

细颗粒度观点挖掘中的观点句识别 与要素抽取研究综述*

陈锋

(南开大学商学院, 天津 300071)

摘要: 细颗粒度观点挖掘需要解决观点句识别、要素抽取等关键问题。论文对国内外相关研究进行梳理, 归纳现有观点句识别方法并找出影响识别率的原因, 分析基于规则和基于统计的两类方法进行要素抽取的优势和不足。研究发现, 机器学习方法或人工和机器相结合的方法是要素识别研究的基本趋势, 观点本体和观点表达语言规律的研究需要进一步加强, 应用领域有待扩展, 观点抽取结果挖掘需进一步深入。

关键词: 观点挖掘; 要素抽取; 观点句识别

中图分类号: G250.2

DOI: 10.3772/j.issn.1673-2286.2015.10.004

随着互联网的发展和应用, 互联网上的文本数量激增, 其中含有大量评论。公司和政府等组织也开始利用这些评论挖掘人们的意见和情感倾向, 用以支持决策。观点挖掘是人们针对某个实体及其特征发表的意见、态度、情感的挖掘和分析^[1]。观点挖掘也称意见挖掘、态度挖掘或评论挖掘。如果进一步对挖掘出的观点进行情感倾向性分析, 可以称为情感计算或文本倾向性分析。观点挖掘的自动化也成为重要的研究领域, 与自然语言处理、信息检索、机器学习、知识组织和知识服务等均有一定的相关性。

按照观点挖掘的颗粒度可以将其分为篇章级别、句子级别和要素级别 (aspect-based) 的观点挖掘。篇章级别的观点挖掘认为一篇文档 (如一条微博或一条商品评论) 是一个基本信息单位, 假定一个篇章表达了一个总体的意见。句子级别的观点挖掘把一个句子作为一个基本信息单位, 假定一个句子表达一条意见。尽管篇章级别和句子级别的文本挖掘在很多领域得到了广泛应用, 但是对文本观察可以看出, 一个总体上表达肯定态度的文章并不意味着对事物的每一个方面都是肯定的。为了获得更精确的观点挖掘结果, 需要更细粒

度的观点挖掘, 即要素级别的观点挖掘。

实现要素级的观点挖掘, 需要解决观点句识别、要素抽取和情感分析等若干关键问题。情感分析的研究较早, 成果较丰富, 相关介绍也比较多。观点句识别和要素抽取是新的研究热点。根据第五届中文汉语倾向性评测会议 (Chinese Opinion Analysis Evaluation, COAE) 的测评结果, 对非限定领域的微博中观点句和要素抽取的准确率和召回率均不理想, 低于50%^[2]。因此, 本文主要介绍观点句识别和要素抽取的主要技术和研究进展, 希望有助于汉语文本观点挖掘的研究。

1 观点及要素的定义

观点 (opinion) 是人们对事物、人物、事件等的意见。观点的形式化定义常用四元组、五元组和六元组。Kim和Hovy使用四元组定义观点, 观点由四个元素组成: 主题 (topic)、持有者 (holder)、陈述 (claim)、情感 (sentiment)^[3]。一条观点就是某人 (holder) 对某个主题 (topic) 发表了一段含有特定情感 (sentiment) 的陈述 (claim)。Liu Bing和Zhang Lei提出观点模型的

* 本研究得到国家社会科学基金重大项目“我国网络社会治理研究” (编号: 14ZDA063) 资助。

五元组定义^[4]: 观点(实体entity, 要素aspect, 意见倾向opinion orientation, 持有者holder, 时间time)。Liu Bing和Zhang Lei也提出六元组的定义^[5]: 观点(实体甲entity1, 实体乙entity2, 要素aspect, 对比表达式PE, 意见持有者holder, 时间time)。六元组针对比较型的意见表达, 比如“三星手机的屏幕比苹果手机的更大”。

要素(aspect)又称为“特征”(feature), 或译为“方面”。在Liu Bing和Zhang Lei提出观点模型的五元组定义中, 要素是实体的部分和属性, 也可以是实体本身; 实体可以是产品、服务、人、事件、组织、话题等等^[4]。

观点挖掘的过程就是在文本中自动抽取这些元素, 其子任务主要有: 观点句识别、实体抽取、要素抽取、情感倾向判断和统计分析。

2 观点句识别方法

观点句是表达观点的句子。在篇幅较长的商品评论或新闻报道中, 或者在微博、论坛等情景下, 用户的观点与无关的陈述(如“我昨天刚买了一部苹果手机”)、广告等混在一起, 需要进行观点句识别。

现有的研究一般把观点句识别看做句子分类问题, 利用评价词或情感词词典、句法特征、上下文、主题等资源, 使用SVM(Support Vector Machine, 支持向量机)^[6-10]、CRF(Conditional Random Fields, 条件随机场)^[7,11-12]、贝叶斯分类器、最大熵、决策树等分类工具^[13]进行分类。几种分类工具综合使用的效果略高于仅使用单一工具^[13-14]。

选用较多的观点句特征可以有效提高识别效果。在对中文微博的观点句识别中, 除了评价词外, 加入语气词、程度副词和词性结构等特征后, 召回率显著提高^[15]。在各类特征中, 评价词是重要的识别依据, 因而评价词词典或情感词词典对性能的影响比较大。丁晟春、赵洁等利用HowNet^[11]、网络新词挖掘方法^[9]对评价词词典进行扩充, 以改善观点句的识别效果。

比较句是一类特殊的观点句。宋锐等针对比较句这一类型, 通过构建中文比较模式库以实现中文比较句的自动识别, 选取比较主体、比较客体及其上下文的词、词性、位置、语义等特征以及领域知识, 利用条件随机场模型进行中文比较关系抽取^[16]。

客观句也可以表达观点。观点句并不一定包含评价词和情感词。一些句子虽然表达某种客观事实, 但隐含了评价和情感倾向。例如“这家酒店在外环路以

外”, 隐含了对酒店位置偏远的评价。许延祥等按表达功能把句子分为4种类型: 客观句、主观句、混合句、组织句, 并且通过人工标注后统计发现, 36%的观点通过客观句表达^[17]。客观句中通常没有评价词和情感词, 成为识别观点句的难点。这也是观点句识别召回率较低的重要原因。

3 要素抽取方法

实体和要素抽取任务与信息抽取任务很相似, 但也有很大不同。传统的信息抽取一般从格式较为固定的文本中抽取, 如新闻、科技报告等等。但是目前的观点挖掘大多处理互联网上的格式自由的文本, 如博客、商品评论等。这类文本通常短小, 无固定行文格式, 语法灵活, 用词新颖, 而且混杂很多无关的内容。一般而言, 要素抽取可以分解为三个子任务: 一是确定要素集合; 二是从待分析文本中抽取要素表达式; 三是对要素或要素表达式进行分类、排序等处理。

3.1 确定要素集合

要素集合是实体所有要素的集合, 或称要素词典、特征词典。要素集合并非是所有要素表达式的集合。

确定要素的方法有三类: 一是人工发现和归纳, 二是机器自动识别, 三是人工归纳和机器发现相结合。研究表明, 一些实体的要素是有限的, 比如产品评论的要素通常就是组件、功能、服务; 这些实体的要素(特征)集合可以使用人工归纳的方法建立^[17]。使用人工方法准确率较高, 其缺陷是费时费力。人工归纳出部分要素后, 也可以使用一些同义词典等语言资源进行扩充, 例如Bloomd等手工构建了一个要素库, 从少量文本样本中发现一些特征词, 然后使用WordNet对特征词进行扩展^[18]。

机器自动识别方法把要素抽取视为聚类问题, 借用文本聚类算法, 例如使用K-mean算法^[19]或Apriori算法^[20]。实验表明使用聚类算法有效但不理想, 优点是不需要人工标注训练语料。另有一些算法利用了词性特征和其他资源来进行判断, 例如Li等使用百度搜索引擎的搜索结果计算候选特征和实体之间的互信息, 用Rcut算法来评价候选特征的阈值^[21]。该方法的优点是人工干预较少, 可以方便地更换实体或领域。

人工和机器相结合的方法可视为一种弱监督的机

器学习方法,首先由人工给出一些规则或要素作为种子,然后机器在语料中学习去发现新的要素。Hu和Liu提出基于关联规则来抽取表达式^[22]。该方法利用高频次出现的要素和评价词的关联关系去发现低频次的要素。Zhuang等使用该方法抽取了电影评论,取得了较好的效果^[23]。Wang等在该方法基础上引入了互信息识别汉语产品评论中的特征词和评价词^[24]。Zhang等对该方法进行了改进,使用了更多语言规则来发现特征^[25]。Li等利用一个标记好的某领域的语料库去抽取另一个新领域的特征^[26]。Popescu和Etzioni利用名词和鉴别短语(discriminator)的互信息(PMI, Pointwise mutual information)来发现新特征,使用WordNet和形态学特征发现实体的部分或属性^[27]。人工和机器学习相结合的方法可以充分利用要素或特征词的使用频次、句法特征、搭配关系等等特性,以较少的人工标注语料,通过种子集和大量的易获取的未标注语料来获得新要素。

汉语与英语的差异之一是没有形态学特征而且书写不分词,汉语文本通常需要机器分词。机器分词的错误会给汉语要素抽取带来负面影响。如果要素是复合词或未登录词(机器分词词表中没有的词),那么分词后的文本一般会被切分成连续的词语或单个汉字(也称分词碎片),进而影响要素词识别。例如“该手机的摄像头对焦速度很快”,分词后的文本是“该/手机的摄像头/对焦/速度/很快”,“摄像头对焦速度”这个要素被切分成了3个词,不容易被作为一个要素识别出来。郑敏洁等提出一种层叠条件随机场的抽取方法来解决该问题,准确率、召回率和F1值分别提升1.62%、5.75%和4.17%^[28]。

3.2 要素表达式的抽取方法

文本中对实体的评价可以有两种形式:一是显性的表达,如“这部电影的情节很有趣”;或者隐性的表达,如“这是一部有趣的电影”或“手机很大”,其要素“情节”和“尺寸”没有出现。对这两类表达,一般采用不同的抽取策略。

3.2.1 显性表达式抽取方法

显性特征表达式的抽取方法可以分为基于语言规则的理性主义方法和基于统计的经验主义方法。经验

主义方法主要使用隐马尔科夫、条件随机场等序列模型(sequence model)和LDA等话题模型(topic model)。

理性主义的方法利用语言规则来构建抽取模板。许延祥通过人工总结规则,构建了一个基于语言规则的意见表达式抽取系统,意见挖掘准确度为85.81%,召回率为39.08%^[29]。

隐马尔科夫模型(HMM, Hidden Markov Model)把词和短语视为观察序列,特征或观点表达式视为隐藏状态。Jin等利用该方法从产品评论中抽取要素和要素表达式^[30-31]。隐马尔科夫模型的缺陷是能够利用的语言特征比较有限。

条件随机场(CRF)可以引入更多的语言特征。Jakob和Gurevych使用条件随机场模型抽取要素,利用了多项语言学特征:词性、依存语法成分、词的距离^[32]。Li等还尝试使用条件随机场抽取联合短语、长句和复句中的特征^[33]。吕品等利用条件随机场实现中文产品评论的方面级(即要素级颗粒度)观点挖掘,结果显示,用CRF对中文评论进行要素级观点挖掘和对英文评论的要素级观点挖掘的性能差异不大,优于词典化的隐马尔科夫模型和关联规则挖掘方法^[34]。张莉等把专家规则和机器学习相结合,首先确定10种基本句法关系作为语言学特征,然后利用CRF进行机器学习,最后用于评价对象抽取^[35]。王荣洋等在使用CRF模型时引入了语义角色信息,如施事者、受事者等,对电影评论的实验表明该方法能够提高抽取性能^[36]。戴敏等引入更多句法特征来训练CRF,性能也有小幅提升^[37]。

话题模型(Topic Model)方法把特征视为文本中潜藏的话题。大量研究在pLSA(Probabilistic Latent Semantic Analysis)模型^[38]和LDA(Latent Dirichlet Allocation)模型^[39]的基础上进行参数调整或扩展,把话题模型应用于要素抽取。Lu等使用pLSA对短文本进行要素表达式的识别和聚类,把要素表达式定义为二元组<首词,修饰语>(head term, modifier)^[40]。LDA模型参数少,可以避免被过度拟合训练文本。Wang等使用LDA模型分两步处理两类不同的训练语料,首先采集并标注电子商务网站上内容完整的专业评论作为种子集来进行机器学习,然后根据得到的种子特征对大量的用户评论进行聚类 and 训练,取得了较好的效果^[41]。但Titov和Mcdonald发现某类产品的评论中每篇文档都讨论同样的要素,使得一般的LDA模型准确率不高,他们提出多粒度话题模型(Multi-grain LDA)弥补该缺陷^[42]。Mcdonald等进一步改进了该模型,以解决

话题到要素的映射问题^[43]。目前已有一些研究使用LDA模型处理汉语文本。李芳等使用LDA模型对中文文本的汽车评论进行了评价主题(特征)挖掘,实验结果中褒义句的正确率为76.2%,召回率为51.7%,而贬义句的正确率为55.4%,召回率为38.2%^[44]。话题模型易于理解,便于计算,参数可以调节,具有较好的弹性。话题模型的缺陷是很难识别低频次的要素,另外一个缺陷是需要训练语料进行机器学习。

另外一些方法综合了规则和统计方法,或者利用了本体、互联网等其他资源。例如Li等把要素抽取作为浅层语义分析问题,对每个句子构建句法树(parse tree),用以识别要素^[45]。Lu等利用了在线的本体库Freebase来抽取要素表达式^[46]。Ma和Wan利用向心理论(Centering theory)^[47]从汉语新闻评论中抽取观点^[48]。刘鸿宇等综合多种方法抽取评价对象,首先使用句法分析结果获取候选评价对象,继而结合基于互联网挖掘的PMI算法和名词剪枝算法对候选评价对象进行筛选^[49]。

3.2.2 隐性要素表达式的抽取

隐性要素表达一般有两种方式:一是使用评价词(sentiment words)直接修饰实体,如“手机很贵”,“这个宾馆的房间很脏”;另一种是陈述事实,如“iPhone 6 plus不适合装进衣服口袋”。陈述事实的表达方式很难抽取隐藏的要素,目前尚未看到相关研究。对第一种方式,基本方法是寻找评价词和要素的对应关系,对应方法有三种:词典法、分类法和聚类法。Fei等使用基于词典的方法来解决形容词类的隐性要素问题^[50]。刘慧慧等把隐性要素表达式的挖掘视为评价词的分类问题,选用词法和依存句法作为特征,用决策树做训练分类器^[51]。Hai等提出一个两阶段共现规则聚类方法,首先挖掘评价词和显性表达式在训练语料中的共现关系,然后对规则聚类以生成每个评价词与要素的关联规则^[52]。Su等提出一种迭代聚类方法,最终得到关系最强的评价词和要素的映射关系^[53]。

3.3 要素的归类问题

人们使用不同的词语表达同一类特征,使用上面介绍的方法进行特征识别后,还需要对特征进行归类。尽管可以使用同义词词典这样的语言学资源,但归类并

不是简单的合并同义词和近义词,因为确定两个词是不是一类特征,是和领域相关的。在归类过程中,根据应用的需求,还可以对要素排序或以树形结构整理要素集合。

要素的归类方法可以分为三类:一是基于同义词词典等语义资源;二是依据领域相关性强的本体;三是利用聚类算法。Liu等^[54]和Carenini等^[55]都使用WordNet做资源解决要素聚类问题,效果并不理想。Zhai等分别使用半监督的算法和非监督的算法来对特征归类^[56-57],两种算法都利用了同义词和近义词词典,后者利用了更多的自然语言知识来发现有助于归类的上下文。Wei和Gulla使用产品的本体知识归类产品的特征(要素)^[58]。冯淑芳和王素格构建了汽车评价本体知识库,可以用于汽车评论的要素聚类和层次化组织^[59]。Guo等提出多级LSA模型来归类产品要素表达式,该方法首先把要素表达式归类为话题集合,然后依据该话题集合的信息和评论中的上下文信息用LDA模型对要素表达式再次分类^[60]。吕韶华等提出一种基于LDA模型的餐馆评论排序方法^[61]。计算相似性和相关性的常用算法也被应用于要素聚类和排序,例如Yu等用计算语义距离的方法来构建特征树^[62],Zhang等利用特征与评价词的相关性和特征的出现频率来决定特征的重要程度^[63],Yu等依据评论者的数量和特征对整条评论情感倾向的影响程度来对特征排序^[64]。

4 结语

在商业应用的驱动下,要素级观点挖掘成为研究热点,然而目前抽取效果依然欠佳。观点句识别和要素抽取等研究有以下不足:

一是对语言规律研究不够深入。观点挖掘的对象是自然语言文本,但是目前从语言学角度对评论文本的研究还不多,研究缺少大规模标注语料的支持。

二是对观点本体的研究不足,缺少可共享的要素集合资源。特定实体的要素通常是个封闭集合,如果有可共享的要素集合资源,研究者可以集中精力进行要素表达式的挖掘。

三是观点挖掘的应用领域需要扩展。目前的应用和研究主要集中于产品和服务评论,服务于市场营销领域,对公共政策、学术思想的观点挖掘研究仍然较少。

四是在观点元组被挖掘后,如何统计分析观点元

组中的信息,实现抽取出的观点最大化利用,目前研究有待深化。

从近年的研究可以发现,半监督的机器学习方法成为研究趋势。客观句的识别和以陈述事实为表达方式的隐性要素表达式抽取依然是非常有挑战的工作。如果要提高观点挖掘的召回率,必须突破该难题。

参考文献

- [1] Pang B, Lee L. Opinion Mining and Sentiment Analysis[J]. Foundations and Trends in Information Retrieval, 2008, 2(1): 1-135.
- [2] 谭松波,王素格,廖祥文,刘康. 第五届中文倾向性分析评测总体报告[C]//谭松波,王素格,廖祥文,刘康. 第五届中文倾向性分析评测报告论文集. 太原: 第五届中文倾向性分析评测研讨会, 2013: 5-53.
- [3] Kim S M, Hovy E. Determining the Sentiment of Opinions[C]// Proceedings of International Conference on Computational Linguistics, 2004.
- [4] Zhang L, Liu B, Liu B, Zhang L. A Survey of Opinion Mining and Sentiment Analysis[C]// Aggarwal C C, Zhai C X. Mining Text Data. New York: Springer, 2012: 434-499.
- [5] Zhang L, Liu B. Aspect and Entity Extraction for Opinion Mining[C]// Data Mining and Knowledge Discovery for Big Data. New York: Springer Heidelberg, 2014: 1-35.
- [6] 杜锐,朱艳辉,鲁琳,等. 基于SVM的中文微博观点句识别算法[J]. 湖南工业大学学报, 2013 (2): 89-93.
- [7] 周海云. 基于多特征融合的中文比较句识别算法[J]. 中文信息学报, 2013(6): 110-116.
- [8] 王乐, 闭应洲. 基于特征模板提取及SVM的观点句识别[J]. 广西师范学院学报(自然科学版), 2014(3): 85-89.
- [9] 赵洁, 温润. 基于新词扩充和特征选择的微博观点句识别方法[J]. 情报学报, 2013(9): 945-951.
- [10] 潘艳茜, 姚天昉. 微博汽车领域中用户观点句识别方法的研究[J]. 中文信息学报, 2014(5): 148-154.
- [11] 丁晟春, 文能, 蒋婷, 等. 基于CRF模型的半监督学习迭代观点句识别研究[J]. 情报学报, 2012(10): 1071-1076.
- [12] 林鸿飞. 基于组块分析的评价对象识别及其应用[J]. 广西师范大学学报(自然科学版), 2011(1): 151-156.
- [13] 吕云云, 李珣, 王素格. 基于BootStrapping的集成分类器的中文观点句识别方法[J]. 中文信息学报, 2013(5): 84-92.
- [14] 郭云龙, 潘玉斌, 张泽宇, 等. 基于证据理论的多分类器中文微博观点句识别[J]. 计算机工程, 2014(4): 159-163, 169.
- [15] 丁晟春, 孟美任, 李霄. 面向中文微博的观点句识别研究[J]. 情报学报, 2014(2).
- [16] 宋锐, 林鸿飞, 常富洋. 中文比较句识别及比较关系抽取[J]. 中文信息学报, 2009, 23(2): 102-107.
- [17] 许延祥, 罗铁坚, 周佳, 等. 评价文本中意见分布规律研究[J]. 中文信息学报, 2014, 28(3): 150-158.
- [18] Bloom K, Grag N, Argamon S. Extracting Appraisal Expressions[C]// Proceedings of the 2007 Annual Conference of the North American Chapter of the ACL, 2007.
- [19] 韩雪婷, 李炜, 沈奇威. 用户评论中产品特征的抽取及聚类[J]. 计算机系统应用, 2013, 22(5): 188-203.
- [20] 周红庆, 吴扬扬. 中文客户评论对象特征的抽取与聚类方法[J]. 微型机与应用, 2014, (15): 69-71.
- [21] Li S, Zhou L, Li Y. Improving Aspect Extraction By Augmenting a Frequency-based Method with Web-based Similarity Measures[J]. Information Processing & Management, 2015, 51(1): 58-67.
- [22] Hu M, Liu B. Mining Opinion Features in Customer Reviews[C]// Proceedings of National Conference on Artificial Intelligence, 2004.
- [23] Zhuang L, Jing F, Zhu X. Movie Review Mining and Summarization[C]// Proceedings of National Conference on Artificial Intelligence, 2006.
- [24] Wang B, Wang H. Bootstrapping both product features and opinion words from Chinese customer reviews with cross-inducing[J]. Computational Linguistics, 2011, 37(1): 9-27.
- [25] Zhang L, Liu B, Lim S H, et al. Extracting and Ranking Product Features in Opinion Documents[J]. Extracting and Ranking Product Features in Opinion Documents. - ResearchGate, 2010.
- [26] Li F, Pan S, Jin Q, et al. Cross-Domain Co-Extraction of Sentiment and Topic Lexicons[C]// Proceedings of Annual Meeting of the Association for Computational Linguistics, 2012.
- [27] Popescu A, Etzioni O. Extracting Product Features and Opinions from Reviews[C]// Proceedings of Conference on Empirical Methods in Natural Language Processing, 2005.
- [28] 郑敏洁, 雷志城, 廖祥文, 等. 基于层叠CRFs的中文句子评价对象抽取[J]. 中文信息学报, 2013(3): 69-76.
- [29] 许延祥. 评价文本中意见表达规律的研究与应用[D]. 北京: 中国科学院研究生院, 2012.
- [30] Jin W, Ho H, Srihari R. Opinionminer: A Novel Machine Learning System for Web Opinion Mining and Extraction[C]// Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009.
- [31] Jin W, Ho H. A Novel Lexicalized HMM-Based Learning Framework for Web Opinion Mining[C]// Proceedings of International Conference

- on Machine Learning, 2009.
- [32] Jakob N, Gurevych I. Extracting Opinion Targets in a Single and Cross-Domain Setting with Conditional Random Fields[C]// Proceedings of Conference on Empirical Methods in Natural Language Processing, 2010 .
- [33] Li F,Han C,Huang M, et al. Structure-aware Review Mining and Summarization[C]// Proceedings of International Conference on Computational Linguistics, 2010 .
- [34] 吕品,钟珞,蔡敦波,等.基于CRF的中文评论有效性挖掘产品特征[J]. 计算机工程与科学, 2014 (2): 359-366.
- [35] 张莉,钱玲飞,许鑫.基于核心句及句法关系的评价对象抽取[J]. 中文信息学报, 2011(3): 23-29.
- [36] 王荣洋,鞠久朋,李寿山,等.基于CRFs的评价对象抽取特征研究[J]. 中文信息学报, 2012(2): 56-61.
- [37] 戴敏,王荣洋,李寿山,等.基于句法特征的评价对象抽取方法研究[J]. 中文信息学报, 2014, 28(4): 92-97.
- [38] Hofmann T. Unsupervised Learning by Probabilistic Latent Semantic Analysis[J]. Machine Learning, 2001, 42(1): 177-196.
- [39] Liu B,Zhang L. Latent Dirichlet Allocation[J]. The Journal of Machine Learning Research, 2003(3) .
- [40] Lu Y,Zhai C,Sundaresan N. Rated Aspect Summarization of Short Comments[C]// Proceedings of International Conference on World Wide Web, 2009.
- [41]Wang T , Cai Y, Leung H, et al. Product Aspect Extraction Supervised with Online Domain Knowledge[J]. Knowledge-based Systems, 2014, 71(0): 86-100.
- [42] Titov I,Mcdonald R. Modeling Online Reviews with Multi-Grain Topic Models.[C]// Proceedings of International Conference on World Wide Web , 2008.
- [43] Titov I,Mcdonald R. A Joint Model of Text and Aspect Ratings for Sentiment Summarization[C]//Proceedings of Annual Meeting of the Association for Computational Linguistics , 2008.
- [44] 李芳,何婷婷,宋乐.评价主题挖掘及其倾向性识别[J].计算机科学, 2012(6): 159-162.
- [45]Li S,Wang R,Zhou R. Opinion target extraction using a shallow semantic parsing framework.[C]//Proceedings of National Conference on Artificial Intelligence, AAAI 2012, 2012
- [46] Lu Y,Duan H,Wang H, et al. Exploiting Structured Ontology To Organize Scattered Online Opinions.[C]//Proceedings of International Conference on Computational Linguistics, 2010 .
- [47] Grosz J,Winstein S, Joshi A. Centering: A Framework for Modeling The Local Coherence of Discourse[J]. Computational Linguistics, 1995, 21(2): 203-225.
- [48] Ma T,Wan X. Opinion Target Extraction In Chinese News Comments.[C]//Proceedings of International Conference on Computational Linguistics, 2010 .
- [49] 刘鸿宇,赵妍妍,秦兵,等.评价对象抽取及其倾向性分析[J].中文信息学报, 2010 (1): 84-88, 122.
- [50] Fei G,Liu B,Hsu M, et al . A Dictionary-Based Approach To Identifying Aspects Implied By Adjectives for Opinion Mining[C]// Proceedings of International Conference on Computational Linguistics , 2012 .
- [51] 刘慧慧,王素格,赵策力等.观点句中评价对象/属性的缺省项识别方法研究[J]. 中文信息学报, 2014, 28(5): 175-182.
- [52] Hai Z,Chang K, Kim J. Implicit feature identification via co-occurrence association rule mining[C]//Computational Linguistic and Intelligent Text Processing, 2011 .
- [53] Su Q,Xu X,Guo H, et al. Hidden Sentiment Association In Chinese Web Opinion Mining[C]//Proceedings of International Conference on World Wide Web, 2008 .
- [54] Liu B,Hu M,Cheng J. Opinion Observer: Analyzing and Comparing Opinions On The Web[C]//Proceedings of International Conference on World Wide Web , 2005 .
- [55] Carenini G,Ng R,Zwart E. Extracting Knowledge from Evaluative Text[C]//Proceedings of Third International Conference on Knowledge Capture, 2005 .
- [56] Zhai Z,Liu B,Xu H, et al . Grouping Product Features Using Semi-Supervised Learning with Soft-Constraints[C]//Proceedings of International Conference on Computational Linguistics , 2011.
- [57] Zhai Z,Liu B,Xu H, et al. Clustering Product Features for Opinion Mining[C]//Proceedings of ACM International Conference on Web Search and Data Mining, 2011 .
- [58] Wei W, Gulla J. Sentiment Learning On Product Reviews Via Sentiment Ontology Tree.[C]//Proceedings of Annual Meeting of the Association for Computational Linguistics, 2010.
- [59]冯淑芳,王素格. 面向观点挖掘的汽车评价本体知识库的构建[J]. 计算机应用与软件, 2011, (5): 45-47.
- [60] Guo H,Zhu H,Guo Z, et al. Product Feature Categorization with Multilevel Latent Semantic Association[C]//Proceedings of ACM International Conference on Information and Knowledge Management, 2009 .
- [61]吕韶华,杨亮,林鸿飞. 基于LDA模型的餐馆评论排序[J]. 计算机工程, 2011, 第19期(19): 62-64.
- [62] Yu J, Zha Z,Wang M, et al. Domain-Assisted Product Aspect

- Hierarchy Generation: Towards Hierarchical Organization of Unstructured Consumer Reviews.[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing, 2011 .
- [63] Zhang L,Liu B,Lim S H, et al. Extracting and Ranking Product Features in Opinion Documents [C]// Proceedings of the 23rd International Conference on Computational Linguistics: Posters Association for Computational Linguistics, 2010:1462-1470.
- [64] Yu J, Zha Z J,Wang M, et al. Aspect Ranking: Identifying Important Product Aspects from Online Consumer Reviews.[C]//ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011:1496-1505.

作者简介

陈锋，男，1982年生，南开大学博士生，研究方向：文本挖掘、知识发现，E-mail: chenfengwork@163.com。

A Survey of Subjectivity Classification and Aspect Extraction in Fine-Grained Opinion Mining

CHEN Feng

(Department of Information Resource Management, Business School of Nankai University, Tianjin 300071, China)

Abstract: The key steps in fine-grained opinion mining are subjectivity classification and aspect extraction . By reviewing of domestic and foreign related research, the formulated definition of the opinion and aspect is made and the reasons of the low recognition rate are summarized. and then the advantages and disadvantages of two kinds of methods(based on rules and statistics) are analyzed. It is found that the semi-supervised machine learning method is a research trend. This paper proposes to strengthen the research on linguistic rules and the ontology of opinion, to expand the application fields and make the comprehensive utilization of the results.

Keywords: Opinion Mining; Subjectivity Classification; Aspect Extraction

(收稿日期: 2015-09-23; 编辑: 王立学)

声明

近日发现有人假冒《数字图书馆论坛》名义，收录稿件并收取高额版面费，此行为严重侵犯了本刊名誉及作者利益。为此，本刊郑重声明，对盗用本刊名义及私刻本刊公章的个人及代理机构，本刊将保留诉诸法律的权利。

敬请广大作者提高警惕，谨防上当受骗。

作者投稿请直接与本刊联系，在线投稿网址: <http://tougao.dlf.net.cn>; 编辑部邮箱: DLF@istic.ac.cn; 编辑部电话: 010-58882324。