基于领域特征指示词的隐式特征识别研究

陈 莹[1,2],叶 宁[1,2],徐 康[1,2],王汝传[1,2]
(1. 南京邮电大学 计算机学院、软件学院、网络空间安全学院,江苏 南京 210046;
2. 江苏省无线传感网高技术研究重点实验室,江苏 南京 210046)

摘 要:网络购物这一领域的迅猛发展带来了海量的在线评论数据,挖掘评论数据中所蕴藏的语义以及情感信息对用户以及商家都有着莫大的价值。在这样的应用需求背景下,出现了针对文本的情感分析(sentiment analysis)技术。但由于中文语言表达的多样性与复杂性,用户会在评论中含蓄地提到评价属性与观点。而现有研究对包含显式特征评论文本的情感分析已趋渐成熟,针对隐式评论句进行特征识别的却较少。因此,文中面向隐式特征识别这一研究难点,提出一种基于领域特征指示词的隐式特征识别方法。该方法首先利用构建的多词型的主题情感联合模型对特定领域内的显式评论句进行特征类别指示词的挖掘;再引入词向量模型作为衡量隐式评论句中线索词与特征指示词集中词项语义相关度的标准;最后分情形来实现对隐式评论句中线索词所属特征类别的指派。通过对不同产品的评论数据集进行实验,结果证明了该方法的有效性。

关键词:产品评论;语义分析;显式特征;隐式特征;主题模型;词向量
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2021)09-0024-07
doi:10.3969/j.issn.1673-629X.2021.09.005

Research on Implicit Feature Identification Based on Domain Feature Indicators

CHEN Ying[1,2], YE Ning[1,2], XU Kang[1,2], WANG Ru-chuan[1,2]
(1. School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210046, China;
2. Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing 210046, China)

Abstract: The rapid development of online shopping has brought a huge amount of online review data. The semantic and emotional information contained in the review data is of great value to both users and merchants. In this context of application requirements, sentiment analysis for text has emerged. Due to the diversity and complexity of Chinese language expression, users will implicitly mention evaluation attributes and opinions in comments. The methods of mining comments with display feature have become more and more mature, but the research on implicit feature identification is less. Therefore, an implicit feature identification method based on domain feature indicators is proposed for implicit feature identification. Firstly, the constructed multi-word thematic affective association model is used to mine the feature category indicators of the display comments in a specific field. The word2vec is used as a criterion to measure the semantic relevance between the clue word and the feature indicator word in implicit comments. Finally, the assignment of the characteristic category of the clue words in the implicit comment is realized by case analysis. The effectiveness of the proposed method is demonstrated by experiments on review data sets of different products.

Key words: product-reviews; semantic analysis; explicit feature; implicit feature; topic-model; Word2Vec

0 引 言

依托互联网+逐层推进的时代背景,足不出户的网购以不可逾越的地位占据人们的内心世界,参与网购的人数与涉及的商品服务种类与日俱增。可见,网购俨然成为一种时尚的潮流[1]。由此衍生出的在线评论数据呈爆发式增长并且蕴含着巨大的潜在价值[2]。对这些数据的有效挖掘可以帮助公司和商家深入了解消费者需求,从而提高产品的质量。但是,仅从文档层面或句子层面分析还不足以探究用户的意见。人们通常对产品的各个方面进行评论,包括产品的组成部分

收稿日期:2020-10-11 **修回日期**:2021-02-18
基金项目:国家自然科学基金(61872194,61872196)
作者简介:陈 莹(1995-),女,硕士研究生,研究方向为自然语言处理;叶 宁,教授,CCF 高级会员(10059S),通信作者,研究方向为无线传感器网络信息与安全。

以及属性。因此,有必要对产品特征级别的观点进行提炼,而不是针对整个产品或整个评论文本。

然而,分析客户评论以获得更细粒度理解的自动化任务面临着许多挑战,特别是由于在客户评论中并不总是显式地提出。已有的技术和研究大都只是致力于从评论语料中挖掘和抽取在评论语句中显式出现了的评价对象[3]。而根据 Kim 和 Flavius 的研究可知,产品中许多重要的特性也会被消费者含蓄地提到[4]。例如评论句“手机很小,可以放进我的口袋里”隐式地表达了关于手机“尺寸”方面的意见。隐式特征的提取是一个复杂的问题。文中主要研究隐式方面识别。

考虑如下关于电子产品领域的评论:

例子 1 “很棒,很顺畅不卡顿。”

例子 2 “还不错,可以随身携带。”

例子 3 “昨天下单,今天收到货了。”

这些评论句不难发现都有一个共通之处即不包含明确的特征词。但在例句 1 中,根据观点词“顺畅、卡顿”可轻易推断出用户是在描述系统这一特征。例句 2 中有观点词“不错”,但由于其适配性很难仅从词语本身识别出特征,结合下文中提到的“随身携带”可知用户是想表达关于“尺寸”这一特征的观点。例句 3 中没有任何评价词,但根据“收到货”这一非观点词可知是在描述“物流”这一特征。所以,根据上述分析可知,借助评论句中的观点词或非观点词可间接地识别出隐式特征。

在现有的研究中,隐式特征识别大致采用共现分析、关联规则、主题模型及分类等方法,其中基于共现和关联规则的关系推断法最普遍[5]。这两种研究方法主要是依赖观点词与特征属性之间的映射关系,利用带有标签的语料库训练模型来提取隐式特征。但随着线上交易量的日益剧增,在线评论的数据量也越来越多,需要消耗大量人力资源。研究者开始专注于无监督方法。主题模型,如 PLSA 和 LDA,在自然语言处理的许多任务中很流行,它们也可以用来识别隐式特征[6-7]。

但这些方法没有考虑在没有观点词的情况下非观点词对识别隐式特征的指导性,而且有的方法也忽视了词的语义信息,使得隐式识别的精度和准确度不是很高。所以,文中面向隐式特征识别这一研究难点,提出了一种基于领域特征指示词的隐式特征识别方法。该方法首先利用多词型的主题情感联合模型自动地从包含显式特征的评论句中挖掘出“特征-情感”和“特征-非观点”词对集,整合成特征指示词集;再引入词向量模型作为衡量隐式评论句中线索词与特征指示词集中词项语义相关度的标准;最后根据线索词的类型对隐式特征分情况进行识别。

1　相关工作

隐式特征首先在 Liu 等人中进行了讨论[5],他们给出了隐式特征的定义。从那时起,一些研究开始关注隐式特征的识别。目前的研究可分为监督识别、无监督识别和半监督识别三类。文中主要基于无监督识别展开研究。

Prasojo 等人扩展了传统的命名实体识别方法,利用形容词到方面的映射将特征集关联到每个实体[8],然后,他们选择频率最高的特征作为目标。Santu 等人结合一个背景语言模型和几个特征语言模型生成评论中的每个单词。他们通过期望最大化(EM)估计参数,并检测最终的隐含特征列表[9]。Xu 等[6]预先定义特征类别,将在包含显式特征的评论句中得到的约束和先验知识纳入主题模型 LDA 得到特征类别的相关词语,以这些词语为特征对评论句建模,通过构建 SVM 分类器识别隐式特征。Sun 等[7]使用联合主题模型进行隐式特征提取。他们将与隐含特征相关的意见词分为两类,即特殊意见词和一般意见词。一般意见词可以与许多不同的特征共同出现,而特殊意见词只与一个特定的特征共同出现。他们计算了两个概率分布,一个是主题的意见分布,另一个是主题和意见的上下文分布。最后,他们使用这些值进行隐式特征提取。张莉等基于领域中的常用词对特征词进行聚类,通过精简意见词和对其进行同义词扩展,构建<特征观点 权重> 三元组字典,用于识别隐式特征[10]。

此外,还有许多其他方法,如关联规则挖掘(Zhang 等)[11]和共现关系(Rana and Cheah[12]; Makadia[13])用于无监督隐式特征识别。

2　方　法

2.1　总体流程

文中所提出的方法具体如图 1 所示,主要包括三个步骤。首先,利用多词型的主题情感联合模型进行特征主题聚类并从显式评论句中提取出“特征-特征指示”词对集;接着,使用语言技术平台 LTP 对隐式评论句进行词性标注,产生候选线索词,利用词向量模型计算线索词与特征指示词的语义相似度为线索词匹配特征指示词;最后,根据所匹配到的特征指示词类型分情况采用不同的方法进行隐式特征的指派。

2.2　多词型的主题情感联合模型

2.2.1　模型概述

ASUM(aspect and sentiment unification model)模型基于 LDA(latent Dirichlet allocation)进行改进,假设每个句子只有一个主题以及这个主题下的情感倾向。因此,模型的主要目的便是从评论文本中提取出每一个句子中的(特征,情感)对,以此作为情感分析的依

据[3]。但是 ASUM 并未区分表示主题的词语是特征词,或特征指示词还是情感词,要想明确得到词语的类型,还需要人为地进行辨别。因此,为了能从显式评论句中自动挖掘出基于领域的“特征–情感”和“特征–非观点”词对集并充分利用主题模型的主题(特征)聚类性质,文中基于 ASUM 模型的假设提出一个多词型的主题情感联合模型。该模型通过加入表示单词类型的隐含变量,建立其与单词的关系,进一步获得类型同单词的概率分布,从而可以识别出单词的类型。

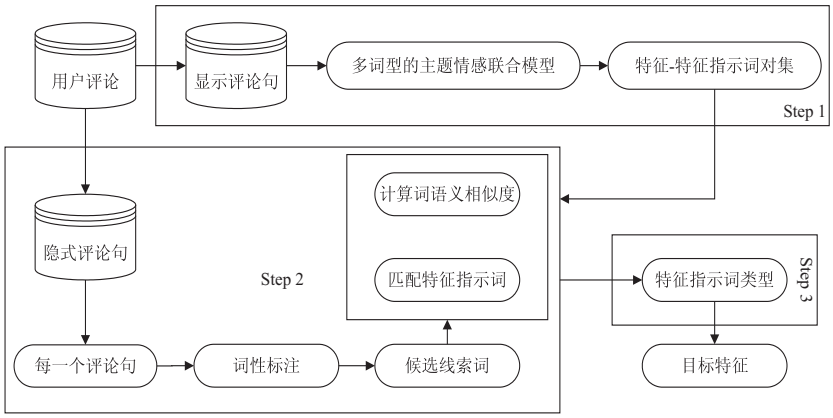

图 1 基于领域特征指示词的隐式特征识别方法框架

多词型的主题情感联合模型的图形化表示如图 2 所示,相关的变量和符号在表 1 中给出解释。

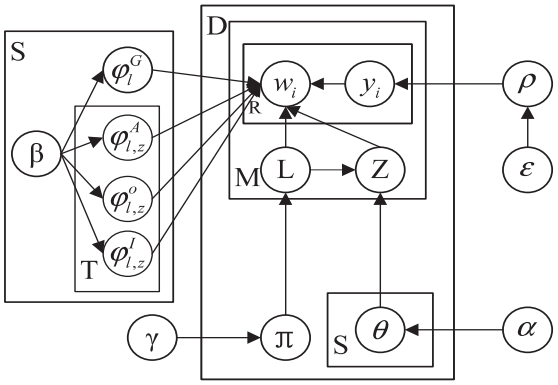

图 2 多词型的主题情感联合模型图形化表示

多词型的主题情感联合模型通过引入一个隐含变量 y 来表示单词的类型。$y \in \{0,1,2,3\}$ 分别表示单词 w 是一个通用情感词,特定的情感词,特征词以及非观点特征指示词。模型根据一个先验的狄利克雷分布生成词语的类型分布,狄利克雷分布是多项式分布的共轭分布,共轭的特性可以使得先验分布和后验分布的形式相同,可以形成一个先验链[8]。

大多数的产品评论其实都是一句话包含一个特征以及对其评价观点,所以为了挖掘针对同一实体产品的评论集中不同特征以及观点,此模型假设同一个句子的单词属于同一个主题(特征)和情感极性,则每一篇文档在此模型下的生成过程如下:

(1)生成一个词的类型分布 $\rho \sim \mathrm{Dir}(\varepsilon)$;

(2)生成一个情感分布 $\pi_d \sim \mathrm{Dir}(\gamma)$;

(3)对每一个情感倾向 l,生成一个主题分布 $\theta_{d,l} \sim \mathrm{Dir}(\alpha)$;

(4)对于每一个情感倾向 l,生成情感 l 下的通用情感词分布 $\varphi_l^G \sim \mathrm{Dir}(\beta_l)$;

表 1 多词型的主题情感联合模型图字母含义

		含义
符号	S	情感倾向的个数
	T	主题的个数
	M	句子的个数
	R	词的个数
	D	文档数目
超参数	γ	π 的狄利克雷分布先验
	ε	ρ 的狄利克雷分布先验
	α	θ 的狄利克雷分布先验
	β	φ 的狄利克雷分布先验
隐含和可观察的变量	w	词语
	l	情感标签
	z	主题
	y	词类型
	ρ	词类型分布
	π	情感倾向分布
	θ	主题分布
	φ_l^G	通用情感词分布
	$\varphi_{l,z}^A$	特征词分布
	$\varphi_{l,z}^o$	特定情感词分布
	$\varphi_{l,z}^I$	非观点特征指示词分布

(5)对于每一个情感倾向 l 和主题 z,生成三种类型的词语分布:

(a)情感 l 和主题 z 下的特征词分布 $\varphi_{l,z}^A \sim \mathrm{Dir}(\beta_l)$;

(b)情感 l 和主题 z 下的特定情感词分布 $\varphi_{l,z}^o \sim$

$\mathrm{Dir}(\beta_l)$;

(c)情感 l 和主题 z 下的非观点特征指示词分布 $\varphi^{l}_{l,z} \sim \mathrm{Dir}(\beta_l)$ 。

(6)对于文档中每一个句子:

(a)选择一个情感标签 $l \sim \mathrm{Multi}(\pi_d)$;

(b)选择一个主题 $z \sim \mathrm{Multi}(\theta_{d,l})$ 。

(7)对于每一个单词 $w_i \in d$:

(a)将它所属文档的情感标签 l 分配给它;

(b)选择一个主题 $z_i \sim \mathrm{Multi}(\theta_{d,l})$;

(c)选择单词的类型 $y_i \sim \mathrm{Multi}(\rho)$;

(d)选择单词 w_i :

选择一个单词 $w_i \sim \mathrm{Multi}(\varphi^{y_i}_{l,z_i})$ 或者 $w_i \sim \mathrm{Multi}(\varphi^{y_i}_{l})$ 。

2.2.2　参数估计

多词型的主题情感联合模型的参数估计使用了吉布斯采样。在采样初始化过程中,引入情感词典、领域情感词典以及领域特征词典作为先验知识,以便能更准确地采样出词语的类型。具体做法就是,在初始化时遍历所有文档中每一个单词,若单词存在于这三个词典里,便对其标注相应的词语类型。

为了获得 π , θ , φ 和 ρ ,在吉布斯的采样过程中会依次采样出每一个单词的主题,情感倾向以及单词的类型。现在大多数产品评论都是内容精短但语义信息丰富的形式,若单纯将每一个评论看作是一篇文档,会因为文本的稀疏性造成采样结果准确率不高的情况。而文中是为了挖掘某一实体产品的不同特征,其评论都是围绕产品不同特征进行评价,评论句之间都有一定的语义相似度。所以为了解决评论文本稀疏性问题,文中在多词型的主题情感联合模型的采样过程中,将所有评论看作是一篇长的伪文档进行采样。首先,为每一个单词采样一个主题和情感标签,主题和情感标签是联合采样,采样条件公式如公式(1)所示,公式中具体符号含义在表 2 中给出解释。

$$p(s_i = j, z_i = k \mid s_{-i}, z_{-i}, w) \propto \frac{C^{\mathrm{DS}}_{dj} + \gamma_j}{\sum_{j=1}^{S} C^{\mathrm{DS}}_{d_j} + \gamma_j} \frac{C^{\mathrm{DST}}_{djk} + \alpha_k}{\sum_{k=1}^{T} C^{\mathrm{DST}}_{djk'} + \alpha_{k'}} \frac{\Gamma(\sum_{w=1}^{W} C^{\mathrm{STW}}_{jkw} + \beta_{jw})}{\Gamma(\sum_{w=1}^{W} C^{\mathrm{STW}}_{jkw} + \beta_{jw} + m_i)} \prod_{w=1}^{W} \frac{\Gamma(C^{\mathrm{STW}}_{jkw} + \beta_{jw} + m_{iw})}{\Gamma(C^{\mathrm{STW}}_{jkw} + \beta_{jw})} \tag{1}$$

接着对词语类型进行采样。基于狄利克雷的先验分布,第 i 个单词的词语类型 y 的采样条件公式如公式(2)所示,公式中具体符号含义也在表 2 中给出解释。

$$p(y_i = t \mid y_{-i}, l, z, w) \propto \begin{cases} \dfrac{(n^{(c)}_{l,t,-i} + \beta_i)(n^{(t)}_{l,d,-i} + \varepsilon_t)}{\sum_{i=1}^{V}(n^{(c)}_{l,t,-i} + \beta_i)} & t = 0 \\ \dfrac{(n^{(c)}_{l,k,t,-i} + \beta_i)(n^{(t)}_{l,k,d,-i} + \varepsilon_t)}{\sum_{i=1}^{V}(n^{(c)}_{l,k,t,-i} + \beta_i)} & t \in \{1,2,3\} \end{cases} \tag{2}$$

表 2　模型参数估计的相关符号

符号	含义
s_i	句子 i 情感
z_i	句子 i 的特征
y_i	单词 i 的词类型
y_{-i}	除了单词 i ,分配给所有单词的类型
s_{-i}	除了句子 i ,分配给所有句子的情感
z_{-i}	除了句子 i ,分配给所有句子的特征
W	单词列表
C^{DS}_{dj}	文档 d 中分配了情感 j 的句子数
C^{DST}_{djk}	文档 d 中分配了情感 j 和特征 k 的句子数
C^{STW}_{jkw}	分配了情感 j 和特征 k 的单词数
$m_{i(w)}$	句子 i 中的所有单词数(或单词 w 的个数)
$n^{(c)}_{l,t,-i}$	在情感标签 l 下,除去位置 i 的单词后的词类型 t 对应的单词 c 的个数
$n^{(t)}_{l,d,-i}$	在情感标签 l 下,除去位置 i 的单词后,文档 d 中词类型 t 出现的个数
$n^{(c)}_{l,k,t,-i}$	在情感标签 l 和特征 k 下,除去位置 i 的单词后的词类型 t 对应的单词 c 的个数
$n^{(t)}_{l,k,d,-i}$	在情感标签 l 和特征 k 下,除去位置 i 的单词后,文档 d 中词类型 t 出现的个数

为了后续隐式特征识别的引用,将从显式评论中挖掘出的特征指示词对集整合成如下形式。每一个特征类别($F_1F_2\cdots F_m$)下对应一般情感词、特征情感词以及非观点情感词三种类型词语,每一种词类型下保留概率 top 20 的词语,对其进行筛选,留下语义相关性强的词语,如表 3 所示。

表 3　特征-词型-指示词

特征类别	词型	指示词
	G (一般情感词)	$w_1, w_2\cdots$
F_1	O (特征评价词)	$w_1, w_2\cdots$
	I (非观点情感词)	$w_1, w_2\cdots$
	G	$w_1, w_2\cdots$
F_2	O	$w_1, w_2\cdots$
	I	$w_1, w_2\cdots$
…	…	……
	G	$w_1, w_2\cdots$
F_m	O	$w_1, w_2\cdots$
	I	$w_1, w_2\cdots$

2.3 隐式特征识别

文中基于特征指示词集识别隐式特征,关键步骤就是为隐式评论句中的线索词寻找到最匹配的指示词。利用多词型的主题情感联合模型所挖掘出的特征指示词集虽然在主题聚类以及自动化方面比较好,但会因为基于词共现的原理而忽视一些低频但语义相关度很高的词语,使得匹配指示词的结果不是很成功。所以,为了能在特征指示词集中成功匹配到与线索词相关联度最高的指示词,引入了词向量模型。

词向量概念 Word2Vec 的核心思想是通过上下文学习词的向量表示。词向量的表示能够反映词的语义信息并且利用其空间距离可测度词项间的语义关联度。词向量有 CBOW(continuous bag of words)和 Skip-gram 两个重要模型,二者主要的区别在于 CBOW 利用上下文预测词项,Skip-gram 则是根据词项预测上下文。文中选择 CBOW 模型,借助 Python 的 Genism 工具包构建词向量,向量维度 100,上下文窗口尺寸 5[10]。

将词语在 R^n 空间的词向量表示为 $\vec{w}=[x_1,x_2,\cdots,x_n]$。用词向量内积度量词语间的语义相关度,如公式(3)所示。

$$\mathrm{Sim}(w_1,w_2)=\cos\theta=\frac{\overrightarrow{w_1}\cdot\overrightarrow{w_2}}{\|\overrightarrow{w_1}\|\cdot\|\overrightarrow{w_2}\|} \tag{3}$$

隐式评论句中的线索词一般为观点词和非观点词两种。文中利用语言技术平台 LTP 对评论句进行词性标注,保留下形容词、名词或名词性短语以及动词或动词性短语作为候选线索词和上下文词。为了提高隐式特征识别的准确率,依据隐式评论句中线索词的类别对隐式特征分情况进行识别。

具体步骤如下:

Step 1:选择线索词。若评论句中有形容词,则将形容词视为线索词。否则,将动词或动名词视为线索词。

Step 2:匹配特征指示词。若线索词是形容词,利用公式(3)计算其与表 3 中 *G* 和 *O* 两种类型下词语的关联度,选择关联度最高的词项作为其特征指示词。若线索词是动词或动名词,利用公式(3)计算其与表 3 中 *I* 类型下词语的关联度,也是选择关联度最高的词项作为其特征指示词。

Step 3:依据特征指示词的类型分情形识别隐式特征。

(1)特征指示词是特定情感词或非观点词,将其所属特征类别直接匹配给线索词。

(2)特征指示词是一般情感词,需要结合线索词的上下文词。选定线索词邻近的名词或动词作为上下文词,并根据公式:

$$\mathrm{Score}(F_i)=\alpha\times\delta_i^{\mathrm{Ind}}+b\times\tau_i^{\mathrm{Ind}} \tag{4}$$

计算候选特征类别的得分,选择候选特征集中得分最高的作为观点词识别的隐式特征。其中,δ_i^{Ind} 表示的是线索词的上下文词 con_j 与特征类别 F_i 的关联度,对已标注特征类别的显式评论句使用加权对数似然几率作为度量指标。

计算公式如下:

$$\delta_i^{\mathrm{Ind}}=\sum_{j=1}^{\mathrm{con_{num}}}p(\mathrm{con}_j\mid F_i)\log\frac{p(\mathrm{con}_j\mid F_i)}{p(\mathrm{con}_j\mid\bar{F}_i)}=\frac{N_{\mathrm{con}_j,F_i}}{N_{\mathrm{con}_j,F_i}+N_{\overline{\mathrm{con}_j},F_i}}\log\frac{N_{\mathrm{con}_j,F_i}(N_{\mathrm{con}_j,\bar{F}_i}+N_{\overline{\mathrm{con}_j},\bar{F}_i})}{N_{\mathrm{con}_j,\bar{F}_i}(N_{\mathrm{con}_j,F_i}+N_{\overline{\mathrm{con}_j},F_i})} \tag{5}$$

式中,N_{con_j,F_i} 表示上下文词 con_j 在特征类别是 F_i 的评论句中出现的频率;符号“-”表示非的含义;$\mathrm{con_{num}}$ 表示线索词的上下文词的数目。τ_i^{Ind} 表示匹配的特征指示词在其给定的情况下属于特征类别 F_i 的概率,根据多词型的主题情感联合的主题采样公式(1)和贝叶斯原理,可由公式(6)求得:

$$\tau_i^{\mathrm{Ind}}=\frac{\varphi^Z*\mathrm{word_{all}}}{T_{\mathrm{all}}} \tag{6}$$

其中,$\mathrm{word_{all}}$ 表示所有词语的数目,T_{all} 表示所有主题(特征)的个数。设定 $a+b=1$,表示 δ_i^{Ind} 和 τ_i^{Ind} 所占的权重,在进行多次实验后可知 a 的值为 0.32,b 的值为 0.68。

算法 1:描述了隐式特征识别的过程。

Algorithm 1:隐式特征识别

输入:线索词集 W^{cue},线索词的上下文词集,特征指示词集 W^{Ind}

输出:相匹配的隐式特征集

1 对 W^{cue} 里的每一个线索词 w^{cue}:

2 如果 w^{cue} 是形容词:

3 对特征指示词集 W^{Ind} 里一般情感词和特定情感词类型下的每一个特征指示词 w^{Ind}:

4 计算余弦相似度 $\mathrm{sim}(w^{\mathrm{cue}},w^{\mathrm{Ind}})$;

5 循环结束

6 否则

7 对特征指示词集 W^{Ind} 里非观点词类型下的每一个特征指示词 w^{Ind}:

8 计算余弦相似度 $\mathrm{sim}(w^{\mathrm{cue}},w^{\mathrm{Ind}})$;

9 循环结束

10 得到线索词语义相似度最大的特征指示词 w^{Ind},特征指示词的类型及其所属特征;

11 如果 w^{Ind} 的类型是特定观点词或非观点词:

12 预测 w^{Ind} 所属的特征为相对应线索词的目标特征;

13 否则

14 利用线索词的上下文词和 w^{Ind} 计算候选特征集的得分;

15 预测得分最高的候选特征为相对应线索词的目标特征

3　实验结果与分析

3.1　数据集选择

文中使用了五个不同产品的用户评论来评估所提出的方法,分别是酒店、手机、平板、计算机和衣服。每种产品的评论数量是 10 000 条。使用 Python 工具包 nltk 和语言技术平台 LTP 对评论进行分句、去除停用词、分词和词性标注等操作。经过筛选,各产品的隐式评论句大约占评论总数的 25% 左右,可见识别隐式特征具有重要的意义,能够更全面捕捉特征信息,进一步提升情感分析的精度。

3.2　参数设置

为了训练多词型情感主题情感联合模型,依据文献[14],将参数 γ 设置为 1,表示各种情感出现的概率相同。参数 β 为了结合种子词,采用非对称取法,负向单词情感采样的时候,正向单词的先验为 0,其他设为 0.001,同理正向采样时,负向单词先验为 0,其他也设为 0.001。参数 α 和参数 ε 则分别设置为 0.1 和 0.25,迭代 1 000 次。

3.3　评价指标

文中使用精准度 precision 以及召回率 recall 作为评价指标,如公式(7)和公式(8)所示。

$$\text{precision} = \frac{\text{正确识别的数目}}{\text{测试语料库总数目}} \tag{7}$$

$$\text{recall} = \frac{\text{正确识别的数目}}{\text{所有测试语料库中人工识别的数量}} \tag{8}$$

3.4　实验结果分析

3.4.1　特征指示词集

在进行隐式特征识别之前,首先需要建立一个“特征-特征指示”词对集。表 4 展示了一个关于酒店的显式评论句的挖掘结果样例。实验设定主题(特征)个数为 8,依据各主题下的特征词描述可知这 8 个特征分别为:地理位置、环境、服务态度、酒店设施、餐饮、价格、网上预订以及人气。可以发现地理位置、价格、服务态度和酒店设施这四个类别下不同类型的词语分布比较均匀,而环境、餐饮、网上预订以及人气这四个类别则是某一类型下的词语分布比较突出。由于主题模型依赖数据质量,使用的数据量不够,出现了一些无效词。

3.4.2　隐式特征识别

文中识别隐式特征很大程度上依赖于多词型的主题情感联合模型的采样结果和词向量模型的训练,而在线索词为一般情感词的情况下又考虑了上下文的权重。文献[7]基于标准 LDA 模型提出了一种改进的主题模型联合主题-意见模型(JTO),用于提取意见词的隐含特征,包括特殊意见词和一般意见词。文献[15]试图通过构建改进矩阵和实现 LDA 主题模型来得到两个概率分布。采用余弦相似度考虑上下文权重,计算意见词候选特征的得分来实现隐式特征识别。所以,文中选择与文献[7,15]中用到的方法进行比较。评估指标的计算依赖于手工注释。结果如图 3 和图 4 所示,其中 JTO 和 CW 分别表示文献[7]和文献[15]中所用的方法,MI 则表示文中方法。

表 4　特征指示词集

特征类别	词型	指示词
地理位置	G	不错 满意 喜欢 很棒 差 失望 极好
	O	明确 近 远
	I	靠着 临近 数百米 交通 离店 走路 依山 老街
环境	G	满意 喜欢 失望 惨不忍睹 极好
	O	舒服 幽雅 别具匠心 狭窄 干净 安静 吵 嘈杂 惬意 脏乱 鸟语花香 宽敞
	I	异味 装修 垃圾 视野
服务态度	G	满意 失望 差 不好
	O	热情 耐心 不近人情 友好
	I	接待 招呼 帮 办妥 代劳 投诉
酒店设施	G	一般般 不好 失望 满意
	O	齐全 破 完好 通透 陈旧 安全
	I	卫生间 电热水壶 隔音 停车场 门锁 淋浴 床垫 枕套 有窗 网速
餐饮	G	一般般 中规中矩 满意 很差
	O	美味
	I	味道 品尝 吃 水果 西餐 中餐 早餐
价格	G	满意 正常
	O	值得 便宜 贵 实惠 合适 超值
	I	钱 元 消费 房价 特价 特惠
网上预订	G	喜欢 糟心 失望
	O	方便
	I	携程 条款 免费送 预订 通知 反馈 代理 退款 订单 网管
人气	G	喜欢 满意 失望
	O	抢手 舍不得 旺
	I	排队 推荐 二星 四星 三星 星级 再次入住 分数

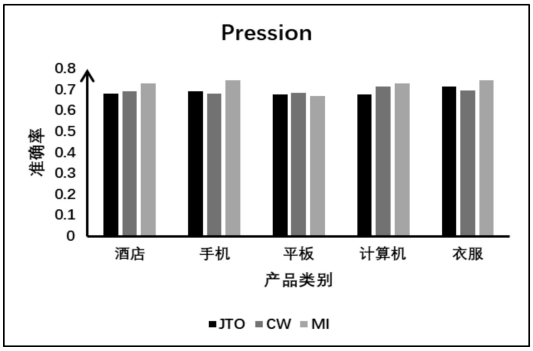

图 3　隐式特征识别的准确率

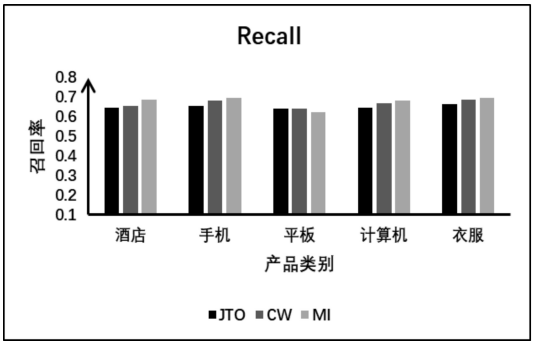

图 4 隐式特征识别的召回率

依据图中数据可以看到,在平板这产品数据集上,MI 方法的准确率和召回率比 JTO 和 CW 要低,有可能是因为多词型的主题情感联合模型在这一数据集上的表现不是很好。而在整体上,经过统计发现 MI 方法在精确率和召回率上比文献[7]中 JTO 方法平均高出 3%,比文献[15]中 CW 方法平均高出 2%,这可能是因为 JTO 和 CW 虽然考虑了观点词和上下文词的权重,但却忽视了词的语义信息和非观点词的指示性。综合上述分析,证明了文中提出的基于领域特征指示词的隐式特征识别方法的有效性。

4 结束语

文中提出了一种基于领域特征指示词的隐式特征识别方法。该方法首先通过在 ASUM 模型中加入表示词语类型的隐含变量构建多词型的主题情感联合模型,利用该模型对特定领域的显式评论句进行特征类别下指示词的挖掘。然后,在隐式特征的识别过程中,引入词向量模型作为衡量隐式评论句中线索词与特征指示词集中词项语义相关度的标准,并根据线索词的类型来分情况实现对隐式特征的指派。实验表明,该方法在隐式特征识别方面有着较好的精确度与召回率。但是该方法只能识别隐式评论句的特征类别,却不能进一步识别其所表达的情感倾向。所以在以后的工作中,将尝试研究评论句中隐式情感的识别,以获得评论用户更全面的情感信息。

参考文献:

[1] 冯希亚. 在线产品评论对消费者购买意愿的影响[D]. 南昌:江西师范大学,2017.

[2] 曾令伟. 产品评论中隐式评价对象的抽取研究[D]. 上海:上海交通大学,2014.

[3] 张 杨. 基于特征的在线评论细粒度情感分析方法研究[D]. 上海:上海交通大学,2017.

[4] SCHOUTEN K,FRASINCAR F. Finding implicit features in consumer reviews for sentiment analysis[C]//Web engineering. Toulouse,France:Springer,2014:130-144.

[5] 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. [s. l.]:Association for Computing Machinery,2005:342-351.

[6] XU H,ZHANG F,WANG W. Implicit feature identification in Chinese reviews using explicit topic mining model[J]. Knowledge-Based Systems,2015,76:166-175.

[7] SUN L,CHEN J,LI J,et al. Joint topic-opinion model for implicit feature extracting[C]//2015 10th international conference on intelligent systems and knowledge engineering (ISKE). Taipei:IEEE,2015:208-213.

[8] PRASOJO R E,KACIMI M,NUTT W. Entity and aspect extraction for organizing news comments[C]//Proceedings of the 24th ACM international on conference on information and knowledge management. [s. l.]:Association for Computing Machinery,2015:233-242.

[9] SANTU S K K,SONDHI P,ZHAI C X. Generative feature language models for mining implicit features from customer reviews[C]//Proceedings of the 25th ACM international on conference on information and knowledge management. [s. l.]:Association for Computing Machinery,2016:929-938.

[10] 张 莉,许 鑫. 产品评论中的隐式属性抽取研究[J]. 现代图书情报技术,2015(12):42-47.

[11] ZHANG W,XU H,WAN W. Weakness finder:find product weakness from Chinese reviews by using aspects-based sentiment analysis[J]. Expert Systems with Applications,2012,39(11):10283-10291.

[12] RANA T A,CHEAH Y N. Hybrid rule-based approach for aspect extraction and categorization from customer reviews[C]//2015 9th international conference on IT in Asia (CITA). Kota Samarahan:IEEE,2015:1-5.

[13] MAKADIA N,CHAUDHURI A,VOHRA S. Aspect-based opinion summarization for disparate features[J]. International Journal of Advance Research and Innovative Ideas in Education,2016,2(3):3732-3739.

[14] JO Y,OH A H. Aspect and sentiment unification model for online review analysis[C]//Proceedings of the fourth ACM international conference on Web search and data mining. [s. l.]:Association for Computing Machinery,2011:815-824.

[15] CHEN J,SUN L,PENG Y L,et al. Context weight considered for implicit feature extracting[C]//2015 IEEE international conference on data science and advanced analytics (DSAA). Paris:IEEE,2015:1-5.