Journal of Computer Applications
计算机应用,2016,36(1):175 -180
ISSN 1001-9081
CODEN JYIIDU
2016-01-10
http://www.joca.cn

文章编号:1001-9081(2016)01-0175-06
DOI:10.11772/j.issn.1001-9081.2016.01.0175

面向产品属性的用户情感模型

贾闻俊[1],张　晖[1*],杨春明[1],赵旭剑[1],李　波[1,2]
(1.西南科技大学 计算机科学与技术学院,四川 绵阳 621010;　2.中国科学技术大学 计算机科学与技术学院,合肥 230027)
(*通信作者电子邮箱 zhanghui@swust.edu.cn)

摘　要: 传统情感模型在分析商品评论中的用户情感时面临两个主要问题:1)缺乏针对产品属性的细粒度情感分析;2)自动提取的产品属性其数量须提前确定。针对上述问题,提出了一种细粒度的面向产品属性的用户情感模型(USM)。首先,利用分层狄利克雷过程(HDP)将名词实体聚类形成产品属性并自动获取其数量;然后,结合产品属性中名词实体的权重和评价短语以及情感词典作为先验,利用潜在狄利克雷分布(LDA)对产品属性进行情感分类。实验结果表明,该模型具有较高的情感分类准确率,情感分类平均准确率达87%。该模型与传统的情感模型相比在抽取产品属性和评价短语的情感分类上具有较高的准确率。

关键词: 情感模型;细粒度;产品属性;分层狄利克雷过程;潜在狄利克雷分布
中图分类号: TP391.1　　**文献标志码:** A

User sentiment model oriented to product attribute

JIA Wenjun[1], ZHANG Hui[1*], YANG Chunming[1], ZHAO Xujian[1], LI Bo[1,2]
(1. *School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang Sichuan* 621010, *China*;
2. *School of Computer Science and Technology, University of Science and Technology of China, Hefei Anhui* 230027, *China*)

Abstract: The traditional sentiment model faces two main problems in analyzing user's emotion of product reviews: 1) the lack of fine-grained emotion analysis for product attributes; 2) the number of product attributes shall be defined in advance. In order to alleviate the problems mentioned above, a fine-grained model for product attributes named User Sentiment Model (USM) was proposed. Firstly, the entities were clustered in product attributes by Hierarchical Dirichlet Processes (HDP) and the number of product attributes could be obtained automatically. Then, the combination of the entity weight in product attributes, the evaluation phrase of product attributes and sentiment lexicon was considered as prior. Finally, Latent Dirichlet Allocation (LDA) was used to classify the emotion of product attributes. The experimental results show that the model achieves a high accuracy in sentiment classification and the average accuracy rate of sentiment classification is 87%. Compared with the traditional sentiment model, the proposed model obtains higher accuracy on extracting product attributes as well as sentiment classification of evaluation phrases.

Key words: sentiment model; fine grain; product attribute; Hierarchical Dirichlet Process (HDP); Latent Dirichlet Allocation (LDA)

0　引言

商品评论往往反映了用户对产品的情感倾向(正面、中性、负面),从这些海量评论文本中自动抽取有价值的信息是情感分析的主要任务。传统的情感分析研究主要面向篇章和句子级别文本,实现篇章或句子的情感极性判定[1-3],但在产品评论文本中,通常包含了用户对产品不同属性的评价,如在手机评论文本"我觉得苹果的屏幕很清晰,拍照效果不错,但就是电池不给力。"中,对于手机"续航能力"属性,文本的情感极性为负面,对于手机"屏幕"属性,文本的情感极性则是正面,因此,需要针对产品属性作更细粒度的情感分析。

细粒度的情感分析是情感分析的重要任务之一,需要从海量的产品评论文本中自动识别出评价对象(产品属性),以及针对评价对象的评价短语(情感词)。产品属性的抽取可以通过人工定义和自动抽取完成,人工定义的产品属性领域性强、可移植性差,一个可行的思路是利用潜在狄利克雷分布(Latent Dirichlet Allocation, LDA)[4]自动抽取产品属性,但产品属性数量须提前确定,在实际中应用不便。另外,在对产品属性进行情感分析时,领域性评价短语由于更具有针对性,应给予更高的权重。例如"清晰"对于"屏幕"的情感评价权重应高于"不错"对于"屏幕"的情感评价权重,因此就需要寻找面向具体产品属性的领域评价短语。

针对上述问题,本文提出了一种面向产品属性的用户情感模型。首先通过对产品评论进行解析,利用词性特征、依存句法关系抽取名词实体和其对应的评价短语,形成一组评价关系〈名词实体,评价短语〉;然后,利用分层狄利克雷过程

收稿日期:2015-07-09;修回日期:2015-09-08。
基金项目:四川省教育厅资助项目(14ZB0113);西南科技大学博士基金资助项目(12zx7116)。
作者简介:贾闻俊(1991-),男,四川广元人,硕士研究生,主要研究方向:情感分析、文本分类;　张晖(1972-),男,安徽宿松人,教授,博士,CCF会员,主要研究方向:数据挖掘、知识工程;　杨春明(1980-),男,云南华坪人,副教授,硕士,CCF会员,主要研究方向:文本挖掘、知识工程;　赵旭剑(1984-),男,四川西昌人,讲师,博士,CCF会员,主要研究方向:文本挖掘、Web信息检索;　李波(1977-),男,四川江油人,讲师,博士研究生,CCF会员,主要研究方向:信息过滤、信息安全。

(Hierarchical Dirichlet Process, HDP)[5]对名词实体聚类形成产品属性;最后,结合产品属性中不同名词实体的权重和情感词典作为先验知识,利用潜在狄利克雷分布对评价短语进行情感分类,形成一组带有情感倾向的评价关系〈产品属性,评价关系〉。

本文的主要工作在于:1)提出一种基于分层狄利克雷过程的产品属性提取模型,通过名词实体聚类自动获取产品属性,有效弥补传统模型需要预先确定产品属性个数的不足;2)利用LDA模型,结合产品属性中不同名词实体的权重、评价短语和情感词典作为先验知识,提出一种评价短语的情感分类模型,能够准确抽取面向产品属性的领域评价短语,提高情感分类的准确率。

1 相关工作

传统的产品情感分析主要针对整个产品,Pang等[1-2]利用朴素贝叶斯、支持向量机、最大熵、图分割方法对产品评论直接进行情感分类。Kennedy等[6]通过情感词语的个数进行情感分类;Prabowo等[7]结合规则和有监督学习方法,提高产品评论的情感分类准确率;Li等[8]针对中文评论信息,利用n元特征,对比朴素贝叶斯、支持向量机、最大熵、神经网络方法的情感分类准确率;Jeyapriya等[9]通过频繁项集挖掘获取产品属性,利用朴素贝叶斯对标注数据情感分类;Liu等[10]针对中文产品评论,利用依存句法获取评价对象,将情感强度作为特征进行情感分类;Zheng等[11]利用文档频率选择特征子集,结合支持向量机对中文产品评论进行情感分析。

近年来,细粒度的情感分析成为研究重点,而主题模型由于在词和句子间加入主题层所带来的灵活性在情感分析领域引起了广泛关注。Titov等[12]构建了一种基于多粒度LDA的情感模型(Multi-aspect Sentiment Model, MAS)。Lin等[13]构建了一种基于LDA的情感模型(Joint Sentiment Topic model, JST),该模型在文章-主题-单词中加入情感层,属于4层主题模型。Fu等[14]提出了一种多粒度的情感模型(Multi-aspect Sentiment Analysis for Chinese Online Social Reviews, MSA-COSR)。该模型结合HowNet词典对中文社会新闻评论进行情感主题分类。Jo等[15]针对产品评论,提出一种的多粒度情感模型(Aspect and Sentiment Unification Model, ASUM),该模型以句子为单位分配主题和情感;同时假设每一句只含有一个主题和一种情感。孙艳等[16]提出了一种无监督的混合情感模型(Unsupervised Topic and Sentiment Unification model, UTSU),该模型提出每个词都与主题和情感相关,对每个句子采样情感标签、对每个词采样主题标签。为了提高情感模型对数据自适应能力更高,Kim等[17]针对在线评论,构建了分层情感主题模型(Hierarchical Aspect-Sentiment Model, HASM)。

传统的产品情感分析,不能对产品不同属性进行情感分类;而在进行细粒度的产品情感分析时,需要提前确定产品属性数量,这在实际应用中是很困难的,因此,本文提出了一种细粒度的面向产品属性的用户情感模型,解决了产品不同属性的情感分类问题和产品属性数量须提前确定的问题。

2 主题情感模型的构建

情感模型的构建主要分为4步:1)产品评论解析,利用词性特征、依存句法关系抽取名词实体和评价短语;2)利用分层的狄利克雷过程对名词实体自动聚类形成产品属性;3)结合名词实体的权重和情感词典作为先验知识,利用LDA对产品属性的评价短语进行情感分类;4)获取带有情感倾向的〈产品属性,评价短语〉评价关系。

2.1 名词实体和评价短语的抽取

名词实体是客观存在并可相互区别产品不同属性的名词,评价短语是可以表达名词实体的情感词语。本文定义一种评价关系〈名词实体,评价短语〉,应用两种不同的抽取模式:一种是基于词性模板的抽取方法;另一种是基于依存句法分析的抽取方法。首先,通过词性模板抽取一部分评价关系;然后,通过句法分析抽取剩余的评价关系。

基于词性模板抽取〈名词实体,评价短语〉:首先,定义抽取词性模板,如表1所示;然后,通过中科院分词器(Institute of Computing Technology, Chinese Lexical Analysis System, ICTCLAS)分词,获取不同词语的词性特征;最后,通过词性模板匹配分词后的评论语料,获取候选的评价关系。

表1 抽取词性模板

词性模板	示例
名词+形容词	包装 n / 不错 adj
名词+副词+形容词	屏幕 n / 很 d / 清晰 adj
名词+副词+动名词	外观 n / 很 d / 喜欢 vn

基于依存句法关系抽取〈名词实体,评价短语〉:首先,定义抽取的依存句法关系模板,如表2所示;然后,通过哈尔滨工业大学开发的语言技术平台(Language Technology Platform, LTP)中的句法分析模块进行句法关系分析;最后,获取候选的评价关系。

表2 抽取依存句法关系模板

句法关系	示例
主谓关系(SuBject Verb, SBV)	屏幕←清晰
动宾关系(Verb OBject, VOB)	喜欢→手机
定中关系(ATTribute, ATT)	流畅←系统
动补结构(CoMplemenT, CMT)	续航→持久

例如手机评论中,“IPhone5的屏幕很清晰,拍照效果很好,续航持久,流畅的系统,但是通话效果不好。店家服务很好,包装不错,外观很喜欢,总体来说比较满意,喜欢新手机”。通过词性抽取模板和依存句法关系抽取模板,可以得到如下的评价关系〈屏幕,清晰〉、〈拍照,好〉、〈续航,持久〉、〈系统,流畅〉、〈通话,不好〉、〈服务,很好〉、〈包装,不错〉、〈外观,喜欢〉、〈总体,满意〉、〈手机,喜欢〉。

2.2 产品属性的构建

产品属性由一系列语义相近或相同的名词实体构成。不同产品的属性数量不同,错误地设定产品属性数量,可能会导致构建的产品属性结果准确率低。本文将分层的狄利克雷过程应用于产品属性的构建,自适应于不同的产品评论数据,确定产品属性的数量。其构造方法包含截棍先验构造(Stick-breaking prior construction)、中式连锁餐厅过程(Chinese restaurant franchise process)[5]。

本文产品评论语料为$X_m(m=1,2,\cdots,M)$,每一篇评论语料包含若干名词实体$X_{mn}(m=1,2,\cdots,M_n)$。第m篇评论语

料中的名词实体的分布服从 $\boldsymbol{\theta}_m \sim \mathrm{Dir}(\alpha\boldsymbol{\tau})$，而每一篇评论语料的先验 $\boldsymbol{\tau}$，通过截棍过程构造，$\boldsymbol{\tau} \sim GEM(\alpha)$ [18] 可以通过排序指示因子 $Z_{mn} = k$ 获得。最后，名词实体聚类成 $k \in [1,\infty)$ 组产品属性，其中第 k 组 Z_{mn} 中的第 j 个名词实体的第 i 评价短语，记为 W_{kji}。采用 Heinrich [19] 提出的直接后验采样方法，如式(1) 所示:

$$p(Z_{mn} = k \mid \cdot) \propto (n_{mk}^{\neg mn} + \alpha\tau_k) \cdot \left(\frac{n_{kt}^{\neg mn} + \beta}{n_k^{\neg mn} + V\beta}\right) \tag{1}$$

其中: $\neg mn$ 表示在第 m 篇评论中，除第 n 个名词实体的其他名词实体; n_{mk} 表示第 m 篇评论中，属于第 k 组产品属性的名词实体的个数; $\alpha\tau_k$ 表示第 k 组产品属性的先验概率; n_{kt} 表示产品评论语料集中，在第 k 组产品属性中名词实体 t 的个数; n_k 表示第 k 组产品属性的名词实体的个数; V 表示产品评论语料集中名词实体的个数。

2.3　情感模型的构建

通过分层狄利克雷过程获取〈产品属性，评价短语〉后，使用评价短语近似计算该产品属性的情感倾向。利用 Gibbs 采样计算出不同评价短语在该组产品属性中的权重，如果权重越大，则说明该评价短语更能代表该组产品属性的情感倾向。由于狄利克雷分布作为一种无监督算法不能区分评价短语的情感倾向，本文利用情感词典 [24] 作为先验知识。该情感词典包含情感词语的情感得分值，如果词语的得分大于 0，则该情感词语为正面情感; 如小于 0，则情感词语为负面词语; 如果该评价短语出现在情感词典中，则使用情感词典中情感词语的分数; 如未出现，则设定一个很小的正数 10^{-8} 作为情感词语的分数。通过分层狄利克雷过程获取的产品属性中，包含名词实体在该属性下的概率，所以，将其也作为先验知识。利用狄利克雷分布进行情感分类，如式(2) 所示:

$$p(O_{kj} = s \mid \cdot) \propto \left(\frac{n_{ks}^{\neg j} + \lambda}{n_k^{\neg j} + E\lambda}\right) \cdot \left(\frac{n_{st}^{\neg j} + \mu}{n_s^{\neg j} + S\mu}\right) \cdot p(W_{kji}) \cdot p(W_{kj}) \tag{2}$$

其中: $\neg j$ 表示除名词实体 j 的其他名词实体; n_{ks} 表示第 k 组产品属性中，属于第 s 级情感层的评价短语个数; n_k 表示第 k 组产品属性中评价短语的个数; n_{st} 表示第 s 级情感层中评价短语的 t 个数; n_s 表示第 s 级情感层中评价短语的个数; E 表示评价短语的个数; S 表示评价短语的情感倾向层级。$p(W_{kj})$ 表示第 k 组 Z 中的第 j 个评价短语的权重，即 $p(W_{kj}) = \varphi_{kj} \circ p(W_{kji})$ 表示该评价短语在情感词典的先验值。如果 $p(O_{kj}) \geqslant 0$，则代表该评价短语的情感倾向为正面，并且情感分值越高，则情感倾向为正面的强度越大; 如果 $p(O_{kj}) < 0$，则代表该评价短语的情感倾向为负面，并且情感分值越小，则情感倾向为正面的强度越大。

构建的情感模型的生成过程如算法 1 所示，该模型的概率图模型如图 1 所示，参数定义如表 3 所示。

Algorithm 1　Generation Process of the Sentiment Mode。

Input: A corpus of product reviews;

Output: A set of sentiment relations {product, evaluation phrase};

Begin

Define a base line $\boldsymbol{H}$, $\boldsymbol{H} \sim DP(\gamma\mathrm{Dir}(\beta))$

Drawing G_m from $\boldsymbol{H}$ according to adapted Dirichlet

For an $\boldsymbol{X}_{mn}$ in the n^{th} entities of the review m, draw a new Z_{mn} from G_m, $Z_{mn} \sim G_m$

Draw a sentiment distribution $\boldsymbol{\omega}_s \sim \mathrm{Dir}(\mu)$ based on evaluation phrase and sentiment lexicon and the weight of entities

For each sentiment assigned to k^{th} Z, draw a distribution $\boldsymbol{\zeta}_{kj} \sim \mathrm{Dir}(\lambda)$

For o_{kj}, Draw an sentiment distribution o_{kj} from $\boldsymbol{\zeta}_{kj}$

For i^{th} evaluation phrase W_{kji} of the j^{th} entity in the k^{th} Z, draw W_{kji} from $p(W_{kji} \mid \boldsymbol{\omega}_s, \boldsymbol{\zeta}_{kj})$

End

首先，产品属性中的名词实体分布符合以 $\gamma\mathrm{Dir}(\beta)$ 为基构造的狄利克雷过程，通过 Gibbs 采样获得基分布 $\boldsymbol{H}$; 然后，G_m(G_m 为第 m 篇产品评论语料) 通过 Gibbs 采样基分布 $\boldsymbol{H}$ 获得，第 m 组产品属性中的第 n 个名词实体 Z_{mn} 通过 Gibbs 采样 G_m 获得; 最后，结合名词实体权重和评价短语在情感词典中的情感分值，利用 Gibbs 采样获取每个评价短语的情感倾向。

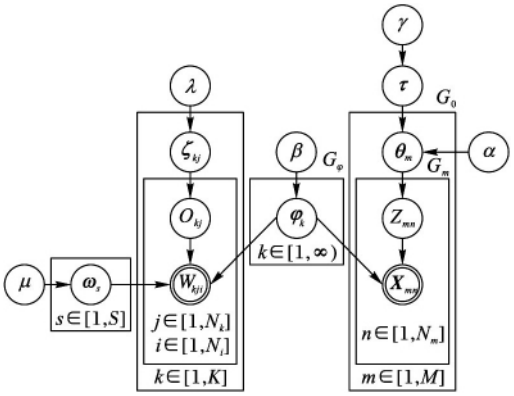

图 1　图模型

表 3　参数定义

参数	含义
M	文档的数量
N_m	每篇文档中含有〈名词实体，评价短语〉评价关系的个数
$\boldsymbol{X}_{mn}$	第 m 篇文档中第 n 个〈名词实体，评价短语〉评价关系中的名词实体
Z_{mn}	第 m 篇文档中第 n 个〈产品属性，评价短语〉评价关系中的产品属性
$\boldsymbol{\theta}_m$	第 m 篇文档中 Z_{mn} 的概率分布
α	$\boldsymbol{\theta}_m$ 的超参数
$\boldsymbol{\tau}$	$\boldsymbol{\theta}_m$ 的先验概率分布
γ	$\boldsymbol{\tau}$ 的超参数
$\boldsymbol{\varphi}_k$	文档集中 Z 的概率分布
β	$\boldsymbol{\varphi}_k$ 的超参数
K	文档集中 Z 的个数
N_k	Z 中评价短语的个数
S	评价短语的情感层级
W_{kji}	第 k 组 Z 中第 j 个 X 的第 i 个评价短语
O_{kj}	第 k 组 Z 中的评价短语
$\boldsymbol{\zeta}_{kj}$	第 k 组 Z 中评价短语的概率分布
λ	$\boldsymbol{\zeta}_k$ 的超参数
$\boldsymbol{\omega}_s$	评价短语的情感概率分布
μ	$\boldsymbol{\omega}_s$ 的超参数

3　实验结果与分析

3.1　实验准备及数据集

本文使用的实验数据来源于“Business to Customer(B2C) 电商手机类目评论信息”[20]、“某门户网站手机评论语料库”[21]、“京东 14 款最热销手机评论”[22] 和“京东商城

iPhone4 Black(黑色)的客户评论数据"[23],去掉不完整和有噪声的数据,最后整理得到关于苹果手机评论6481条。情感词典是使用boson[24]从微博、新闻、论坛等数据的上百万篇情感标注数据当中自动构建的情感极性词典,共含有114767条情感词语。

3.2 参数选择

本文的参数包括β、λ和μ,参数λ采用经验值,$\lambda = 50/S$,S是评价短语的情感级别,本文只作正、负面情感分类,所以,$S = 2$。参数μ采用经验值,$\mu = 0.01$。

参数β影响产品属性的抽取效果,本文通过计算困惑度和产品属性数量来确定参数β,参数β与产品数量曲线如图2所示。参数β与模型困惑度曲线如图3所示。困惑度是一种信息理论的测量方法,用于概率模型的比较,困惑值越小,说明模型越好,计算公式如式(3)所示。同时,产品属性数量不应该过多,因此,本文结合困惑度和产品属性数量综合考虑,设置模型参数。

$$\begin{cases} perplexity = \exp\left(-\left(\sum_{p(w)=1}^{N_m} \lg(p(w))/(N)\right)\right) \\ p(w) = \sum_{Z} p(Z \mid d) * p(w \mid Z) \end{cases} \tag{3}$$

观察图2,当$\beta \in [0.5,1]$,产品属性数量趋于稳定$Z \in [17,20]$,所以应该选择该区间内,困惑值最低的β值。观察图3,当$\beta \in [0.5,1]$,$perplexity \in [16.51,22.13]$。所以,本文综合考虑参数$\beta$对模型困惑度和产品属性数量的影响,设置$\beta = 0.5$,此时,$perplexity = 16.51$,$Z = 20$,模型困惑度和产品属性数量都取得较小值。

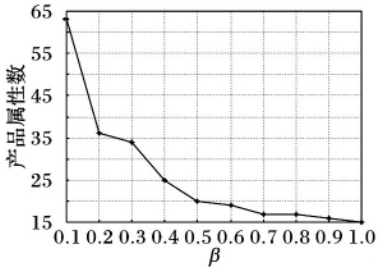

图2 参数β对产品属性数量的影响

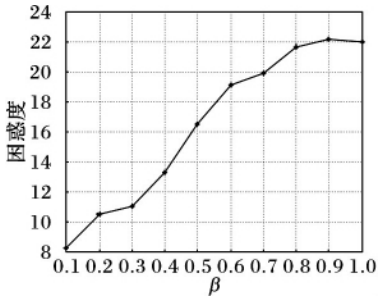

图3 参数β对困惑度的影响

3.3 实验评测

本文采用信息检索领域标准的准确率对实验结果评估。由于实验中产品属性的评价短语众多,所以选择产品属性中评价短语权重前30的词语,使用准确率进行评估。

抽取的产品属性共20组,选取权重前7的产品属性,产品属性中的名词实体如表4所示,对应每组产品属性的评价短语如表5所示。

观察表4后得出结论,每组产品属性都有较好的解释性,例如属性3是关于手机屏幕,属性4是关于手机电池电量,属性6是关于手机的通话质量,属性7是关于手机物流和产品包装的。

表4 前7个产品属性的名词实体

产品属性	名词实体
属性1	游戏 软件 安卓 越狱 上网 功能 操作 iOS 程序
属性2	系统 外观 运行 行货 iOS7 整体 速度 反应 开机
属性3	屏幕 划痕 触屏 外观 反应 触摸屏 灰尘 后壳 手感
属性4	电池 用电 电量 正品 充电 待机 4s 智能机 移动电源
属性5	苹果 产品 三星 质量 价钱 4s 小米 体验 消费者
属性6	信号 通话 声音 耳机 杂音 听筒 电信 移动 联通
属性7	包装 速度 发票 物流 盒子 卖家 商城 售后服务 行货

表5 前7个产品属性的评价短语

产品属性	评价短语	
	正面	负面
属性1	不错 实用 满意 喜欢 华丽 完美 一流 实惠 定制 灵活	坑爹 死机 无语 爱不释手 伤心 卡住 不好 不行 发热 麻烦
属性2	定制 简约 简单 独特 极致 便捷 先进 成熟 简洁 过硬	无语 费劲 伤心 繁琐 复杂 果断 别扭 不清楚 混乱 不稳定
属性3	丰富 亮点 旗舰 高贵 自然 超高 典雅 绚丽 轻松 惊艳	错误 变态 卡机 奇葩 不多说 不灵 不和 不行 惊讶 差劲
属性4	细心 迅速 不卡 不赖 有用 知足 精美 五星 安全 ok	差点 卡顿 差劲 恶心 小气 一天到晚 不满 恼火 紧张 头痛
属性5	人性化 满分 优雅 成功 专业 细致 旗舰 高贵 认真 突出	折腾 气愤 轻微 头疼 窝火 尴尬 要命 揪心 悲催 奇葩
属性6	最高 精致 清晰 无可挑剔连续 杠杠的 优秀 完美 强大	失望 清楚 不好用 悲剧 吓人 模糊 无语 不多说 气愤 吃力
属性7	正好 简便 不容易 踏实 激动 惊喜 秀气 可靠 提升 物美价廉	差劲 坑爹 蛋疼 寒酸 好不容易 粗劣 不容易 马虎 受不了 郁闷

观察表5后得出结论,本文提出的模型能够提取一些产品属性的特有评价短语,例如:属性3中的绚丽和惊艳。这些评价短语在特定领域的情感倾向性更强。

本文对抽取产品属性中评价短语权重前20的词语计算其情感分类的准确率如图4所示。

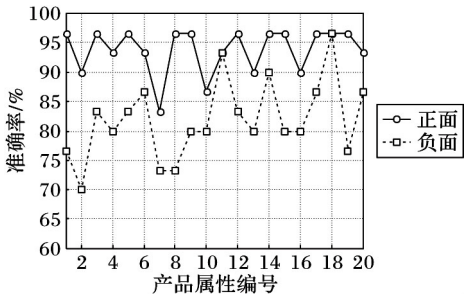

图4 产品属性中评价短语的情感分类准确率

为了验证本文所提出的模型的有效性,本文采用的对比模型有3种:模型1是LDA模型,不能对产品属性进行情感分类。模型2是Lin提出的JST,该模型首次将LDA应用在情感分析领域,将原有的3层主题模型增加情感层后,形成4层图模型,为主题模型在情感领域的应用奠定了基础;同时,该模型假设名词实体和评价短语的权重相同。模型3是Jo提出的ASUM,该模型针对在线评论语料,假设名词实体和情感的采样单位是句子,构建了情感模型。由于模型1、模型2

和模型 3 都不能确定产品属性的数量,所以,产品属性的数量是本文算法计算结果为 20 组。LDA 模型的实验结果如表 6 所示。模型 2、模型 3 和本文模型都能形成〈产品属性,评价短语〉的评价关系,所以最终评测产品属性中评价短语的准确率如图 5 所示。

表 6　非情感倾向的产品属性

产品属性	名词实体、评价短语
属性 1	苹果 体验 确实 价格 三星 产品 质量 用户 操作
属性 2	不错 系统 性价比 流畅 外观 靓丽 包装 屏幕 电池
属性 3	性能 速度 不错 强大 满意 分辨率 像素 屏幕 电池
属性 4	屏幕 一般 不错 灵敏 反应 触屏 边缘 清晰 不错
属性 5	外观 靓丽 漂亮 不错 喜欢 好看 手感 外壳 速度
属性 6	功能 软件 强大 iPhone 越狱 操作 价格 应用 完美
属性 7	系统 流畅 运行 不错 死机 速度 国行 评价 稳定

观察表 6 得到,传统的 LDA 模型抽取的结果是名词实体和评价短语的组合,不能利用评价短语对名词实体进行情感分析。

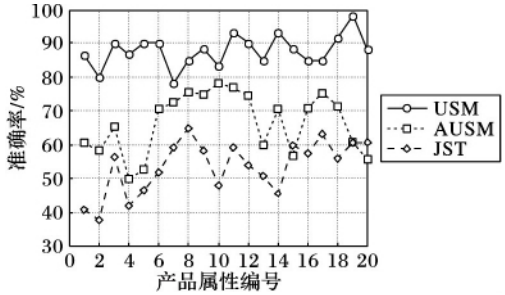

图 5　USM 与 JST、ASUM 情感分类准确率对比

观察图 5 得到,本文提出的 User Sentiment Model(USM) 在产品属性中评价短语的情感分类准确率高于 JST 和 ASUM。分析 USM 实验效果优于 JST 和 ASUM 的原因是其假设更符合实际情况,具体体现在: 首先,JST 和 ASUM 都是基于 LDA 的情感模型,因此需要提前确定产品属性的数量,由于数据领域性不同导致提前确定产品属性数量困难,从而会影响产品属性的抽取的准确率,而且产品属性的情感倾向不一定一致,会进一步降低情感分类的准确率。USM 可以自适应于实验数据,自动确定产品属性的数量,克服了以上缺点。其次,在产品评论语料中,描述产品属性的名词实体是没有情感色彩的,需要通过修饰名词实体的评价短语体现。本文将名词实体与其对应的评价短语分开采样,优于 JST 和 ASUM 将名词实体与评价短语一起采样。最后,在产品评论语料中,一个句子中包含多个名词实体与评价短语,并且不同的名词实体表达不同的情感倾向。ASUM 采样是以句子为单位,一个句子只能包含一个名词实体和一种情感色彩,而 USM 假设以词为单位进行采样,更符合实际情况。

4　结语

针对传统的产品评论情感分析中,缺乏细粒度的产品属性情感分析、产品属性抽取数量必须提前确定、产品属性词语与情感词语赋予相同的权重等问题,本文提出了一种细粒度面向产品属性的用户情感模型。该模型首先定义了两组评价关系〈名词实体,评价短语〉和〈产品属性,评价短语〉。然后利用 HDP 对〈名词实体,评价短语〉进行聚类,形成多组的〈产品属性,评价短语〉,组的数量自适应于数据。最后结合情感词典和狄利克雷分布作为评价短语的先验知识进行情感分类。实验结果表明,在抽取产品属性和评价短语的情感分析中,该模型与传统的情感主题模型相比具有较高的准确率。

本文所提出的模型接下来可以在以下两个方面进行改进: 首先,本文在抽取〈名词实体,评价短语〉的评价关系中,简单地使用词性模板和句法分析来进行评价关系的抽取,在下一步将使用条件随机场来抽取这组评价关系; 其次,该模型在构建情感主题模型中是分步完成的,对联合模型的构建和推理有待进一步研究。

参考文献:

[1] PANG B, LEE L, VAITHYANATHAN S. Thumbs up?: sentiment classification using machine learning techniques [C]// EMNLP 2002: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2002: 79 – 86.

[2] PANG B, LEE L. A sentimental education : sentiment analysis using subjectivity summarization based on minimum cuts [C]// Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2004: 271 – 280.

[3] FU G, WANG X. Chinese sentence-level sentiment classification based on fuzzy sets [C]// Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Stroudsburg, PA: Association for Computational Linguistics, 2010: 312 – 319.

[4] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. The journal of machine learning research, 2003, 3: 993 – 1022.

[5] TEH Y W, JORDAN M I, BEAL M J, et al. Hierarchical Dirichlet processes [J]. Journal of the American statistical association, 2006, 101(476): 1566 – 1581.

[6] KENNEDY A, INKPEN D. Sentiment classification of movie reviews using contextual valence shifters [J]. Computational intelligence, 2006, 22(2): 110 – 125.

[7] PRABOWO R, THELWALL M. Sentiment analysis: a combined approach [J]. Journal of informetrics, 2009, 3(2): 143 – 157.

[8] YE Q, SHI W, LI Y. Sentiment classification for movie reviews in Chinese by improved semantic oriented approach [C]// HICSS'06: Proceedings of the 39th Annual Hawaii International Conference on System Sciences. Piscataway, NJ: IEEE, 2006: 53b.

[9] JEYAPRIYA A, SELVI C S. Extracting aspects and mining opinions in product reviews using supervised learning algorithm [C]// Proceedings of the 2015 2nd International Conference on Electronics and Communication Systems. Piscataway, NJ: IEEE, 2015: 548 – 552.

[10] LIU L, SONG W, WANG H, et al. A novel feature-based method for sentiment analysis of Chinese product reviews [J]. China communications, 11(3): 154 – 164.

[11] ZHENG L, WANG H, GAO S. Sentimental feature selection for sentiment analysis of Chinese online reviews [J]. International journal of machine learning and cybernetics, 2015: 1 – 10.

[12] TITOV I, MCDONALD R. Modeling online reviews with multi-grain topic models [C]// Proceeding of the 17th International Conference on World Wide Web. New York: ACM, 2008: 111 – 120.

[13] LIN C, HE Y. Joint sentiment/topic model for sentiment analysis [C]// Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York: ACM, 2009: 375 - 384.

[14] FU X, LIU G, GUO Y, et al. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon [J]. Knowledge-based systems, 2013, 37: 186 - 195.

[15] 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. New York: ACM, 2011: 815 - 824.

[16] **孙艳,周学广,付伟. 基于主题情感混合模型的无监督文本情感分析**[J]. **北京大学学报(自然科学版)**,2013,49(1):102 - 108. (SUN Y, ZHOU X G, FU W. Unsupervised topic and sentiment unification model for sentiment analysis [J]. Acta scientiarum naturalium universitatis Pekinensis, 2013, 49(1): 102 - 108.)

[17] KIM S, ZHANG J, CHEN Z, et al. A hierarchical aspect-sentiment model for online reviews [C]// Proceedings of the Twenty-Seventh Conference on Artificial Intelligence. Menlo Park, California: AAAI, 2013: 526 - 533.

[18] SETHURAMAN J. A constructive definition of Dirichlet priors [J]. Statistica sinica, 1994, 4: 639 - 650.

[19] HEINRICH G. Infinite LDA implementing the HDP with minimum code complexity [EB/OL]. [2014-11-15]. http://arbylon.net/publications/ilda.pdf.

[20] B2C **电商手机类目评论信息**[EB/OL]. [2015-03-22]. http://www.datatang.com/data/15516/. (Business to customer mobile phone category review information [EB/OL]. [2015-03-22]. http://www.datatang.com/data/15516/.)

[21] **某门户网站手机评论语料**[EB/OL]. [2015-03-23]. http://www.datatang.com/data/44156/. (A portal website mobile phone reviews corpus [EB/OL]. [2015-03-23]. http://www.datatang.com/data/44156/.)

[22] **京东14款最热销手机评论**[EB/OL]. [2015-03-25]. http://www.datatang.com/data/46145/. (Jindong's 14 most popular mobile phone reviews [EB/OL]. [2015-03-25]. http://www.datatang.com/data/46145/.)

[23] **京东商城黑色苹果4的评论数据**[EB/OL]. [2015-03-24]. http://www.datatang.com/data/46145/. (Jindong's black iPhone4 reviews. [EB/OL]. [2015-03-24]. http://www.datatang.com/data/46145/.)

[24] **泊松. 情感词典**[EB/OL]. [2015-03-29]. http://www.searchforum.org. (Boson. Sentiment lexicon [EB/OL]. [2015-03-29]. http://www.searchforum.org.)

Background

This work is partially supported by the Education Department of Sichuan Province (14ZB0113) and the Fundamental Research Funds for the Doctor of the Southwest Science and Technology University (12zx7116).

JIA Wenjun, born in 1991, M. S. candidate. His research interests include sentiment analysis, text classification.

ZHANG Hui, born in 1972, Ph. D., professor. His research interests include data mining, knowledge engineering.

YANG Chunming, born in 1980, M. S., associate professor. His research interests include text mining, knowledge engineering.

ZHAO Xunjian, born in 1984, Ph. D., lecturer. His research interests include text mining, Web information search.

LI Bo, born in 1977, Ph. D. candidate, lecturer. His research interests include information filtering, information security.

(上接第174页)

[8] YE M, YIN P, LEE W, et al. Exploiting geographical influence for collaborative point-of-interest recommendation [C]// SIGIR'11: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2011: 325 - 334.

[9] BERJANI B, STRUFE T. A recommendation system for spots in location-based online social networks [C]// SNS'11: Proceedings of the 4th Workshop on Social Network Systems. New York: ACM, 2011: Article No. 4.

[10] YE M, YIN P, LEE W. Location recommendation for location-based social networks [C]// GIS'10: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM, 2010: 458 - 461.

[11] PARK M, HONG J, CHO S. Location-based recommendation system using Bayesian user's preference model in mobile devices [C]// UIC 2007: Proceedings of the 4th International Conference on Ubiquitous Intelligence and Computing. Berlin: Springer, 2007: 1130 - 1139.

[12] YUAN Q, CONG G, MA Z, et al. Time-aware point-if-interest recommendation [C]// SIGIR'13: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2013: 363 - 372.

[13] LEVANDOSKI J J, SARWAT M, ELDAWY A, et al. LARS: a location-aware recommender system [C]// ICDE'12: Proceedings of the 2012 IEEE 28th International Conference on Data Engineering. Washington, D.C.: IEEE Computer Society, 2012: 450 - 461.

[14] LIU H, HU Z, MIAN A, et al. A new user similarity model to improve the accuracy of collaborative filtering [J]. Knowledge-based systems, 2014, 56: 156 - 166.

[15] **邓爱林,朱扬勇,施伯乐. 基于项目评分预测的协同过滤推荐算法**[J]. **软件学报**,2003,14(9):1621 - 1628. (DENG A L, ZHU Y Y, SHI B L. A collaborative filtering recommendation algorithm based on item rating prediction [J]. Journal of software, 2003, 14(9): 1621 - 1628.)

Background

This work is partially supported by the Six Talents Peak Project in Jiangsu Province.

WANG Fuqiang, born in 1989, M. S. candidate. His research interests include data mining.

PENG Furong, born in 1987, Ph. D. candidate. His research interests include data mining, recommendation system.

DING Xiaohuan, born in 1991, M. S. candidate. Her research interests include data mining.

LU Jianfeng, born in 1969, Ph. D., professor. His research interests include intelligent system, data mining.