情报学报 ISSN 1000-0135
第29卷第5期858-863，2010年10月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol.29 No.5, 858-863 October 2010

doi:10.3772/j.issn.1000-0135.2010.05.014

面向网络评论的观点主题识别研究[1)]

周 杰 林 琛 李弼程
（信息工程大学信息工程学院，郑州 450002）

摘要 网络评论的观点分析为及时掌握广大民众的真实观点提供了渠道。观点主题识别作为观点分析的重要组成部分，用以确定观点所指的对象。本文设计了一种领域无关的观点主题识别算法，该算法以网络评论中观点主题产生的方式为依据，采用由内到外的识别过程，分四个部分完成观点主题识别：内部主题词识别、内部主题构建、外部主题识别和主题的组织。算法能够克服分词和短语类主题带来的影响，识别出语义完整的观点主题。对实际网络评论语料进行测试的结果表明，本文的算法能够有效地识别网络评论中的观点主题。

关键词 中文信息处理 观点分析 网络评论 观点主题识别

Research on the Identification of Opinion Topic Expressed in Web Comments

Zhou Jie, Lin Chen and Li Bicheng
(*Information Engineering Institute, Information Engineering University, Zhengzhou* 450002)

Abstract Opinion analysis of network reviews provides a channel to find out the viewpoint of the common people in time, and opinion topic identification, as a significant part of opinion analysis, is aimed at identifying the objects for the expressed opinions. This paper proposes an opinion topic identification algorithm within independent domain. The algorithm, based on the mode of building opinion topic in network reviews, employs the process to identify the topics from inner to outer, and the process is finished through four steps: Inner topic word identification, Inner topic constitution, Outer topic identification and Topic organization. This algorithm can overcome the influence of word-segmentation error and phrase-topic, so it can get opinion topics which have the integrated semantic information. Experiment in real network reviews corpus proves that the algorithm can identify opinion topic in network reviews effectively.

Keywords Chinese information processing, opinion analysis, network reviews, opinion topic identification

1 引 言

文本的观点分析（Opinion Analysis）是利用计算机处理文本语言信息的一个新的研究方向，它自动识别文本发表者对事物、人物、事件等的个人（或群体）看法或情感倾向。目前，以网络评论为代表的主观性文本数量迅速增长，针对网络评论的观点分析具有广泛的实用价值，如网络舆情分析、产品质量评价、影视和服务评价等。

早期的观点分析研究主要是从整体角度判断句子或评论的情感倾向，它们假定所有评论都针对同一给定主题发表观点。但是评论中可能存在着多个评论主题，为了解决该类问题，相关研究提出一种细

收稿日期：2009年7月1日

作者简介：周杰，男，1984年生，硕士研究生，研究方向：网络舆情话题的观点分析。E-mail: zhoujie-0001@163.com。林琛，女，1981年生，博士研究生，研究方向：网络舆情信息分析。李弼程，男，1970年生，教授，博士生导师，研究方向：数据挖掘、海量信息检索、模式识别。

1) 国家863项目“网络舆情态势分析与预警关键技术研究”（No.2007AA01Z439）资助。

粒度观点分析(Fine-Grained Opinion Analysis)方法。该方法将观点细化为四个元素[1]:主题(Topic)、持有者(Holder)、陈述(Claim)和情感(Sentiment)。

观点主题一般认为是观点所指的对象。在产品评论领域,观点主题常常特指产品及其相关属性,比如"汽车 A"和"A 的造型"。在新闻评论中,观点主题涉及的范围更加广泛,它可以是社会问题、政府政策、热点事件或个人意见等。评论的主题不能片面理解,需要借助上下文确定主题信息。从这种角度看,评论的主题可以定义为一种依赖上下文信息的观点持有者所指的对象,它们往往是真实世界中的物体、事件或抽象的实体[2]。自动识别评论的观点主题,能更加直观地展示网民评论的焦点,并能有效提高多主题事件中观点分析的准确率。

由于观点分析不再局限于产品领域,观点主题也随着评论数据变化而改变,如网络舆情分析中不同事件对应不同的观点主题。因此,人工构建主题本体库的方法不再有效,需要对观点主题进行自动识别。然而针对单条评论的观点主题识别仍存在困难。另外,从用户的角度来看,他们更希望知道评论者讨论的焦点,对于部分与主旨不相关的评论并不关注。本文从整体主题考虑,以网络评论中观点主题产生的方式为依据,提出一种由内到外的观点主题识别方法,并对其有效性进行验证。

本文的布局如下:第 2 节介绍网络评论的观点主题识别的相关研究,第 3 节对相关知识进行描述,第 4 节着重介绍观点主题识别的思路和算法流程,第 5 节通过实验验证算法的合理性并对算法性能进行评估,最后对全文进行总结。

2 相关研究

目前,观点主题识别的研究多集中在英文文本领域,特别是在产品评论的观点主题(即产品特征)识别方面[3~5]。总体来说,相比于观点的持有者、陈述和情感等要素的相关研究,观点主题识别仍处于起步阶段,这主要由于任务本身的困难度和缺乏合适的标注语料库的支持。

在产品评论领域,主题的定义限定为产品的特征[3~6],主要包括产品名称、产品元件以及相关属性。由于特定产品的特征相对固定,可以通过全自动或半自动的方式构建特征本体库[6],再通过查询匹配确定评论的主题。Hu 等[3]首先利用关联规则挖掘确定频繁项作为候选特征,然后进行紧密度和冗余度修剪。Popescu 等[4]计算抽取的名词短语与特定鉴别短语之间的 PMI (Point-wise Mutual Information)值,并利用 WordNet 的 IS-A 层次关系和词形态线索来区分"部件(part-of)"或是"属性(property-of)"类别。

Stoyanov[7]指出观点主题标注需要联系上下文,并在 MPQA 语料的基础上,为每个观点句指定主题类别,再进行主题标注。依据观点主题的上下文相关特性,Stoyanov 在文献[2]中将观点主题识别转换为主题共指消解问题。Kim 和 Hovy[8]主要识别新闻数据的观点主题,他们首先确定观点词,并判断观点词对应的 FrameNet 的语义结构,将语义结构中的角色与观点的元素对应来识别观点主题。

在英文的相关研究中,往往抽取名词和名词短语作为候选主题,并借助外部资源(如 FrameNet、Charniak Parser 等)进行结构分析。汉语中观点主题的成分更加复杂,外部资源也相对缺乏,因此需要结合自身特点建立观点主题识别方法。

3 相关知识描述

一般情况下,主观性评论的发表都由源事件引起,如网络新闻评论中的新闻、论坛中的首帖、产品评论中的产品信息或产品属性列表等。这里将它们统称为评论源。网络评论的观点主题往往来源于评论源,但又不局限于评论源。根据该原则,将评论源中的观点主题称为内部主题,不属于评论源而存在于评论中的观点主题称为外部主题,它们之间的关系如图 1 所示。

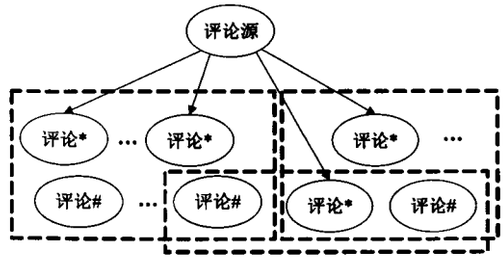

图 1 评论源和评论的观点主题关系图

图中,存在内部主题的评论表示为"评论 * ",全部为外部主题的评论表示为"评论",虚线框内的评论至少存在一个相同的主题。

设 s 表示评论源,由一组有序的词语序列构成,记为 $s = < w_{s1}, w_{s2}, \cdots, w_{sk} >$,其中 w 表示词语(包

含标点符号)。评论集 $R = \{r_1, r_2, \cdots, r_n\}$，$n$ 为评论总数，其中评论 $r_i = < w_{i1}, w_{i2}, \cdots, w_{il} >$。

定义1 内部主题：如果候选主题 $t = \langle w_1', w_2', \cdots, w_p' \rangle$ 满足：①词语序列 s 包含 t，即存在 $i(1 \leqslant i \leqslant k-p)$，使得 $w_j' = w_{s(i+j)}(j = 1, \cdots, p)$；② t 中存在词语 w_j' 符合关联规则约束，且词语权重值 $W(w_j') > \theta_w$，其中 w_j' 称为内部主题词；③主题权重值 $W(t) > \theta_{threshold}$，且 t 满足设定的内部主题规则。

定义2 外部主题：如果候选主题 $t = < w_1', w_2', \cdots, w_p' >$ 满足：①词语序列 s 不包含 t，即不存在 $i(1 \leqslant i \leqslant k-p)$，使得 $w_j' = w_{s(i+j)}(j = 1, \cdots, p)$；② t 中存在词语 w_j' 符合与线索词的关联规则约束；③主题权重值 $W'(t) > \theta'_{threshold}$，且 t 满足设定的外部主题规则。

任务：识别网络评论整体的观点主题 $T = \{t_i, i = 1,2,\cdots,m\}$。

本文主要利用关联规则挖掘算法[9]，具体表述如下：

设有项目集 $I = \{i_1, i_2, \cdots, i_u\}$，事务集 $D = \{T_1, T_2, \cdots, T_v\}$，其中每个事务 T 都是一些项目的集合，即 $T \subseteq I$。如果存在 $X \subset T$，则称事务 T 支持 X。关联规则是形如 $X \Rightarrow Y$ 的蕴含式，其中 $X \in I$，$Y \in I$，并且 $X \cap Y = \varphi$。如果 D 有 $a\%$ 的事务同时支持 X 和 Y，$a\%$ 成为关联规则 $X \Rightarrow Y$ 的支持度；如果 D 中支持 X 的事务中，有 $b\%$ 的事务同时也支持 Y，$b\%$ 成为关联规则 $X \Rightarrow Y$ 的置信度。支持度是对关联规则重要性的衡量，置信度是对关联规则准确度的衡量。

关联规则挖掘就是从事务集 D 中找出超过指定的最小支持度 $a\%$ 和最小置信度 $b\%$ 的关联规则。一般来说，关联规则挖掘包括两步过程：①发现所有的频繁项集，即项集的频度不小于 $a\%$；②由频繁项集产生强关联规则，即规则必须满足最小支持度和最小置信度。

4 观点主题识别

依据网络评论的观点主题往往来源于评论源，并且评论之间存在相同的主题(图1)，本文设计一种由内而外的观点主题识别方法。总体上，将网络评论的观点主题识别算法分为四个部分：内部主题词识别，确定评论源中受到评论者关注的主题词；内部主题构建，将零散的主题词合并为观点主题；外部主题识别，确定由评论源引申的观点主题；主题的组织，对结果进行简单的组织分析。

4.1 内部主题词识别

网络评论的观点主题来源于评论源，但并非评论源中每一个对象都是评论者关注的主题。观点主题也不再仅限于名词和名词短语，还可以是动词、动词短语等。同时未登录词和简称的影响，都会导致观点主题无法作为整体切分。例如，“央视在中国全面停播NBA”事件中，“央/视，中国，停/播，NBA，停/播/NBA”都作为观点主题。因此，需要首先识别出评论者关注的主题词。

本文通过以下步骤识别内部主题词：

(1)预处理

首先进行分词和词性标注，得到词语序列 $s = < w_{s1}, w_{s2}, \cdots, w_{sk} >$。再去除停用词、标点符号和语气助词等特定词性的词语。合并相同词语后得到 $s' = < w_{c1}, w_{c2}, \cdots, w_{cr} >$，$(w_{cj} \in s, j = 1, \cdots, r)$ 且 $w_{ci} \neq w_{cj}, (i, j = 1,2,\cdots,r; i \neq j)$。

(2)权重计算

分别确定各个词语 $w_{cj}(j = 1, \cdots, r)$ 反映观点主题的能力，用词语的权重值 $W(w_{cj})(j = 1,2,\cdots, r)$ 评估。其中，对观点主题判断产生影响的因素主要包括：

1) 整体词频。即词语 $w_{cj}(j = 1, \cdots, r)$ 在集合 $\{s, R\}$ 中出现的频率，它反映评论者关注的程度，用 $Freq(w_{cj})$ 表示，$Freq(w_{cj}) \in \{1,2,\cdots\}$。

2) 位置信息。指词语 $w_{cj}(j = 1, \cdots, r)$ 出现在评论源的位置信息(如标题、正文起始等)，表明评论源的发表者自身对词语重要性的评估，表示为 $Loc(w_{cj})$，$Loc(w_{cj}) \in \{1, \frac{3}{2}, 2\}$。

3) 词性信息。不同的词性对反映主题的能力不同，本文认为名词和动词具有最好的表征能力，用 $Pos(w_{cj})$ 表示，$1 \leqslant Pos(w_{cj}) \leqslant 2$。

4) 词语长度。单字词和多字词在表现具体含义的能力上存在差异，用 $Len(w_{cj})$ 表示长度差异影响，$1 \leqslant Len(w_{cj}) \leqslant 2$。

综合以上因素，权重值表示为

$$W(w_{cj}) = Loc(w_{cj}) \cdot Pos(w_{cj}) \cdot Len(w_{cj}) \cdot Freq(w_{cj}), (j = 1,2,\cdots,r)$$

利用均值动态设定阈值 $\theta_w = \alpha \frac{\sum_{j=1}^{r} W(w_{cj})}{r}$，其

中 α 取实验最优值$\frac{2}{3}$。当 $W(w_{cj}) > \theta_w$ 时，确定为内部主题词。

4.2 内部主题构建

如上文提到的，内部主题识别需要克服两个方面的问题：第一，未登录词和简称形式；第二，主题的词语组合形式。它们都具有位置上的连续特性，因此可以利用位置相关的关联规则算法构建。

经典的关联规则算法中，事务的项目集之间无位置约束。这里为 Apriori 算法添加位置约束条件，即要求频繁 k 项集只与它在事务中相邻的上一项和下一项来产生两个不同的 $k+1$ 项候选集。为了提高 Apriori 算法的效率，以内部主题词为线索词构建频繁项集。具体步骤如下：

1)扫描评论源 $s = < w_{s1}, w_{s2}, \cdots, w_{sk} >$，以线索词 w_i 为中心获取处理窗口。根据网络评论简洁的特性，仅需选择中心左右各两个词语组成窗口，即 $[w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}]$。选择包含 w_i 的 2-频繁项和 3-频繁项。

2)将集合 $\{s, R\}$ 作为事务集 D，利用 Apriori 算法挖掘满足最小支持度 $a\%$ 的频繁项集。由频繁项集产生关联规则，这些规则需要满足最小置信度 $b\%$。

在识别位置相关的频繁项后，还需要进一步优化结果。设 Apriori 算法挖掘到最高为 n-频繁项，① 对于 k-频繁项 X（$k < n$，且 $k > 1$），如果 $f(X) - \sum f(Y)$ 仍满足最小支持度 $a\%$，则认为 X 可作为内部主题。其中 Y 表示包含 X 的 p - 频繁项（$k < p \leqslant n$），$f(\cdot)$ 表示出现频率(不重复计算)。② 对于 1 - 频繁项 X，如果 X 不为单字符，根据上式判断。③ 过滤候选主题中除名词以外的单字符主题。

这里，最高为 3 - 频繁项（$n = 3$），优化过程对高频繁项中频繁出现的子集进行判断，确定是否能够独立作为主题。例如，3 - 频繁项“范/跑/跑”和“停/播/NBA”，两个主题的子集及其频率如表 1 所示。2 - 频繁项中“停播”和“跑跑”的 $f(X) - \sum f(Y)$ 仍满足最小支持度 $a\%$ 的要求，1 - 频繁项增加单字符和词性限制条件，“范”和“NBA”符合要求。因此，添加子集“停播”、“跑跑”、“范”和“NBA”为内部主题。

表 1 频繁项及其子集信息表

3 - 频繁项	子集(2 - 频繁项)	子集(1 - 频繁项)
范跑跑(345)	范跑(346) 跑跑(485)	范(834) 跑(1264)
停播 NBA(106)	停播(373) 播 NBA(132)	停(701) 播(639) NBA(826)

4.3 外部主题识别

网络评论的观点主题不局限于评论源，人们由评论源引申出相关的主题进行评论，但又与评论源存在内在的联系。本节对受关注程度高的外部主题进行识别，主要步骤如下：

1)合并集合 $\{s, R\}$ 中的分词结果，使识别的内部主题构成连续字符串。由内部主题和少量高频词作为线索词 $c_i(i = 1,2,\cdots,q)$ 进行关联规则挖掘，这里不对位置关系进行限制。

2)对线索词 c_i 关联得到的词语 w 进行判决，如果 $c_i \in c_j$ 且 $w \in c_j(j = 1,2,\cdots,q)$，则删除词语 w。并利用位置相关的关联规则挖掘识别未登录词和词语组合，组成候选外部主题集合。

3)计算各个候选项的权重。综合考虑词频、词性、词语长度和相关联的内部主题数，本文侧重选择名词词性、未登录词和词语组合，并设定阈值进行筛选。对于词语组合，过滤不符合主题要求的中文词性组合，如“V + A(动词 + 形容词)”、“N + V(名词 + 动词)”等。

4.4 主题的组织

在识别出网络评论的观点主题之后，确定主题之间的关系将有利于后续针对主题的观点分析。本文仅提供一种简单的识别方法，确定观点主题中动词作用的对象信息。例如，“取消范跑跑教师资格”事件中，存在评论“坚决赞成解聘!”，识别动词主题“解聘”作用的对象“范跑跑”，将有助于观点分析。

设观点主题集合 $T = \{t_i, i = 1,2,\cdots,m\}$，如果主题 t_i 为一个动词，并与名词(短语)或命名实体主题 $t_j(j \neq i)$ 顺序组成字符串 $s = t_i t_j \in T$，则认为动词 t_i 可以作用于对象 t_j。当评论中动词 t_i 对象缺省时，可以使用 t_j 进行补充。

5 实 验

实验所用数据来源于强国论坛、网易新闻评论

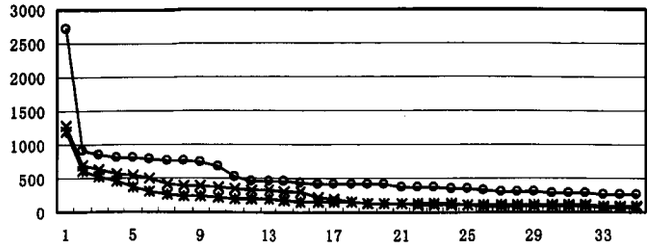

图2 评论数据中的词频信息

等网站,共选取17个话题的网络评论数据,总计一万六千余条。本文采用中科院分词系统ICTCLAS进行词语切分和词性标注,完成预处理过程后统计词频信息。词语词频变化具有明显的趋势。如图2为三个话题的词频信息,横坐标表示词频由高到低的词语序列,其中部分词语的频率远远高于其他词语。

分析各个话题中频率最高的10个词语。评论源中包含的词语平均占54.7%,其中包含大量的观点主题词,表明一般情况下网络评论确实是围绕评论源展开的。另一方面,非评论源的高频词中含有较多常用词汇,如"会"、"有"等。本文使用有效率为指标对识别的观点主题进行人工评价,并从三个方面进行判断:①识别主题是否正确切分和组合;②是否与话题相关联;③在网络评论中是否作为评论的主题。如果满足,则认为识别出有效主题。有效率计算公式为:

有效率 = 有效主题数/识别主题总数 × 100%

对所有识别主题进行判断,得到平均有效率为88.7%。分析识别结果,错误识别主要是一些常用的高频动词,如"应该"、"反对"。当评论的主题相对集中时,识别有效率能够达到95%以上。为了直观地展示识别结果,表2列出"央视在中国全面停播NBA"事件中识别的前10个观点主题。

表2 事件中识别的前10个观点主题

	Word	Source	Freq
1	NBA	评论源	826
2	中国	评论源	562
3	停 播	评论源	373
4	国家	评论	350
5	政治	评论	324
6	体育	评论源	289
7	CCTV	评论	221
8	球迷	评论	146
9	篮球	评论源	130
10	停 播 NBA	评论源	106

为了测试评论中针对已识别主题发表评论的数量,即已识别主题在评论数据中的覆盖范围,采用主题覆盖率进行评价。对于一条评论,如果其主题存在于已识别主题中,则认为在覆盖范围内。覆盖率计算公式如下:

覆盖率 = 处于覆盖范围内的评论数/总评论数 × 100%

同时,为了测试评论数量的影响,将话题分为三类:类型1,评论数量大于1000条,有5个话题;类型2,评论数量介于500和1000之间,有6个话题;类型3,评论数量小于500条,有6个话题。从每个话题中各抽取100条评论进行判断,得到结果如表3所示。

表3 三种类型中已识别主题的覆盖率

类型	平均覆盖率	省略主题评论	去除后的覆盖率
类型1	72.0%	4.0%	75.0%
类型2	73.2%	4.0%	76.2%
类型3	60.8%	5.7%	64.5%

网络评论数据中,部分评论仅发表观点而省略观点主题信息,一般情况下可将主题认为是评论源中的最频繁主题。当去除省略主题的评论后,重新计算覆盖率得到结果如表3所示。实验结果表明,识别的观点主题能够表示评论的整体主题信息。当评论数量较少时,主题信息的频率信息不够显著,致使观点主题识别性能降低。另外,从话题自身来讲,这类话题一般不是矛盾的聚焦点,评论的观点主题体现不明显;当评论到达一定数量后,主题的覆盖率与话题自身有关。

分析未覆盖的评论数据,主要包含三种类型:①与话题无关的评论;②非频繁主题,主要由暗指和同义词引起;③使用指示词,如"这种行为"、"这个问题"和"他(她)"等。同时,本文识别的观点主题也可

以作为判断话题无关评论的依据。

6 结束语

近年来,互联网中网民的意见受到越来越多的关注,领域无关的观点分析能够自动分析网民对网络中新事件的整体意见,从而实现及时预警和合理疏导。观点主题识别是领域无关的观点分析的重要组成部分,识别出网民观点所指的对象。本文对细粒度观点分析中的观点主题识别方法进行初步研究,根据评论产生的基本过程构造整体主题识别模型。对网络话题的评论数据进行实验,算法取得较好的识别效果,但仍有许多需要完善的地方。在下一步的工作中,我们将加入句子结构信息用以提高观点主题的识别性能,并设计出一种有效的无关评论判别方法,过滤无关信息,从而提高识别覆盖率。

参 考 文 献

[1] Kim S M, Hovy E. Determining the sentiment of opinions [C]// Proceedings of the 20th International Conference on Computational Linguistics (COLING-04). Switzerland, 2004: 1367-1373.

[2] Stoyanov V, Cardie C. Topic identification for fine-grained opinion analysis[C]// Proceedings of the 22th International Conference on Computational Linguistics (COLING-08). Manchester, 2008: 817-824.

[3] Hu M Q, Liu B. Mining opinion features in customer reviews [C]// Proceedings of the 19th National Conference on Artificial Intelligence (AAAI-04). California, 2004: 755-760.

[4] Popescu A M, Etzioni O. Extracting product features and opinions from reviews [C]// Proceedings of the Joint Conference of Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP-05). Vancouver, 2005: 339-346.

[5] Liu B, Hu M Q, Cheng J S. Opinion observer: analyzing and comparing opinions on the Web[C]// Proceedings of the 20th International World Wide Web Conference (WWW-05). Chiba, 2005: 342-351.

[6] 姚天昉, 聂青阳, 李建超. 一个用于汉语汽车评论的意见挖掘系统[C]// 中国中文信息学会二十周年学术会议, 2006: 260-281.

[7] Stoyanov V, Cardie C. Annotating topics of opinions[C]// Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-08). Marrakech, 2008: 3213-3217.

[8] Kim S M, Hovy E. Extracting opinions, opinion holders, and topics expressed in online news media text[C]// Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (ACL/COLING-06). Sydney, 2006: 1-8.

[9] Han J W, Kamber M. Data Mining Concepts and Techniques [M]. Beijing: Higher Education Press, 2001: 234-250.

(责任编辑 许增棋)