并以此选取种子情感词,通过使用迭代的方法可以获取名词词性的情感词,这一方法弥补了以往情感词抽取方法词性局限于形容词词性而忽略其他词性的情感词的缺点。Turney 等[10] 提出了点互信息(Point Mutual Information,PMI)的方法判别某个词语是否是评价词语,这种方法适用于各种词性的情感词的识别,但是较为依赖种子情感词集合。Ding 等[11] 提出特定领域情感词的配对,比如在相机领域中,"电池时间长"中"长"是褒义的,"聚焦时间长"中"长"是贬义的,通过组成"电池时间""长")、("聚焦时间""长")配对,判断配对的情感极性。Ganapathibhotla 等[12] 的研究也采用了相似的方法。Lu 等[13] 运用了 Turney 等[14] 提出的句法模式,通过最优化问题求解确定每一配对为褒义或贬义。

基于词典的情感词抽取及判别方法主要是使用词典中的词语之间的词义联系来挖掘评价词语,常用词典包括 WordNet[15]、General Inquirer(GI)词典等。Kim 等[16-17] 利用词典将手工采集的种子情感词集进行扩展来获取大量的情感词,但是这种方法的效果对种子情感词集的个数和质量依赖性比较大。Yu 等[18] 使用构造情感词种子词集的方法来计算新词语与种子集合中个别词的共现概率,但该方法仅仅把共现概率高的词对作为有相同极性的判断依据,没有考虑到文本中的领域以及上下文语境对观点词极性变化的影响。实际在不同的领域,不同词语的共现概率并不均等,而随着语境的变化,词语的极性也有极大差异。Andreevskaia 等[19-22] 使用词典中词语的注释信息来识别与判断情感词的极性。Kamps 等[23-24] 沿用了 Turney 等[4] 的基于点互信息的思想,使用了基于语义词典的方法,通过计算 WordNet 中的所有形容词与种子情感词之间的关联度来识别出情感词,该方法仍然属于无监督的学习方法,能够从统计的意义上处理大量观点词之间的极性依赖关系。该方法在一定程度上使用了词语的语义知识,但因为没有从领域和语境的角度出发,所以并不能适应在变化语境状况下的观点词极性判断。

基于词典的方法具有获取情感词全面、准确的优点,但是由于存在一词多义现象,构建的情感词典往往含有较多的歧义词。

3　语料库、情感词典构建

语料库和情感词典是文本情感分析所需的重要资源,学者们对此给予了极大的关注。

语料库可以通过人工标注[7,25] 方法构成,但是这种方法存在语料库规模有限、情感词在语料库中的分布不均匀、工作量大和容易产生人为差错的缺点。也有学者提出自动标注的方法,比如借助表情符号标注情感[26],借助股票走势和其他金融指标判断文本极性[27-28],还有学者提出借助其他标注语料的方法,例如借助烂番茄(著名的影评网站,提供电影信息与影评检索 API)的影评及其影评单句汇总来收集语料。这类方法的普遍缺点为精确度不高。

目前比较有代表性的语料库有:

1)康奈尔影评数据集(Cornell movie-review datasets)[8,29-30],由 IMDB 的电影评论构成,包含篇章级的 1 000篇褒义评论和 1 000 篇贬义评论,句子级的 5 331 个褒义句子和5 331个贬义句子。但由于未进行更细粒度的标注,它不适合细粒度的情感分析的要求。

2)多视角问答(Multiple-Perspective Question Answering,MPQA)[1,7,31] 语料库对 535 篇广泛来源的新闻进行了语句级人工标注,而且对语句的低层次进行了标注,标出了情感文本的持有者、对象、极性、强度等要素。除了情感,还标注了情绪、推断、信念等。

3) Blog06 是格拉斯哥大学的 TREC 测试集[32],主要由系列话题的主流博客组成。其特点是数据量大,有 25 GB。

4) NTCIR 多语言语料库(NTCIR multilingual corpus)[33] 由英文、中文和日文新闻组成,训练集标注了意见持有者、意见持有者的所有意见、情感极性、根据系列主题预设的相关信息。

5)美国伊利诺斯大学的 Hu 等[34] 的电子产品评论数据集对亚马逊和 cNet 上的评论进行了人工标注,该数据集标注了评价对象、倾向性以及强度,被广泛应用在细粒度的情感分

析中。

目前比较有代表性的情感词典有:

1) General Inquirer lexicon。Stone 等[33]收集了1915个褒义词和2293个贬义词,并按照极性、强度、词性等打上不同的标签,对于词汇还列出不同的义项,可以区别不同义项和词性下的褒贬极性,也相当于对每个单词都构建了一组语义消歧规则。

2) Sentiment lexicon。Hu 等[34]提供,现已包括6800个情感词,特别需要指出的是词典中的一些拼写错误的词汇是因为其出现频率高而存在的。

3) MPQA subjectivity lexicon。Wilson 等[31]提供,包括超过8000项情感词和组合短语。

4) SentiWordNet,Esuli 等[21-22]提供,为 WordNet 的每个同义词集提供褒义、贬义、主观性的评分,评分的取值范围为[0,1.0]。某一同义词集的三类评分可能都不为0,但是评分总和为1,以此表示其在三种情感属性中的倾向。最新的 SentiWordNet 版本为3.0。

5) Emotion lexicon。Mohammad 等[37]通过亚马逊的土耳其机器人(Mechanical Turk)低成本、中规模、高质量地标注词性、极性、属性。

4 情感句判断

语句既可以看作是词汇的自然聚合体,也可以看作是单一语句的短小篇章,而且篇章的整体情感极性是以语句情感极性为基础的。语句级情感分析可以分为三个方面:主客观分析并提取客观性文本;语句情感属性的识别,包括观点持有者和评价对象的抽取;极性和极性强度分析。

Wiebe 等[38]用一系列二元特征通过朴素贝叶斯分类器进行主客观分类。文献[4]介绍了无监督学习方法,首先对少量种子语料进行详细人工标注,通过词汇分布相似度来划分词类,进而进行主客观分类。Hatzivassiloglou 等[39]进一步在 Wiebe 的方法中增加了倾向性和等级属性,使用 SIMFINDER[40]对语句计算相似度,然后依据相似度用朴素贝叶斯分类器进行主客观分类。Riloff 等[9]提出对有监督学习的训练集自动标注的自举(bootstrapping)方法。Pang 等[29]提出了基于图的最小切割法。Benamara 等[41]对主客观性进行了详细讨论,指出存在有情感的客观句和无情感的主观句。

Turney 等[42]使用点互信息判定语句情感极性,提出了先抽取出主观句,进而再对其进行情感分类的方法。该方法使用形容词种子集为语句中的词评分,然后根据等分判断语句情感倾向。Hu 等[34]通过 WordNet 的同义词、反义词关系,得到情感词汇及其情感极性,然后由句子中占优势的情感词汇的语义倾向决定该句子的极性。Kim 等[43]的研究也与之相似。

情感句判断的方法主要可以归纳为有两大类:一类是基于情感词的方法,另一类是机器学习方法。

5 评价对象与意见持有者分析

在语句情感分析中,分析意见的其他属性和语句上下文对正确理解和判断情感极性至关重要。Kim 等[43]认为意见由四个元素组成,即主题(Topic)、持有者(Holder)、陈述(Claim)和情感(Sentiment)。这四个元素之间存在着内在的联系,意见持有者针对某主题发表了具有情感的意见陈述。情感分析的过程就是要在文本中抽取出这些元素并分析它们之间的关系,它的主要任务包括:

1) 主题抽取(Topic Extraction):识别主题或评价对象;

2) 意见持有者识别(Holder Identification):确定意见表述的作者;

3) 陈述筛选(Claim Selection):针对文本的主客观分类;

4) 情感分析:确定文本的语义倾向(Semantic Orientation),即情感极性(Polarity)。

Pang 等[44]举例说明了评价对象和意见持有者对情感极性判定起着非常重要的作用。对于同一条评论"去读原著吧",根据意见领域的不同,也可以说根据评价对象或意见持有者的不同,可以得到不同的情感极性。如果这一评论来自于书评这一领域,或者说评价对象是某一本书,再或者说意见持有者是某一本书的读者,即表示评价者认为这本书值得一读,是肯定的评价,也是显式评价;而如果这一评论来自于影评这一领域,或者说评价对象是某一电影,再或说意见持有者是某一电影的观众,则表示评价者认为这个电影不如原著精彩或偏离了原著,是否定的评价,也是隐式评价。评价者提供了不但看过电影而且读过原著这一暗示。

Kim 等[43]提出的意见持有者和评价对象归类为命名实体识别,直接借助自然语言处理技术处理,如借助语义角色标注来完成观点持有者的抽取,或借助于命名实体识别技术来获取观点持有者。Kim 等[17]还将所有名词短语都视为候选观点持有者,使用 ME(Maximum Entropy)模型来进行计算。Choi 等[45]使用 CRF(Conditional Random Field)模型融合各种特征来完成观点持有者的抽取,这种方法的关键在于分类器和特征的选取。Bethard 等[46]认为观点持有者一般是和观点同时出现的,在抽取出情感句中的评价对象之后,分析句中观点和动词的句法关系,即可同步获取观点持有者。

6 篇章级情感分析

篇章级情感分析需要考虑更多上下文语义理解和领域知识,有时不同极性的语句综合在一起可能产生某一特定的篇章情感极性,有时没有情感极性的语句集合可能产生强烈的情感极性。例如,马克·吐温在一段文本最后写到"每次我阅读'傲慢与偏见',我就想挖出来她,用她自己的胫骨敲打她的头盖骨"。虽然表面上这是一个没有负面情感词的句子,但是却使整个篇章表现出强烈的情感倾向。

当前,篇章级情感分析多采用基于统计机器学习方法,比如,文本分类中常用的 K 近邻算法、贝叶斯分类器、最大熵分类器和支持向量等。Pang 等[8]利用统计机器学习方法对电影评论进行分类,实验结果显示支持向量取得了最佳分类效果。多数研究结果表明,支持向量与贝叶斯分类器具有较好的文本分类性能,特征选择通常对于性能的影响更为关键,篇章级情感分析通常采用词的 unigram 和 bigrams 构造特征向量,词性、否定词、句法结构等特征往往能改善系统的性能。

Pang 等[47]使用 Yahoo 搜索引擎分析评论,他们通过编制搜索集合,然后比较集合中各项在搜索引擎中的返回结果(Top k)。他们的方法是非监督学习方法,可以不依靠情感词典,但是这种方法要求对搜索集合和搜索引擎的要求较高。

值得注意的是,篇章往往并非只有单一观点。多观点篇

章的细分、聚类和消歧有待进一步研究。

7 文本情感分析应用

近年来,很多研究机构将情感分析技术应用于现实的生活中,开发了很多实用的意见挖掘系统。Dave 等[48] 开发的 ReviewSeer 是世界上第一个情感分析工具。Liu 等[49] 研究并开发的产品信息反馈系统 Opinion Observer,利用网络上的顾客评论资源,对评论的主观内容进行分析处理,提取并统计消费者对产品特征的评价,并采用可视化方式显示统计和比较结果。Wilson 等[50] 研究并开发的 OpinionFinder 是一个自动识别主观性句子以及句子中各种主观性元素的系统。微软 Gamon 等[51] 研究并开发的 Pulse 系统可以从大量文本数据中,利用文本聚类技术提取出用户对产品细节的看法,并挖掘出有关汽车评价中的情感极性和强度。Cherry 等[52] 开发了自杀倾向的预防检测系统。

8 结语

综上所述,文本情感分析研究尚处于起步阶段,取得了部分研究成果,但也需要多学科知识的协同研究,面临着许多亟待解决的理论难题。

文本情感倾向分析研究实质上是自然语言处理的一个应用方向。自然语言处理通过多年研究积累了大量有价值、有深度的研究成果,显然这些研究成果对于情感倾向分析也有很重要的指导意义和应用价值。但从当前的研究状况来看,自然语言处理的研究成果在文本情感倾向分析中还没有得到更广泛和贴切的应用。因此,研究发现情感倾向分析与传统自然语言处理之间的关系和继承性,充分利用自然语言处理的研究成果,一方面可以使得更多的语言技术被用于情感倾向分析,提高该技术的研究水平; 另一方面也使得自然语言处理的成果在具体的应用显示出其重要的价值。

在情感倾向分析中,常常需要使用各种语言工具和资源,例如各种情感词典、褒贬义词典、评价信息数据库等,以及各种不同用途的训练和测试语料。这说明,这些资源是研究中不可缺少的一部分,它们既可以给分析研究提供有价值的知识和信息,又可以为研究提供各种应用领域的统计数据。但是,本文归纳的语言工具、词典和实验语料没有统一的规范,基本上都在使用各自构造的词典和语料来进行自己的研究,研究结果相互之间是难以比较的。虽然现在有了情感倾向评测会议,并提供了一些资源,但还远远不够。同时,我们发现在日常生活中使用的各种字典、词典、工具书本身就包含了大量语言知识和信息,这些工具书在编撰的过程中耗费了大量的人力和时间来保证相关信息的准确与权威,所提供的例句也非常规范准确。但是这些工具书目前尚未在研究工作中得到重视。因此,如何有效利用现有语言工具和相关资源,如何规范、快速地构造语言工具和相关资源,并使之在研究中被广泛采纳应用,是进一步研究的一个重点。

本文介绍了在文本情感倾向分析中的多种特征以及相应的方法。针对从词语级别的特征,到句子级的特征,再到文本篇章级别的特征,人们提出了各种各样的算法。但在这些人工选取的特征和方法中,还没有哪一种能够完美地解决文本情感倾向识别这一问题。这也从侧面说明这一研究课题的复杂性。所以今后还应进一步细分具体问题及其应用领域,研究发现文本情感倾向的特征,继续寻找更有针对性、更简捷方便的识别方法。

文本情感倾向分析是目前在自然语言理解技术中最贴近实际应用需求的研究方向。但是由于语言处理技术在这一应用领域发展得还不够成熟,因此现有的许多情感分析系统,或意见挖掘系统在实际商业化的应用中还难以担当大任。目前在应用中最大的困难就是,还没有某个系统能够在情感倾向分析判断中具有较高的准确率,所以未来的研究工作显然还是要以提高系统的识别准确率为核心,争取把系统发展成为多领域、跨语言的强健的情感分析处理平台。

总之,文本情感分析研究是近几年自然语言理解领域的一个新的研究方向,不但具有重要的理论研究意义,而且具有广阔的应用前景。

参考文献:

[1] HATZIVASSILOGLOU V, MCKEOWN K. Predicting the semantic orientation of adjectives[C]// Proceedings of Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 1997: 174 – 181.

[2] WIEBE J, BRUCE R, MATTHEW B, *et al*. A corpus study of evaluative and speculative language[C]// Proceedings of the Second SIGdial Workshop on Discourse and Dialogue. Stroudsburg, PA: Association for Computational Linguistics, 2001: 186 – 195.

[3] HATZIVASSILOGLOU V, WIEBE J. Effects of adjective orientation and gradability on sentence subjectivity[C]// Proceedings of the 18th Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2000: 299 – 305.

[4] WIEBE J. Learning subjective adjectives from corpora[C]// Proceedings of National Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2000: 735 – 741.

[5] WIEBE J, WILSON T, BELL M. Identifying collocations for recognizing opinions[EB/OL]. [2012-06-20]. http://wenku.baidu.com/view/24e5e11cb7360b4c2e3f6416.html.

[6] WIEBE J, RILOFF E. Creating subjective and objective sentence classifiers from unannotated texts[C]// Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing. Berlin: Springer-Verlag, 2005: 486 – 497.

[7] WIEBE J, WILSON T, CARDIE C. Annotating expressions of opinions and emotions in language[J]. Language Resources and Evaluation, 2005, 39(2/3): 164 – 210.

[8] PANG B, LILLIAN L, SHIVAKUMAR V. Thumbs up: sentiment classification using machine learning techniques[C]// Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2002: 79 – 86.

[9] RILOFF E, WIEBE J. Learning extraction patterns for subjective expressions[C]// Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2003: 105 – 112.

[10] TURNEY P, LITTMAN M L. Measuring praise and criticism: Inference of semantic orientation from association[J]. ACM Transactions on Information Systems, 2003, 21(4): 315 – 346.

[11] DING X, LIU B, YU P S. A holistic lexicon-based approach to opinion mining[C]// Proceedings of the Conference on Web Search and Web Data Mining. New York: Association for Computing Machinery, 2008: 231 – 240.

[12] GANAPATHIBHOTLA M, LIU B. Mining opinions in comparative sentences[C]// Proceedings of International Conference on Compu-

tational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2008: 241 – 248.

[13] LU Y, CASTELLANOS M, DAYAL U. Automatic construction of a context-aware sentiment lexicon: an optimization approach[C]// Proceedings of the 20th International Conference on World Wide Web. Stroudsburg, PA: Association for Computational Linguistics, 2011: 347 – 356.

[14] TURNEY P D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews[C]// Proceedings of Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2002: 417 – 424.

[15] MILLER G A, BECKWITH R, FELLBAUM C, *et al*. WordNet: An on-line lexical database [J]. International Journal of Lexicography, 1990, 3(4): 235 – 244.

[16] KIM S M, HOVY E. Automatic detection of opinion bearing words and sentences[C]// Proceedings of the International Joint Conference on Natural Language Processing. Berlin: Springer, 2005: 61 – 66.

[17] KIM S M, HOVY E. Identifying and analyzing judgment opinions [C]// Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference. Stroudsburg, PA: Association for Computational Linguistics, 2006: 200 – 207.

[18] YU H, HATZIVASSILOGLOU V. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences[C]// Proceedings of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2003: 129 – 136.

[19] ANDREEVSKAIA A, BERGLER S. Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from WordNet glosses[C]// Proceedings of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2006: 209 – 216.

[20] SU F, MARKERT K. Subjectivity recognition on word senses via semi-supervised mincuts[C]// NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2009: 1 – 9.

[21] ESULI A, SEBASTIANI F. Determining the semantic orientation of terms through gloss analysis[C]// Proceedings of the ACM SIGIR Conference on Information and Knowledge Management. New York: ACM, 2005: 617 – 624.

[22] ESULI A, SEBASTIANI F. Determining term subjectivity and term orientation for opinion mining[C]// Proceedings of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2006: 193 – 200.

[23] KAMPS J, MARX M, MOKKEN R J. Using WordNet to measure semantic orientation of adjectives[C]// Proceedings of the Conference on Language Resources and Evaluation. Lisbon, Portugal: Conference on Language Resources and Evaluation, 2004: 1115 – 1118.

[24] KAMPS J, MARX M. Words with attitude[C]// Proceedings of the 1st International Conference on Global WordNet. Mysore, India: Central Institute for Indian Languages, 2002: 332 – 341.

[25] KU L W, LO Y S, CHEN H. Test collection selection and gold standard generation for a multiply-annotated opinion corpus[C]// Proceedings of the Association for Computational Linguistics Demo and Poster Sessions. Stroudsburg, PA: Association for Computational Linguistics, 2007: 89 – 92.

[26] READ J. Using emoticons to reduce dependency in machine learning techniques for sentiment classification[C]// Proceedings of the ACL Student Research Workshop. Stroudsburg, PA: Association for Computational Linguistics, 2005: 115 – 124.

[27] KOPPEL M, SHTRIMBERG I. Good news or bad news? Let the market decide[C]// Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications. Stanford: AAAI Press, 2004: 86 – 88.

[28] GHOSE A, IPEIROTIS P G, SUNDARARAJAN A. Opinion mining using econometrics: A case study on reputation systems[C]// Proceedings of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2007: 416 – 423.

[29] PANG B, LEE L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts[C]// Proceedings of Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2004: 271.

[30] PANG B, LEE L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales[C]// Proceedings of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2005: 115 – 124.

[31] WILSON T, WIEBE J, HOFFMANN P. Recognizing contextual polarity in phrase-level sentiment analysis[C]// HLC '05: Proceedings of the Human Language Technology Conference and Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2005: 347 – 354.

[32] MACDONALD C, OUNIS I. The TREC Blogs 06 collection: Creating and analysing a blog test collection[R]. Glasgow, Scotland: University of Glasgow, Department of Computer Science, 2006.

[33] SEKI Y, EVANS D K, KU L W, *et al*. Overview of opinion analysis pilot task at NTCIR-6[C]// Proceedings of the Workshop Meeting of the National Institute of Informatics Test Collection for Information Retrieval Systems. Tokyo: National Center of Science, 2007: 265 – 278.

[34] HU M, LIU B. Mining opinion features in customer reviews[C]// Proceedings of the National Conference on Artificial Intelligence. San Jose, California: AAAI Press, 2004: 755 – 760.

[35] STONE P. The general inquirer: A computer approach to content analysis[J]. Journal of Regional Science, 1968, 8(1): 113 – 116.

[36] HU M, LIU B. Mining and summarizing customer reviews[C]// Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC: ACM, 2004: 168 – 177.

[37] MOHAMMAD S M, TURNEY P D. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon[C]// Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. Stroudsburg, PA: Association for Computational Linguistics, 2010: 26 – 34.

[38] WIEBE J, BRUCE R F, O'HARA T P. Development and use of a gold-standard data set for subjectivity classifications[C]// Proceedings of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 1999: 246 – 253.

(下转第1607页)

4 结语

本文研究轨迹数据的概化方法,针对轨迹数据概化中区域划分的区域范围不能有效控制以及覆盖网格尺度不能合理选择的问题,提出了 LMG 方法来多次划分样本密集的区域,使各个区域之间的轨迹点密度基本一致。在此基础上提出了 TRAGenLMG 方法,在时间约束下合并连续往复通过的邻接区域,生成概化轨迹。使用真实数据集进行实验,从信息丢失量和概化轨迹聚类与原始轨迹聚类的结果相似性两个方面进行概化效果的衡量。实验结果显示: TRAGenLMG 在一定程度上保持了原始轨迹信息的同时,对于后续挖掘处理具有较好的适用性,效率相对于原始轨迹较高。

参考文献:

[1] ANDRIENKO N, ANDRIENKO G. Spatial generalization and aggregation of massive movement data[J]. Visualization and Computer Graphics, 2011, 17(2): 205 – 219.

[2] ZHANG L, YANG G, WANG Z C. Trajectory clustering based on spatial generalization[J]. Journal of Information & Computational Science, 2012, 9(2): 315 – 32.

[3] GIANNOTTI F, NANNI M, PEDRESCHI D, *et al*. Trajectory pattern mining[C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2007: 330 – 339.

[4] 唐良, 唐常杰, 姜页希, 等. TRAODGrid: 基于 Grid 空间划分的高效离群轨迹检测方法[J]. 计算机研究与发展, 2008, 45(10): 185 – 190.

[5] SHERKAT R, LI J, MAMOULIS N. Efficient time stamped event sequence anonymization[EB/OL]. [2012-08-20]. http://www.cs.hku.hk/research/techreps/document/TR-2011-02.pdf.

[6] ASSAM R, SEIDL T. Preserving privacy of moving objects via temporal clustering of spatio-temporal data streams[C]// Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS. New York: ACM, 2011: 9 – 16.

[7] Unisys Weather [DB/OL]. [2012-05-23]. http://weather.unisys.com/hurricane/atlantic/.

[8] STEFANAKIS E. Trajectory generalization under space constraints [EB/OL]. [2012-08-20]. http://www.giscience.org/proceedings/abstracts/giscience2012_paper_74.pdf.

[9] MASCIARI E. A framework for trajectory clustering[C]// Proceedings of the 3rd International Conference on GeoSensor Networks. Berlin: Springer-Verlag, 2009: 102 – 111.

[10] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48 – 61.

[11] 袁冠,夏士雄,张磊,等. 基于结构相似度的轨迹聚类算法[J]. 通信学报, 2011, 32(9): 103 – 110.

[12] MARTINEZ-BEA S. Trajectory anonymization from a time series perspective[C]// 2011 IEEE International Conference on Fuzzy Systems. Piscataway: IEEE, 2011: 401 – 408.

[13] ANDRIENKO G, ANDRIENKO N, GIANNOTTI F, *et al*. Movement data anonymity through generalization[J]. Journal of Transactions on Data Privacy, 2010, 3(2): 91 – 121.

(上接第 1578 页)

[39] HATZIVASSILOGLOU V, WIEBE J. Effects of adjective orientation and gradability on sentence subjectivity [C]// Proceedings of International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2000: 299 – 305.

[40] HATZIVASSILOGLOU V, KLAVANS J L, HOLCOMBE M L, *et al*. SIMFINDER: A flexible clustering tool for summarization [C]// Proceedings of the Workshop on Summarization in NAACL-01. Stroudsburg, PA: Association for Computational Linguistics, 2001: 41 – 49.

[41] BENAMARA F, CHARDON B, MATHIEU Y, *et al*. Towards context-based subjectivity analysis [C]// Proceedings of the 5th International Joint Conference on Natural Language Processing. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, 2011: 1180 – 1188.

[42] TURNEY P D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews [C]// Proceedings of Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2002: 417 – 424.

[43] KIM S M, HOVY E. Determining the sentiment of opinions [C]// Proceedings of the 20th International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2004: 1367 – 1373.

[44] PANG B, LEE L. Opinion mining and sentiment analysis [J]. Journal Foundations and Trends in Information Retrieval, 2008, 2(2): 1 – 135.

[45] CHOI Y, CARDIE C, RILOFF E. Identifying sources of opinions with conditional random fields and extraction patterns [C]// Proceedings of HLT/EMNLP-2005. Stroudsburg, PA: Association for Computational Linguistics, 2005: 355 – 362.

[46] BETHARD S, YU H, THORNTON A. Automatic extraction of opinion propositions and their holders [C]// Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text. Stanford: American Association for Artificial Intelligence, 2004: 22 – 24.

[47] PANG B, LEE L. Using very simple statistics for review search: An exploration [C]// Proceedings of International Conference on Computational Linguistics. Manchester, UK: Coling 2008 Organizing Committee, 2008: 75 – 78.

[48] DAVE K, LAWRENCE S, PENNOCK D M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews [C]// Proceedings of the 12th International World Wide Web Conference. New York: ACM, 2003: 519 – 528

[49] LIU B, HU M, CHENG J. Opinion observer: analyzing and comparing opinions on the Web [C]// Proceedings of the 14th International Conference on World Wide Web. New York: ACM, 2005: 342 – 351.

[50] WILSON T, HOFFMANN P, SOMASUNDARAN S, *et al*. OpinionFinder: A system for subjectivity analysis [C]// Proceedings of HLT/EMNLP on Interactive Demonstrations. Stroudsburg, PA: Association for Computational Linguistics, 2005: 34 – 35.

[51] GAMON M, AUE A, CORSTON-OLIVER S, *et al*. Pulse: Mining customer opinions from free text [C]// Proceedings of the 6th International Symposium on Intelligent Data Analysis. Berlin: Springer-Verlag, 2005: 121 – 132.

[52] CHERRY C, MOHAMMAD S. Binary classifiers and latent sequence models for emotion detection in suicide notes [J]. Journal of Biomedical Informatics Insights, 2012, 5(S1): 147 – 154.