基于主题模型的(Aspect,Rating)摘要生成方法研究

吕 品,汪 鑫,罗宜元,计春雷
(上海电机学院电子信息学院,上海 201306)

摘 要: 提出基于短语参数学习的主题模型 TMPP(Topic Model based on Phrase Parameter)对在线评论中被评价实体的 aspect 和与之对应的 rating 进行抽取. TMPP 具有三个特点:1)评论用"短语袋"表示;2)将标准的 LDA 中表示文档-主题的参数扩展为(aspect,rating)集;3)融合了先验知识. 介绍了 TMPP 模型参数的物理含义、模型的生成过程以及先验知识的获取和表示方法;阐述了在 TMPP 模型中引入方面集聚类使用先验知识的原因与好处、TMPP 模型提取(方面,等级)对形成(aspect,rating)摘要的原理. 以真实的在线产品评论数据集为实验对象,在实验过程中引入先验知识的方面识别分析和等级预测精度分析,列出了五类产品相关方面和对立的情感词的实验结果. 通过与已有的基线方法比较,实验表明若评论集中每篇评论有一个总体等级,TMPP 能产生高质量的(aspect,rating)摘要.

关键词: 主题模型;(aspect, rating)摘要;短语袋;TMPP
中图分类号: TN911 **文献标识码:** A **文章编号:** 0372-2112 (2016)12-3036-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.12.032

(Aspect,Rating) Summarization Based on Topic Model

LÜ Pin,WANG Xing,LUO Yi-yuan,JI Chun-lei
(*School of Electronic and Information,Shanghai Dianji University,Shanghai* 201306,*China*)

Abstract: This paper proposes a topic model TMPP (Topic Model based on Phrase Parameter), which can extract the aspects and associated with their ratings for the evaluated entities in online reviews. TMPP has three characterisitcs: (1) It assumes the review is represented as a bag-of-phrase. (2) It extends the document-topic parameter from the standard LDA as a set of (aspect ,rating). (3) It incorporates the prior knowledge. We introduce the physical meaning of each parameter for the TMPP, the generative process for the TMPP and the representation of the prior knowledge. Furthermore, the reason and advantage of incorporating the aspect cluster into the TMPP are presented; the mechanism of obtaining the (aspect,rating) is also given by extracting the aspects and associated with their ratings from the online product reviews. We conduct extensive experiments on a very large real life dataset from taobao.com and find that TMPP can produce high quality (aspect,rating) summarization if each review has an overall rating by comparing the performance between existing baseline models and TMPP.

Key words: topic model; (aspect,rating) summarization; bag-of-phrase; topic model based on phrase parameter (TMPP)

1 引言

Web 技术的发展使在线评论成为决策支持的有价值资源[1]. 然而,阅读者要从海量评论中找到所有不同甚至可能相反的观点几乎不可能. 因此,对在线评论进行挖掘,生成(aspect,rating)摘要,方便用户对目标实体获得不同视角的评价必然成为情感分析研究不可或缺的一部分. Aspect(方面)指的是被评价实体的某个物理组成部分、功能或性质、亦可以是被评论事件的某一个特征等[2]. Rating(等级)是用 1 到 5 之间的整数表示的情感满意度. 一般,评论网站只要求用户对被评价实体给出一个用不同星号个数表示的总体评价(总体等

收稿日期: 2014-12-24;修回日期:2016-08-22;责任编辑: 郭游
基金项目:国家自然科学基金青年基金(No. 61402280);上海电机学院计算机科学与技术优势学科(No. 16YSXK04);上海电机学院科研计划项目(No. B1-0227-16-032-031)

级).尽管总体等级对潜在用户的决策有帮助,但提供的信息并不充分,因为不同用户可能有完全不同的需求.若能从在线评论中挖掘得到(aspect,rating)摘要,便能让潜在的购买者更深入的了解产品质量,通过对同类型的不同产品进行权衡,最终做出明智的购买决定.本文提出基于短语参数学习的主题模型 TMPP 对在线评论进行(aspect,rating)摘要挖掘.挖掘(aspect,rating)摘要由 2 项任务构成:1) 方面识别,即抽取被评价实体的相关方面集;2)等级预测,即为每一个方面分配一个整数,表示评论者对该方面的情感满意度.

2　相关研究工作

近年来,绝大多数观点文摘挖掘研究工作的重心是方面识别[1,3],只有极少数工作是在方面识别的同时给出方面的预测等级[4,5].方面识别的经典方法有 2 类:频率方法和主题模型方法.频率方法采用在高频率名词短语上应用一些约束识别被评价实体(产品)的方面[2,6~8].该方法的局限性在于:1)可能会丢失低频率的方面和它们的变化形式[9],并产生许多不是表示被评价实体方面的名词;2)需要人工调整各种参数,移植性差.主题模型方法能克服以上不足[5,10~12],但通常采用先识别方面,后再对方面进行等级预测的手段,方面识别和其相应的等级预测以串行方式进行,其结果会导致挖掘过程中的错误累积.

另外,评论文本有两种表示模型:"词袋"模型和"短语袋"模型.短语是对原始的评论经过预处理后得到的(t,s)信息对,t 表示 aspect,s 表示与某一 aspect 对应的观点.研究工作[9~11,13,14]使用"词袋"模型表示评论,虽然它们也采用了一些技术挖掘评论中的局部主题或子主题(方面),但研究重心是将识别的方面按情感进行聚类.然而,(aspect,rating)文摘挖掘目标是识别方面和其对应的等级,即尝试从同一被评价实体的评论集合中推断出被评价实体的方面和其对应的评价等级.文献[4~5]使用"短语袋"模型表示评论,前者使用主题模型 PLSA 和 Structured PLSA 对短评论挖掘,产生(aspect,rating)文摘;后者提出了方面与其等级具有依赖关系的 ILDA 主题模型.然而,尽管主题模型能输出表示某一主题的词集,但词集中的词往往在语义上不相干,即主题的质量不高[1].

为了解决挖掘过程中的错误累积和主题质量欠佳问题,本文设计了主题模型 TMPP,它用"短语袋"模型表示评论,将标准的 LDA 中表示文档-主题的参数 θ 扩展为(aspect,rating)集,对 aspect 和 rating 同时建模,以减少错误累积;引入潜在聚类变量 c 表示领域先验知识,指导模型产生质量更好的方面.

3　TMPP 模型

3.1　TMPP 引入方面集聚类的原因与好处

对被评价实体进行评价,就是从在线评论中抽取评价实体的各方面(aspect),并基于评论的总体等级,用 1 到 5 之间的整数预测评价实体各方面的情感满意度(情感等级 rating),于是产生了形成(aspect,rating)摘要的方面和与之相应的情感等级对.一个评价实体有许多(aspect,rating)对,故要进行方面聚类,这就是 TMPP 引入方面聚类的原因.为了克服总体等级的片面性,TMPP 模型整合了一个方面聚类变量 c,将总体等级分解成每个方面对应的情感等级,产生一个有利于潜在用户进行决策支持的(aspect,rating)摘要,体现了 TMPP 引入方面聚类的好处.

3.2　TMPP 输出(aspect,rating)信息对的原理

TMPP 模型获取(aspect,rating)信息对的原理简述如下:

(1) 利用整合了先验知识的 TMPP 寻找被评价实体中语义上更连贯的方面.

(2) 通过聚类算法对相同聚类的等级预测对数量和不同聚类的等级预测对数量进行分类.x 表示相同聚类的等级预测对数量,y 表示不同聚类的等级预测对数量.

(3) 引用等级预测的聚类相似度的度量标准公式来预测等级相似度值.

(4) 最终,获取如本文表 3 至表 6 的被评价实体的评论摘要表.

3.3　TMPP 模型参数的物理含义

TMPP 模型的盘子示意图如图 1 所示.

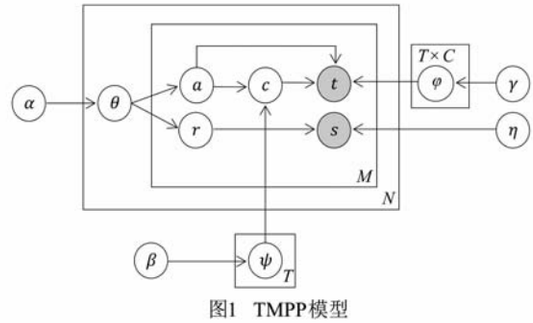

图1　TMPP模型

其中,模型参数的物理含义如下:

a:潜在方面(aspect);

r: 方面对应的潜在等级(rating);

c: 潜在的聚类变量;

t: 重要的方面词,是被观察变量;

s: 重要方面词所对应的情感词,是被观察变量;

(t_m,s_m): 第 m 对观点短语,$m=1,2,\cdots,M$;

α,β: Dirichlet 参数;

θ:服从参数为 α 的狄利克雷分布的随机变量,是文档层的(aspect,rating)集. 对每一对(aspect,rating),θ 包含了产生 aspect 和 rating 组合的概率,每一篇评论抽样一次 θ。固定 θ 后,再为该评论产生观点短语,且假定潜变量 a_m 和 r_m 被独立抽样;

$T\times C$:聚类结果,T 为方面的个数,C 为聚类的个数;

ψ,η: 多项式分布参数;

ψ:对 $p(\text{cluster}|\text{aspect})$ 分布建模,ψ 是服从参数为 β 的 Dirichlet 分布;

φ:对 $p(t|\text{aspect},\text{cluster})$ 分布建模,φ 是服从参数为 γ 的 Dirichlet 分布.

3.4 TMPP 模型的生成过程

与文献[4]不同,TMPP 整合了潜在聚类变量 c 连接潜在方面 a 和被观察词 t,输入是 N 篇评论,T 个方面,C 个聚类,每一篇评论有 M 个短语. TMPP 用随机变量 ψ 对 $p(\text{cluster}|\text{aspect})$ 分布建模;用随机变量 ψ 对 $p(t|\text{aspect},\text{cluster})$ 分布建模,并把随机变量 θ 作为高层的(aspect,rating)集. 对每一个 (aspect,rating) 对,θ 包含了产生 aspect 和 rating 组合的概率. TMPP 为每一篇评论抽样一次 θ,固定 θ 后,再为该评论产生观点短语,且假定潜在变量 a_m 和 r_m 被独立抽样,其生成过程如下:

1. 选择 $\theta\sim\text{Dir}(\alpha),\psi\sim\text{Dir}(\beta),\varphi\sim\text{Dir}(\gamma)$

2. 选择 $c\sim\text{Multi}(\psi)$

3. 对于每一对观点短语$(t_m,s_m),m\in\{1,2,\cdots,M\}$

(a)选择 $a_m\sim P(a_m|\theta)$ 和 $r_m\sim P(r_m|\theta)$

(b)选择 $c\sim P(c|a_m)$

(c)选择 $t_m\sim P(t_m|a_m,c,\varphi)$ 和 $s_m\sim P(s_m|r_m,\eta)$

$P(t_m|a_m,c,\varphi)$ 和 $P(s_m|r_m,c,\eta)$ 分别是以 a_m,c 和 r_m 为条件的多项式分布. 因此,联合概率分布如公式(1)所示.

$$P(a,r,t,s,\theta,c|\alpha,\beta,\gamma,\eta)=p(\theta|\alpha)\prod_{m=1}^{M}[p(a_m|\theta)p(r_m|\theta)p(c|a_m)p(t_m|a_m,c,\varphi)p(s_m|r_m,\eta)] \tag{1}$$

已知一篇评论有 M 个短语,关键的推断是计算式(2)所示的潜在变量的后验概率.

$$P(a,r,\theta,c|t,s,\alpha,\beta,\gamma,\eta)=\frac{P(a,r,t,s,c,\theta|\alpha,\beta,\gamma,\eta)}{P(t,s|\alpha,\beta,\gamma,\eta)} \tag{2}$$

4 先验知识

4.1 先验知识的获取

领域先验知识可从 Web 上中获取,因为通过调查发现,尽管评论文本的领域不同,但不同领域上许多被评价实体的方面是相同的. 因此,可把从不同领域集中挖掘出的相同方面作为主题模型的先验知识,让这些先验知识指导 TMPP 模型产生高质量的方面. 算法 1 给出了先验知识获取的具体方法.

算法 1 先验知识获取方法

Input:

多个领域的评论语料

Output:

知识 K

方法:

1. for each $D_i\in D_L$ do
2. $A_i\leftarrow\text{LDA}(D_i)$; //在每一个评论语料 D_i 上运行 LDA,并将得到的主题集赋给方面集 A_i
3. endfor
4. $A\leftarrow\cup_i A_i$;
5. $TC\leftarrow k\text{-means}(A)$; //对所有领域产生的方面集 A 执行 k-means 聚类,产生一些连贯的主题簇
6. for each $T_j\in TC$ do
7. $K_j\leftarrow\text{FPM}(T_j)$; //对每一个聚类 T_j 执行频繁项集挖掘产生频繁 2 - 模式集表示知识
8. endfor
9. $K\leftarrow\cup_j K_j$;

算法 1 包含 3 个步骤:1)在每一个领域的语料上运行 LDA[15];2)对 LDA 运行得到的主题集进行聚类;3)从每一个聚类中挖掘出频繁模式. 第 1 步执行后,算法 1 获得一个主题集,选取每一主题下概率较高的词表示主题. 由于质量高的知识应该跨领域共享主题,所以可利用频率方法识别频繁出现的词作为先验知识,以保证知识的质量. 但是,对于先验知识,还存在 2 个需要解决的问题:1)特定的方面可能仅出现在该方面所在领域. 如果在频率方法中简单使用一个频率阈值,将无法区分一般的方面和特定的方面;2)词在不同的领域可能具有不同的含义. 算法的第 2 步是对每一个主题执行 k-means 算法,得到主题聚类. 为了实现第 3 步中的知识挖掘,采用了频繁模式挖掘[16],其目标是找到所有满足最小支持度计数的模式. 一个模式就是一个词集合,所有模式组成了先验知识集合,简称先验知识基.

4.2 先验知识的表示方法

由于知识从每一个主题聚类中抽取,所以把经过频繁模式挖掘得到的先验知识基表示为聚类的集合. 每一个聚类由一个频繁 2-模式集组成.

例如:聚类 1:{电池,寿命},{电池,小时},{电池,长}

聚类 2:{服务,支持},{支持,顾客},{服务,顾客}

实验中挖掘了频繁 2-模式和频繁 3-模式,在利用它们指导主题模型生成方面的连贯性评估中发现,频繁 2-模式的性能优于频繁 3-模式. 与此同时,人工观察

发现若属于相同主题的两个词出现在同一集合中,则更能体现词的语义关系. 这也说明模式越长,包含错误的可能性越大.

5　先验知识的使用

TMPP 使用阻塞式 Gibbs 进行推理[17]. 对文档中的每一个词 w_i,Gibbs 能自动减少方面 a 和聚类 c 的关联. 基于 Gibbs 的条件分布(式(6))能同时抽样方面 a_m 和包含了 w_i 的聚类 c. 除了考虑 a_m 与词 w_i 之间的匹配外,在计算该条件分布的过程中,还考虑了如下两个问题:

(1)聚类 c 的作用

聚类变量的作用:1)判定 c 是否是词 w_i 的先验知识;2)控制词 w 和 w'概率的增加. 已知某一领域的评论语料,c 是 w_i 的先验知识意味着:包含 w_i 的聚类 c 中的频繁 2-模式也是实际领域评论语料的先验知识. 如果 c 是 w_i 的先验知识,则认为 c 中的先验知识有用,且能被提供给 TMPP 模型用于指导生成较高质量的 a_m;否则,对于 w_i,c 不是合适的先验知识,不能用于指导 TMPP 模型. 基于共文档频率[18],度量了 c 中 w_i 之间的共现,如式(3)所示.

$$\mathrm{Co\_Doc}(w,w') = \frac{D(w,w')+1}{(D(w)+D(w'))\times\frac{1}{2}+1} \tag{3}$$

其中,(w,w')表示聚类 c 中的频繁 2-模式. $D(w,w')$表示同时包含词 w 和 w'的评论数量,$D(w)$表示只包含词 w 的评论数量. 公式(3)中分子与分母同时加 1 平滑是避免共文档频率为 0 的情况.

此外,给 w_i 分配方面 a_m 和聚类 c 不仅增加了 a_m 和 c 与 w_i 相关的概率,而且还可能使 a_m 和聚类 c 与 w'有关联. 本文利用 Generalized Plya urn(GPU)模型表示语义相关的词[19]. w'与 c 中的 w_i 共享了一个频繁 2-模式. 概率增加量由公式(4)定义的矩阵 $A_{c,w',w}$ 来确定[20]:

$$A_{c,w',w}=\begin{cases}1, & \text{if } w=w'\\ \delta, & \text{if}(w,w')\in c, w\neq w'\\ 0, & \text{otherwise}\end{cases} \tag{4}$$

观察公式(4)中的 w,值 1 控制了 w 的概率的增加;值 δ 控制了 w'的概率增加.

(2)c 与 a_m 的一致性

c 和 a_m 的一致性表示聚类 c 中所有频繁 2-模式是否反映方面 a_m. 如果 c 和 a_m 一致,那么 c 中所有频繁 2-模式中的词应该是方面 a_m 中的热点词. 本文使用对称的 KL-Divergence 作为聚类 c 分布 Dist_c 和方面 a_m 的分布 Dist_a 之间的一致性度量. 对于 Dist_c,由于 c 中的词没有先验偏好,所以对 c 中的所有词都使用均匀分布. 对于 Dist_a,使用排名前 15 的词表示方面 a_m. 一致性计算如公式(5)所示.

$$\mathrm{Agreement}(c,a)=\frac{1}{\mathrm{KL}(\mathrm{Dist}_c,\mathrm{Dist}_a)} \tag{5}$$

Dist_c 和 Dist_a 差别越小,c 和 a_m 之间的一致性越高.

式(3),(4)和(5)一起形成阻塞式 Gibbs,如式(6)所示,它能在确定先验知识有用性的同时对 TMPP 模型产生较好质量的方面提供指导.

$$\begin{aligned}P(a_m=a,c_j=c\mid a^{-j},c^{-j},w,\alpha,\beta,\gamma,A)\propto&\\ \sum_{(w,w')\in c}\mathrm{Co\_Doc}(w,w')\times\mathrm{Agreement}(c,a)&\\ \times\frac{n_{m,t}^{-j}+\alpha}{\sum_{a'=1}^{T}(n_{m,a}^{-j}+\alpha)}\times\frac{\sum_{w'=1}^{V}\sum_{v'=1}^{V}A_{c,w',w}\times n_{a,c,v'}^{-j}+\beta}{\sum_{c'=1}^{C}(\sum_{w'=1}^{V}\sum_{v'=1}^{V}A_{c,w',w}\times n_{a,c,v'}^{-i}+\beta)}&\\ \times\frac{\sum_{w'=1}^{V}A_{c,w',w_i}\times n_{a,c,w'}^{-j}+\gamma}{\sum_{v'=1}^{V}(\sum_{w'=1}^{V}A_{c,w',w_i}\times n_{a,c,w'}^{-j}+\gamma)}&\end{aligned} \tag{6}$$

其中,n^{-j}表示除 a_m 和 c_j 的当前分配以外的计数,例如:a^{-j}和 c^{-j}. $n_{m,a}$表示方面 a 被分配到评论 m 中的词的次数. $n_{a,c}$表示聚类 c 出现在方面 a 中的次数. $n_{a,c,v}$表示词 v 同时出现在方面 a 和聚类 c 中的次数. α,β,γ 是预先定义好的超参数.

尽管以上阻塞式 Gibbs 能区分有用的知识和不合适的知识,但可能存在对于某一特定词,该词不在任何一个聚类中,即该词没有任何对应的先验知识. 为解决这种问题,定义单一聚类概念,即为词 w 增加一个只有频繁 1-模式的聚类$\{w,w\}$. 由于单一聚类并不包含任何知识,仅仅只有词本身,所以式(3)和式(5)不成立. 实验中就使用所有非单一聚类的共文档频率的平均值和一致性平均值作为单一聚类的式(3)和式(5)的计算值.

6　实验

与 TMPP 模型比较的两种基线方法分别是 LDA[4] 和 ILDA[4]. 这 3 个模型都使用“短语袋”模型表示评论,不同的是 TMPP 模型增加了领域先验知识. 因此,比较的目的是观察主题模型使用先验知识是否能产生更高质量的方面.

6.1　实验设置

为获得先验知识,从淘宝上采集了 30 个领域的评论,如表 1 所示. 每一个领域包含 1000 篇评论. 使用中国科学院计算机所的中文分词与词性标注工具 ICTCLAS 对评论语料进行分词与词性标注,并利用哈工大中文停用词表过滤了评论中的无关词. 由于获取先验

知识的评论语料是在标准的 LDA 模型上运行,所以不需要将评论文本预处理为观点短语集. 对于 LDA 模型,设置参数 $\alpha=1,\beta=0.1,T=15$,潜在变量 θ 和 z 的后验估计共执行了 1000 次迭代,得到的每一个主题(方面)只取概率排序在前 15 的词. 在运用 k-means 聚类算法对得到的方面集进行划分时,设置的聚类数目为 30,即评论领域的数量. 利用频繁模式挖掘先验知识时,最小支持度设置为 $\min(5,0.4\times\#T)$[1],$\#T$ 表示一个聚类中事务的数量. 本文中的事务是指所有领域的主题数量.

表 1 30 个领域名称

手机	显示器	MP4	电话机	U 盘
数码相机	DVD 播放器	无线路由器	电机	收音机
平板电脑	蓝光/DVD 影碟机	耳机	遥控器	投影仪
移动电源	HIFI 音箱	话筒	摄像头	键盘
笔记本	硬盘	主板	打印机	移动电源
鼠标	网络播放器	扩音器	相机电池	GPS 导航仪

为了比较基线方法与 TMPP 模型,实验只从 30 个评论语料中选取了笔记本,手机,数码相机,平板电脑和 MP4 这五种类型产品的评论语料作为测试语料集. 由于 3 个模型都以“短语袋”表示评论,所以首先要对这五类产品进行预处理. 预处理后得到的观点短语数量分别是:17540,7852,14320,4317,6591. 对于这三个模型,潜在变量的后验估计执行 2000 次迭代. 并且设置 $\alpha=1,\beta=0.1,T=15,\sigma=0.2$,对于每一个主题聚类,$\gamma$ 设置为这个聚类中词数量的比例.

6.2 引入先验知识后的方面识别分析

(1)主题连贯性评估

主题模型的评估常采用困惑度评价,但困惑度并不能反映语义连贯性. 近年来,主题连贯性度量已成为一个实际的评估标准[18,21]. 主题连贯性的评估值越高,意味主题的可解释性越好. 因而本文也采用主题连贯性度量来观察使用了先验知识的 TMPP 模型产生的方面在质量上是否优于两个基线方法. 针对 15 个主题分别计算了主题连贯性之后的平均值,其中 LDA 模型在评论语料上的运行作为初始的迭代(即第 0 次迭代).

图 2 给出了三个模型的在五类数据上运行后得到的主题连贯性评估曲线. 从图 2 能观察到:1)用领域先验知识指导的 TMPP 模型具有最高的主题连贯性. 这表明 TMPP 找到了最具有解释性的方面;2)ILDA 的主题连贯性优于 LDA,这说明尽管 ILDA 没有使用先验知识,但由于对方面及相应的等级之间的依赖进行了建模,所以可能更有利于发现语义上连贯的主题.

(2)人工评估

人工评估阶段让两位研究生充当专家角色,对三个模型在五个领域上产生的主题是否具有连贯性进行了手工标注. 如果专家一致认为大多数热点词是连贯的,且能表达现实世界,就将这个主题标记为连贯;否则,为不连贯. 对于一个连贯的主题,如果热点词反映了主题所表示的方面,就将其标注为正确;否则,为不正确.

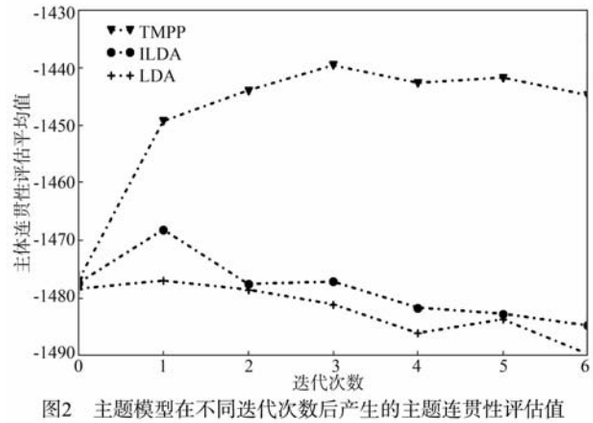

图2 主题模型在不同迭代次数后产生的主题连贯性评估值

实验采用 Precision@n 度量人工评估结果. 图 3 给出 $n=5$ 和 $n=10$ 的 Precision@n 值. 从图 3 可观察到,相比于基线方法,TMPP 在 5 个领域上都有改进. 改进最大的是数码相机领域,最小的是 MP4 领域. 这是因为先验知识中有较多的方面与数码相机领域的方面有重叠,而与 MP4 领域的方面重叠较少,即如果一个领域与许多其它领域共享了方面,那么利用先验知识就能较大程度地改进主题模型产生的主题质量;否则,改进较小.

表 2 以主题质量改进最大的数码相机领域和改进最小的 MP4 领域的评论语料为例,列出了由 TMPP 和 2 个基线模型产生的方面样例“电池”和该方面的前 10 个热点词. 从表 2 可知,TMPP 发现了更多正确的和有意义的热点方面词. 表 2 中用粗黑体标注的词是方面样例“电池”中不符合语义的词.

表 2 三个模型在数码相机领域和 MP4 领域产生的方面样例“电池”

数码相机			MP4		
TMPP	ILDA	LDA	TMPP	ILDA	LDA
电池	电池	电池	小时	电池	物流
充电	长	功能	时间	长	电池
分钟	性价比	声音	电池	小时	小时
小时	充电	小时	充上电	喜欢	价位
正品	质量	连接	充电	便宜	充电头
长	正品	充电	长	充电	精致
好	小时	精致	充电头	方便	内存
原装	时间	性价比	小巧	发货	宝贝
宝贝	读卡器	非常	物流	外音	价格
短	接口	Wifi 卡	满意	好	满意

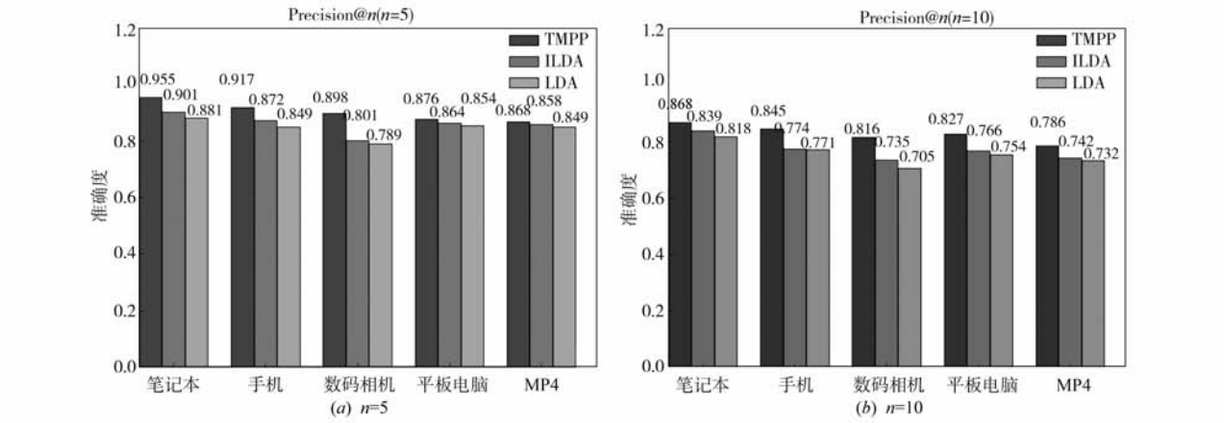

图3　三个模型在五个领域上平均的Precision@n(n=5,10)值

6.3　等级预测分析

(1)等级预测精度分析

采用公式(7)所示的聚类相似度衡量等级预测精度[4]. 对三个模型,k 值固定为 5,表示方面等级的聚类数量;P_i 表示主题模型 i 产生的等级预测;P_m 表示人工标注产生的等级预测. P_i 与 P_m 的一致性要在 $k\times(k-1)$ 个等级预测对上进行检验. 对每两个等级预测对,P_i 和 P_m 可能把它分配到相同的聚类或不同的聚类. 因此,公式(7)中的 x 表示在两个划分中,属于相同聚类的等级预测对数量;y 表示在两个划分中,属于不同聚类的等级预测对数量.

$$\text{RandIndex}(P_i,P_m)=\frac{2(x+y)}{k\times(k-1)} \tag{7}$$

TMPP 模型具有将评论的总体等级分解为单个等级的功能,所以相同聚类的等级预测对数量 x 与不同聚类的等级预测对数量 y 均高;*LDA* 模型的总体等级不能分解为单个等级,所以相同聚类的等级预测对数量 x 与不同聚类的等级预测对数量 y 均低;尽管 ILDA 模型也能将评论的总体等级分解为单个等级,但由于没有使用先验知识,只利用了方面与等级之间的依赖关系,所以相同聚类的等级预测对数量 x 与不同聚类的等级预测对数量 y 介于 TMPP 和 LDA 之间. 对三个模型输出的每一个词聚类,人工标注 P_m 就是从词聚类中找出所有的形容词,根据褒贬形容词的个数确定方面的情感等级. 褒义词越多,情感等级就越高. 实验假定一个方面所在词聚类中有 3 个及以上褒义词就认为情感等级较高. 针对以上五类产品,分别计算了三种不同模型的 RandIndex 值,并列于表 3 中.

从表 3 可知:TMPP 的等级预测相似度值最高,这说明在语义更连贯的方面中,属于相同方面聚类的等级预测对数量较多. ILDA 模型的等级预测相似度值比 LDA 高,原因是 ILDA 模型中方面和等级之间的潜在语义关联建模也有利于等级预测. 此外,所有模型在数码相机这类产品数据集上等级预测精度最好,这是因为一方面先验知识中有较多的方面与数码相机的方面重叠,另一方面这类产品的训练数据集最大.

表 3　等级预测的聚类相似度 RandIndex 值

产品类型	TMPP 模型	ILDA 模型	LDA 模型
笔记本电脑	0.70	0.51	0.43
手机	0.75	0.60	0.50
数码相机	0.78	0.57	0.55
平板电脑	0.72	0.53	0.41
MP4	0.75	0.46	0.48

(2)不同品牌数码相机的方面抽取和等级预测

表 4 和表 5 是不同品牌数码相机 Canon/佳能 PowerShot A23 和 Sony/索尼 DSC-W690 的评级方面总结示例. TMPP 模型能依据先验知识产生语义上更连贯的方面,根据这些方面把评论的总体等级分解为单个等级,以便用户能获得目标产品的不同视角. 尽管两个不同品牌的数码相机有相同的总体等级 3,但是 Canon/佳能 PowerShot A23 有更好的“放大”,而 Sony/索尼 DSC-W690 有更好的“屏幕”和“声音”,为用户提供更详细的信息,相比于产品的总体等级,这种方式有助于用户做出购买决定.

表 4　Canon/佳能 PowerShot A23

抽取的方面	对应等级
镜头焦距	5
价格	4
相片质量	4
电池寿命	2
屏幕	1
总体等级	3

表 5 Sony/索尼 DSC-W690 的方面和相应等级的方面和相应选级

抽取的方面	对应等级
镜头焦距	2
价格	4
相片质量	4
电池寿命	3
屏幕	2
总体等级	3

(3)五类产品的相关方面及对应的正情感词

通过设置不同的主题数目和 10 次交叉验证,按概率从高到低列出了五类产品中排名前八位的相关方面和与该主题相关的概率最高的正情感词,如表 6 所示. 从表 6 中相关方面所对应的正情感词分析得知,淘宝网的用户对所评价对象的某一个方面有较好的购买体验时,使用频率最高的正的情感词分别是"好"、"高"等最简单,最常用的形容词. 这形成了淘宝在线产品评论的一个显著特点. 该特点为在线评论中情感词的抽取研究提供了一定的事实依据[22].

表 6 产品的相关方面及对应的正情感词

产品类型	相关方面	正情感词	相关方面	正情感词
笔记本电脑	包装	不错	预装系统	方便
	性价比	高	散热	好
	性能	好	音质	好
	外观	酷	屏幕	舒服
手机	性能	不错	外观	精美
	待机时间	长	显示屏	好
	上网功能	强大	音乐功能	强大
	通话质量	好	软件支持	好
数码相机	像素	高	画质	精美
	价格	便宜	颜色	漂亮
	质量	好	携带	方便
	款式	好	服务态度	好
平板电脑	质量	好	携带	方便
	性能	好	样式	好
	性价比	高	屏幕	好
	是否正品	正品	到货速度	快
MP4	音效	好	画质	好
	外观	好看	包装	不错
	性价比	高	整体感觉	不错
	实用性	好	到货速度	快

7 结论

Web 技术的发展使在线评论成为决策支持的有价值资源. 本文提出基于短语参数学习的主题模型 TMPP,它能同时抽取在线评论中被评价实体的 aspect 和其对应的 rating. 此外,TMPP 还整合了一个潜在的聚类变量,用于指导产生更连贯的方面. 聚类变量表示从大量已知领域中学习到的知识. 这种知识是通过在已知评论语料上执行标准的 LDA 模型后,对其产生的主题进行分类,然后通过频繁模式挖掘得到的. 在实际的评论语料上比较了提出的 TMPP 和基线模型,TMPP 模型产生的方面质量高. 通过等级预测的聚类相似度度量标准,发现 TMPP 模型方面的等级预测也优于基线模型.

参考文献

[1] Zhiyuan Chen, Arjun Mukherjee, Bing Liu. Aspect Extraction with automated prior knowledge learning[A]. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics[C]. Baltimore: ACL, 2014. 347 – 358.

[2] Hu Minqing, Bing Liu. Mining and summarizing customer reviews[A]. Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining[C]. Seattle: ACM, 2004. 168 – 177.

[3] 吕品,钟珞,蔡敦波,吴云韬. 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,2014,36(2):359 – 366.
LÜ Pin, ZHONG Luo, CAI Dun-bo, WU Yun-tao. Effective mining product featuers from Chinese review based on CRF [J]. Computer Engineering & Science, 2014, 36 (2): 359 – 366. (in Chinese)

[4] Samaneh Moghaddam, Martin Ester. ILDA: Inerdependent LDA model for learning latent aspects and their ratings from online product reviews[A]. Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011[C]. Beijing, China: ACM, 2011. 665 – 674.

[5] Yue Lu, ChengXiang Zhai, Neel Sundaresan. Rated aspect summarization of short comments[A]. Proceedings of the 18th International Conference on World Wide Web [C]. Madrid: ACM, 2009. 131 – 140.

[6] Liu Bing, Minqing Hu, Junsheng Cheng. Opinion observer: Analyzing and comparing opinions on the web[A]. Proceedings of the 14th International Conference on World Wide Web[C]. Chiba: ACM, 2005. 342 – 351.

[7] Moghaddam Samaneh, Martin Ester. Opinion digger: an unsupervised opinion miner from unstructured product reviews [A]. Proceedings of the 19th ACM Conference on Information and Knowledge Management[C]. Toronto: ACM, 2010. 1825 – 1828.

[8] Ana-Maria Popescu, Oren Etzioni. Extracting product features and opinions from reviews[A]. Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing[C]. Vancouver: ACL, 2005. 339 – 346.

[9] Guo Honglei, Huijia Zhu, Zhili Guo, Xiaoxun Zhang, Zhong Su. Product feature categorization with multilevel latent semantic association[A]. Proceedings of ACM International Conference on Information and Knowledge Management[C]. HongKong: ACM, 2009. 1087 – 1096.

[10] Titov Ivan, Ryan McDonald. Modeling online reviews with multi-grain topic models[A]. Proceedings of the 17th International Conference on World Wide Web[C]. Beijing:ACM,2008. 111 – 120.

[11] Hongning Wang, Yue Lu, Chengxiang Zhai. Latent aspect rating analysis on review text data:a rating regression approach[A]. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. Washington:ACM,2010. 783 – 792.

[12] Wong Tak-Lam, Wai Lam, Tik-Shun Wong. An unsupervised framework for extracting and normalizing product attributes from multiple web sites[A]. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Singapore:ACM,2008. 35 – 42.

[13] Titov Ivan, R. McDonald. A joint model of text and aspect ratings for sentiment summarization[A]. The 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies[C]. Columbus: ACL, 2008. 308 – 316.

[14] Mei Qiaozhu, Xu Ling, Matthew Wondra, et al. Topic sentiment mixture: modeling facets and opinions in weblogs [A]. Proceedings of the 16th International Conference on World Wide Web[C]. Banff:ACM,2007. 171 – 180.

[15] Blei David M , Andrew Y Ng, Michael I Jordan. Latent dirichlet allocation[J]. The Journal of Machine Learning Research,2003(3):993 – 1022.

[16] Jiawei Han, Hong Cheng, Dong Xin, Xifeng Yan. Frequent pattern mining: current status and future directions [J]. Data Mining and Knowledge Discovery,2007,15(1):55 – 86.

[17] Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, Mark Steyvers. Learning author-topic models from text corpora[J]. ACM Transactions on Information Systems,2010,28(1):1 – 38.

[18] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, Andrew McCallum. Optimizing semantic coherence in topic models[A]. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing[C]. Edinburgh:ACL,2011. 262 – 272.

[19] Hosam Mahmoud. Polya Urn Models. Chapman & Hall/ CRC Texts in Statistical Science[M]. USA: CRC Press,2008.

[20] Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, Riddhiman Ghosh. Exploiting domain knowledge in aspect extraction[A]. Proceedings of EMNLP[C]. Seattle:ACL,2013. 1655 – 1667.

[21] Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, Michael Zhu. A practical algorithm for topic modeling with provable guarantees[A]. Proceedings of the 30th International Conference on Machine Learning[C]. Atlanta: JMLR, 2013. 280 – 288.

[22] 吕品,钟珞,唐琨皓. 在线产品评论用户满意度综合评价研究[J]. 电子学报,2014,42(4):740 – 745.
LÜ Pin, ZHONG Luo, TANG Kun-hao. Customer satisfaction degree evaluation of online product review[J]. Acta Electronica Sinica,2014,42(4):740 – 745. (in Chinese)

作者简介

吕　品　女,1973 年 3 月出生,湖北鄂州人,现为上海电机学院副教授、博士,研究方向为数据挖掘、观点挖掘与情感分析.
E-mail:lvp@ sdju. edu. cn

汪　鑫　男,1978 年 3 月出生,安徽黟县人,现为上海电机学院讲师、硕士,研究方向为数据挖掘、云计算.
E-mail:wangx@ sdju. edu. cn

罗宜元　男,1986 年 9 月出生,河南信阳人,现为上海电机学院讲师、博士,研究方向为密码学与计算机安全.
E-mail:luoyy@ sdju. edu. cn

计春雷　男,1964 年 1 月出生,上海人,现为上海电机学院教授、博士、硕士生导师,研究方向为大数据、数据挖掘.
E-mail:jicl@ sdju. edu. cn