北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis

基于主题情感混合模型的无监督文本情感分析

孙艳　周学广†　付伟

海军工程大学信息安全系，武汉　430033; † 通信作者, E-mail: zxg196610@hotmail.com

摘要　有监督、半监督的文本情感分析存在标注样本不容易获取的问题。为此，通过在 LDA 模型中融入情感模型，提出一种无监督的主题情感混合模型(UTSU)。UTSU 模型对每个句子采样情感标签，对每个词采样主题标签，无须对样本进行标注，就可以得到各个主题的主题情感词，从而对文档集进行情感分类。情感分类实验对比表明 UTSU 模型的分类性能比有监督情感分类方法稍差，但在无监督的情感分类方法中效果最好，情感分类综合指标比 ASUM 模型提高了约 2%，比 JST 模型提高了约 16%。

关键词　主题模型; LDA; 情感分析; 混合模型

中图分类号　TP391

Unsupervised Topic and Sentiment Unification Model for Sentiment Analysis

SUN Yan, ZHOU Xueguang†, FU Wei

Deparment of Information Security, Naval University of Engineering, Wuhan 430033;
† Corresponding author, E-mail: zxg196610@hotmail.com

Abstract　Supervised and semi-supervised sentiment classification methods need label corpora for classifier training. To solve this problem, an unsupervised topic and sentiment unification model (UTSU model) is proposed based on the LDA model. UTSU model imposes a constraint that all words in a sentence are generated from one sentiment and each word is generated from one topic. This constraint conforms to the sentiment expression of language and will not limit the topic relation of words. UTSU model is compeletly unsupervised and it needs neither labeled corpora nor sentiment seed words. The experiments of sentiment classification show that UTSU model comes close to supervised classification methods and outperforms other topic and sentiment unification models. UTSU model improves the F_1 value of sentiment classification 2% than ASUM model and 16% than JST model.

Key words　topic model; latent Dirichlet allocation(LDA); sentiment analysis; unification model

现代信息技术赋予了传统社会经济活动前所未有的社会化、网络化内涵，极大地提高了效能。越来越多的用户乐于在互联网上分享自己对于某事件、产品等的观点或体验，这类评论信息迅速膨胀，仅靠人工的方法难以应对网上海量信息的收集和处理。如何有效地管理和使用这些评价信息成为当前的迫切需求，这促进了自动文本情感分析技术的发展[1−2]。

情感分析中的两个重要任务是情感信息抽取和情感信息分类，目前主要有基于规则和基于统计两种方法。新词的不断出现、表达方式的变化以及复杂的语言处理都使得基于规则的情感分析方法难以适用。

机器学习方法和文本表示模型是基于统计的情感分析方法的两个核心内容。机器学习方法包括有监督、半监督和无监督情感分析。有监督和半监督

国家自然科学基金(61100042)资助
收稿日期: 2012−06−02; 修回日期: 2012−08−20; 网络出版时间: 2012-10-26 17:55
网络出版地址: http://www.cnki.net/kcms/detail/11.2442.N.20121026.1755.022.html

的机器学习方法中分类器的训练需要一定数量经过标注的训练样本, 然而人工标注过程相对耗时费力, 成本昂贵, 无监督的机器学习则无需标注的训练样本。

长期以来文本表示的主要方法是向量空间模型(vector space model, VSM)。VSM 认为文档都是在词典空间中进行表示的, 即一个文档是一个一对多的映射, 表示为文档→词。随着人们对文本认识的发展, 发现向量空间模型没有考虑词的同义和多义情况, 忽视了词与词之间的语义联系。为挖掘文本的潜在语义, 人们开始寻找更能表示文本语义的文本表示模型。潜在语义分析(latent semantic analysis, LSA)就是一种能探查词与词之间内在语义联系的方法, 打破了文档都是在词典空间进行表示的思维定式, 在文本和词之间加入了一个语义维度, 采用线性代数的方法提取语义维度。随着概率统计分析的发展, 基于概率统计分析模式逐渐取代了基于线性代数的分析模式。概率潜在语义分析(probabilistic latent semantic analysis, pLSA)就是 LSA 的概率拓展, 它比 LSA 具有更坚实的数学基础。但是 pLSA 模型中的参数随着文本集的增长而线性增长, 容易出现过拟合情况, 且模型中的文档概率值与特定的文档相关, 没有提供文档的生成模型, 对于训练集外的文本无法分配概率。pLSA 存在的问题促发了人们寻找更好的主题模型, 2003 年, Blei 等[3]在 pLSA 的基础上提出潜在狄里克雷分配(latent Dirichlet allocation, LDA)模型。LDA 模型是一个完全的生成模型, 具有良好的数学基础和灵活拓展性, 已经应用到文本分析的很多领域中。

本文结合了无监督机器学习和 LDA 主题模型的优点, 提出了一个无监督的主题情感混合模型(unsupervised topic and sentiment unification model, UTSU), 通过对每个句子采样情感标签, 对每个词采样主题标签, 解决了文本主题发现和主题情感分类问题。

1 相关工作

LDA 模型是全概率生成模型, 参数空间的规模与文档数量无关, 适合处理大规模语料库。目前已有研究将 LDA 模型应用到情感分析中。

Titov 等[4]提出了一个多粒度 LDA 模型(multi-grain LDA, MG-LDA), 并应用于基于主题的情感摘要生成中, 提出多主题情感模型(multi-aspect sentiment model, MAS)[5]。虽然 Titov 等用实验证明了 MG-LDA 模型对于提取细粒度的主题有很好的效果, 但是 MG-LDA 需要对已标注好训练集进行训练, 属于有监督学习, 具有样本不容易获取和领域移植性差的缺点。同样需要监督学习的还有文献[6]提出的 ME-LDA 模型(MaxEnt-LDA), 该模型结合了最大熵组件和主题模型, 需要监督学习。

为使主题模型既能获得细粒度的主题又保持无监督学习的特征, 很多学者对主题模型进行了改进。Brody 等[7]直接将句子作为一个文档, 建立“句子-主题-词”关系。这种方法将 LDA 模型没有考虑文档和文档之间的关系进一步扩大, 没有考虑句子和句子之间的关系, 事实上在不同的句子中同一个主题可以有着完全不同的词。而且该方法只对主题词进行了情感词识别, 并没有得到文档或句子的情感分布, 即没有建立情感模型。Jo 等[8]认为一个句子中的所有词都由同一个主题和同一个情感产生, 因此采样主题标签时, 对每个句子采样主题标签, 而不是对每个词采样主题标签, 建立“文档-主题-句子”关系, 这种方法硬性地缩小了词之间的主题联系。

主题情感混合模型在语言模型上有两种表示方法。第一种是将主题和情感描绘成一个单一的语言模型, 在模型中一个词可能同时与主题和情感都相关, 如文献[8]提出的 ASUM 模型(aspect and sentiment unification model)和文献[9]提出的 JST 模型(joint sentiment/topic model)。另一种是将情感与主题作为分开的语言模型, 一个词要么是情感词要么是主题词, 只能二选一, 如文献[10]提出的 TSM 模型(topic sentiment mixture)。TSM 模型将词分为主题词和情感词, 认为情感词对主题发现没有作用, 而事实上情感词是表示主题的重要词汇, 应该是主题词的一部分。

本文提出的 UTSU 模型中的每个词都与主题和情感相关, 这一点是与 TSM 模型最大的区别。文献[7]只对主题词进行了情感词识别, 并没有得到文档或句子的情感分布, 即没有建立情感模型, 而本文的 UTSU 模型是一个主题情感混合模型。ASUM 模型采样主题标签和情感标签时, 对每个句子进行采样, 而不是对每个词采样, 而 JST 模型是对每个词进行采样主题标签和情感标签。本文的 UTSU 模型对每个句子采样情感标签, 对每个词采样主题标签, 这种采样方式即符合语言的情感表达, 又不会

缩小词之间的主题联系。

2 UTSU 模型

2.1 UTSU 模型的生成过程

UTSU模型是在LDA模型的基础上添加了情感模型而构建的。由于自然语言中的情感都是以句子为单位进行表达的(转折句除外), UTSU 模型假设一个句子的所有词由一种情感产生, 故对句子进行情感标签采样, 建立“文档-情感-句子”关系。沿用LDA 模型中每个词有不同的主题, 对每个词进行主题标签采样, 建立“文档-主题-词”关系。在运行UTSU 模型前, 先对文本进行预处理, 将转折句从转折处分为两句。

UTSU 模型的框图如图 1 所示。图中符号说明如表 1 所示。

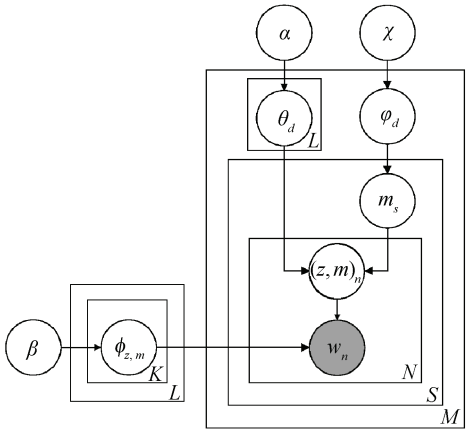

图 1 UTSU 模型框图

Fig. 1 Graphical model for UTSU

表 1 UTSU 模型符号说明

Table 1 Meanings of the notations for UTSU mode

符号	说明	符号	说明
θ_d	文档 d 的主题分布	α	主题分布的 Dir 参数
φ_d	句子 s 的情感分布	χ	情感分布的 Dir 参数
$\phi_{z,m}$	主题情感对 (z,m) 的单词分布	β	单词分布的 Dir 参数
m_s	句子 s 的情感	M	文档数
$(z,m)_n$	单词的主题和情感	S	文档的句子数
w_n	句子中的单词	N	句子的词汇数
K	主题数	L	情感数

UTSU 模型是一个 4 层盘子模型, 其产生过程的伪代码描述如下所示。

Name: Generative model for UTSU
//“topic plate”
for all topics $k \in [1,\cdots,K]$ and sentiments $j \in [1,\cdots,L]$ do
  sample mixture components $\phi_{z,m} \sim \mathrm{Dir}(\beta)$
end for
//“document plate”
for all documents $d \in [1,\cdots,M]$ do
  for all sentiments $j \in [1,\cdots,L]$ do
    sample mixture proportion $\theta_{dj} \sim \mathrm{Dir}(\alpha)$
  end for
  sample mixture proportion $\varphi_d \sim \mathrm{Dir}(\chi)$
  //“sentence plate”
  for all sentences $s \in [1,\cdots,\mathrm{Nds}]$ in document d do
    sample sentiment index $m_s \sim \mathrm{Multi}(\varphi_d)$
    //“word plate”
    for all words $n \in [1,\mathrm{Ns}]$ in sentence s do
      sample topic index $z_{s,n} \sim \mathrm{Multi}(\theta_{dm})$
      sample term for word $w_n \sim \mathrm{Multi}(\phi_{(z,m)_{s,n}})$
    end for
  end for
end for

给定所有参数, UTSU 模型所有潜在变量和可观察变量的联合概率为

$$p(\phi_{(1,1):(K,L)},\theta_{1:M},\varphi_{1:M},(z,m)_{1:M},w_{1:M}) = \underbrace{\prod_{1,1}^{K,j} p(\phi_{(i,j)}|\beta)}_{\text{主题-情感层}} \underbrace{\prod_{d=1}^{M} p(\theta_d|\alpha)p(\varphi_d|\chi)\left(\prod_{s=1}^{\mathrm{Nds}} p(m_s|\varphi_d)\left(\underbrace{\prod_{n=1}^{Ns} p(w_{s,n}|\phi_{(z,m)_{s,n}})p\left((z,m)_{s,n}|m_s,\theta_d\right)}_{\text{词层}}\right)\right)}_{\text{句层}} \quad (1)$$

文档层

其中 Nds 表示文档 d 内的句子数, Ns 表示句子 s 内的词数。

2.2 UTSU 模型求解

用 i 来表示词汇记号的索引号, $i=(d,s,n)$, 词汇记号 $w_i = w_{d,s,n}$ 表示与文档位置、句子位置相关的词汇, s_i 表示词汇记号 w_i 所在的句子。m_{s_i} 表示词汇记号 w_i 所在句子的情感分配, m_{-s_i} 表示除当前词汇记号所在句子外其他词汇记号所在句子的情感分配。z_i 表示词汇记号 w_i 的主题分配, z_{-i} 表示除当前词汇记号外的其他所有词汇记号的主题分配。有

$w=\{w_i=t,w_{-i}\}$, $z=\{z_i=k,z_{-i}\}$, $m=\{m_{s_i}=j,\ m_{-s_i}\}$。利用 Gibbs 采样算法进行采样，当前词汇记号 w_i 的主题为 k，情感为 j 的概率可通过式(2)得到。

$$
\begin{aligned}
&p(z_i,m_{s_i}\,|\,z_{-i},m_{-s_i},w)\\
&=\frac{p(z,m,w)}{p(z_{-i},m_{-s_i},w)}\\
&=\frac{p(w|z,m)p(z,m)}{p(w_{-i}\,|\,z_{-i},m_{-s_i})p(w_i)p(z_{-i},m_{-s_i})}\\
&\propto\frac{B(n_{k,j}+\beta)}{B(n_{k,j,-i}+\beta)}\frac{B(n_d+\chi)}{B(n_{d,-s_i}+\chi)}\frac{B(n_{d,j}+\alpha)}{B(n_{d,j,-i}+\alpha)}\\
&\propto\frac{\Gamma(n_{k,j}^{(t)}+\beta_t)}{\Gamma\left(\sum_{t=1}^{V}(n_{k,j}^{(t)}+\beta_t)\right)}\frac{\Gamma(n_d^{(j)}+\chi_j)}{\Gamma\left(\sum_{j=1}^{L}(n_d^{(j)}+\chi_j)\right)}\\
&\cdot\frac{\Gamma(n_{d,j}^{(k)}+\alpha_k)}{\Gamma\left(\sum_{k=1}^{K}(n_{d,j}^{(k)}+\alpha_k)\right)}\frac{\Gamma\left(\sum_{t=1}^{V}(n_{k,j,-i}^{(t)}+\beta_t)\right)}{\Gamma(n_{k,j,-i}^{(t)}+\beta_t)}\\
&\cdot\frac{\Gamma\left(\sum_{j=1}^{L}(n_{d,-s_i}^{(j)}+\chi_j)\right)}{\Gamma(n_{d,-s_i}^{(j)}+\chi_j)}\frac{\Gamma\left(\sum_{k=1}^{K}(n_{d,j,-i}^{(k)}+\alpha_k)\right)}{\Gamma(n_{d,j,-i}^{(k)}+\alpha_k)},
\end{aligned}
\tag{2}
$$

其中 $B(\alpha)$ 是 Beta 函数，$B(\alpha)=\dfrac{\prod_{i=1}^{k}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k}\alpha_i)}$，$\Gamma$ 为 Gamma 函数，$\Gamma(x)=\int_0^{\infty}\mathrm{e}^{-t}t^{x-1}\mathrm{d}t$。

因为 $\Gamma(x+1)=x\Gamma(x),\ x>0,$

所以

$$
\begin{aligned}
p(z_i=k,m_{s_i}=j\,|\,z_{-i},m_{-s_i},w)\propto&\frac{n_{k,j,-i}^{(t)}+\beta_t}{\sum_{t=1}^{V}(n_{k,j,-i}^{(t)}+\beta_t)}\\
&\frac{n_{d,-s_i}^{(j)}+\chi_j}{\sum_{j=1}^{L}(n_{d,-s_i}^{(j)}+\chi_j)}\frac{n_{d,j,-i}^{(k)}+\alpha_k}{\sum_{k=1}^{K}(n_{d,j,-i}^{(k)}+\alpha_k)}。
\end{aligned}
\tag{3}
$$

$n_{k,j,-i}^{(t)}$ 表示除当前词汇记号外，其他与 w_i 内容相同的词 w 的主题和情感分别为 k 和 j 上的词汇记号个数，$n_{d,-s_i}^{(j)}$ 表示除当前词汇记号所在句子外文档 d 中情感为 j 的句子数，$n_{d,j,-i}^{(k)}$ 表示除当前词汇记号外，文档 d 中情感 j 主题为 k 的词汇记号数。

从式(3)可以看出，词汇记号 w_i 的情感 $m_{s_i}=j$，主题 $z_i=k$ 的条件概率由 3 部分组成，左半部分对应着 w_i 的主题为 k 情感为 j 的概率，中间部分对应着情感 j 在文档 d 的情感分布出现的概率，右半部分对应着在文档 d 的当前主题分布中，情感为 j 主题为 k 出现的概率。在整个文档集中，如果一个单词的很多词汇记号分配在主题和情感分别为 z 和 j 上，那么这个单词的其他任何一个词汇记号分配在主题和情感分别为 z 和 j 上的概率就会增加。如果情感 j 在同一文档中多次出现，那么在该文档中出现的任何句子分配给情感 j 的概率也会增加。同理，如果主题 k 在同一文档中多次出现，那么在该文档中出现的任何单词分配给主题 k 的概率也会增加。

舍弃词汇记号，用 w 表示唯一性词，θ、φ 和 ϕ 的估计如下:

$$
\begin{aligned}
\hat{\theta}_{d,j,k}&=\frac{n_{d,j}^{(k)}+\alpha_k}{\sum_{k=1}^{K}(n_{d,j}^{(k)}+\alpha_k)},\\
\hat{\varphi}_{d,j}&=\frac{n_d^{(j)}+\chi_j}{\sum_{j=1}^{L}(n_d^{(j)}+\chi_j)},\\
\hat{\phi}_{k,j,w}&=\frac{n_{k,j}^{(w)}+\beta_t}{\sum_{t=1}^{V}(n_{k,j}^{(w)}+\beta_t)},
\end{aligned}
\tag{4}
$$

$\hat{\theta}_{d,j,k}$ 表示文档 d 的当前主题分布中，主题为 k 情感为 j 出现的概率，$\hat{\varphi}_{d,j}$ 表示情感 j 在文档 d 的情感分布中的概率估计，$\hat{\phi}_{k,j,w}$ 表示词汇 w 分配在主题和情感分别为 z 和 j 上的概率估计，$n_d^{(j)}$ 表示文档 d 中分配在情感 j 上的句子数，$n_{d,j}^{(k)}$ 表示文档 d 中分配在主题为 k 情感为 j 上的词数，$n_{k,j}^{(w)}$ 表示 w 分配在主题为 k 情感为 j 上的次数。

3 实验结果与分析

3.1 实验数据集

从大众点评网上下载关于快递、烧烤的评论网页，下载中国科学院谭松波博士公布的关于酒店和计算机的情感分类数据集，整理共得到 9180 个文本。正类(Pos)文本都是从三星级以上评论中整理得到的，负类(Neg)文本都是从三星级以下评论中整理得到的。每种数据集的大小和正负情感分布如表 2 所示。

表 2 数据集

Table 2 Dataset

极性	快递 (Corp1)	烧烤 (Corp2)	酒店 (Corp3)	计算机 (Corp4)
Pos	1150	910	1130	1270
Neg	1140	1230	1200	1150

预处理数据集: 1) 对含有“但”、“但是”、“可是”等转折词的句子进行切分, 从转折处将句子分为两句; 2) 统计实验所需的文档–词共现信息, 其中中文分词采用中国科学院的 ICTCLAS 开源工具包, 统计时剔除停用词, 但是保留“不”、“没”、“都”等对情感判断产生影响的词。

3.2 主题–情感词发现

本文实验的情感只考虑褒义和贬义两种, 不考虑中性情感。利用 UTSU 模型进行主题情感发现, 参数设置如下: $\alpha=1$, $\chi=1$, $\beta=0.01$, $L=2$, 以上参数均为经验最优值, 主题数 $K=4$, 迭代次数 $N=1000$, 得到的主题–情感词按照在文档集中的概率大小, 排列如表 3 所示。限于空间, 只列出了前 39 个关于计算机的主题–情感发现词。

从表3中可以看出, 正负情感词在主题–情感发现中分的比较明显, 如左边表示贬义的情感词“郁闷、慢、重、一般”等, 右边表示褒义的情感词有“不错、漂亮、小巧、喜欢、舒服、精致”等。形容词“大”同时出现在两边靠前的位置, 这是由于“大”可以表达褒义也可以表达贬义, 如“电脑轻巧, 电池强劲, 键盘尺寸够大”和“钢琴烤漆很容易留指纹印, 并且进入不了系统, 开机时声音很大”中的“大”的情感完全相反。

通过对主题–情感词进行分析, 发现有很多无主题无情感的单字高频词, 这种词可以看作是情感分类中的噪声干扰, 在此称作噪声词, 其会对后续情感分类产生影响, 表4为去噪后获得的主题–情感词汇表。

表 3 主题情感发现词汇表

Table 3 Example topic-sentiment words discorved by UTSU

分类	计算机(−)			计算机(+)		
词汇	不	京东	卡	不错	带	起来
	系统	高	开机	键盘	上网	价格
	装	重	显示	电池	方便	功能
	有点	驱动	机	没有	性价	散热
	买	手	内存	小时	使用	手感
	大	郁闷	坏	做工	性能	喜欢
	散热	笔记本	太	机器	长	开
	屏幕	运行	换	漂亮	单	本本
	时	声音	胆包	本	实	硬盘
	面	触摸板	行	速度	小巧	强
	感	慢	需要	大	时间	摄像头
	问题	分	鼠标	配置	比较	白色
	一般	分区	不能	外观	舒服	精致

表 4 去噪后的主题情感发现词汇表

Table 4 Example topic-sentiment words discorved by UTSU after denoising

分类	计算机(−)			计算机(+)		
词汇	不	驱动	内存	不错	上网	价格
	系统	郁闷	坏	键盘	方便	功能
	有点	笔记本	胆包	电池	性价	散热
	大	运行	需要	没有	使用	手感
	散热	声音	鼠标	小时	性能	喜欢
	屏幕	触摸板	不能	做工	长	本本
	感	慢	速度	机器	实	硬盘
	问题	分区	拒绝	漂亮	小巧	强
	一般	卡	存储	速度	时间	摄像头
	京东	开机	响应	大	比较	白色
	高	显示	风扇	配置	舒服	精致
	重	反应	收到	外观	起来	快速
	垃圾	死机	付款	细致	接口	优雅

3.3 情感分类

利用 UTSU 模型的 $\hat{\phi}_{d,j}$ 可以得到情感 j 在文档 d 的情感分布中的概率估计, 取每种情感在文档 d 的情感分布中的概率估计的最大值可得到文档 d 的情感, 即 $m_d=\arg\max_j\left\{\hat{\varphi}_{d,j}\,\middle|\,j\in[1,\cdots,L]\right\}$ 作为文档 d 的情感。

将本文提出的 UTSU 模型与 ASUM 模型、JST 模型和 Pang 等[11]的方法进行了对比。ASUM 模型和 JST 模型的原文中都用到了种子情感词作为先验知识。由于种子情感词的不同对结果影响较大, 本文统一采用无先验知识。Pang 方法中使用信息增益选取了2000个特征, 分类器采用SVM, 分类时采用 10 重交叉验证。各种方法在 4 个数据集上的情感分类正确率(precision)、召回率(recall)值和 F 综合指标如图 2 所示。

4 种方法中 Pang 方法是有监督的学习方法, 其他 3 种都是无监督的主题情感混合模型。从图 2 中可以看出, 综合考虑准确率和召回率, 效果最好的是 Pang 方法。但由于 Pang 方法是基于向量空间模型的有监督学习方法, 需要先对标注好的样本进行训练才能测试。其他 3 种主题情感混合模型中, 效果最好的是 UTSU 模型, 其情感分类在 4 个数据集上综合指标平均值比 ASUM 模型高约 2%, 比 JST 模型高约 16%。这也证明了本文提出的对每个句子

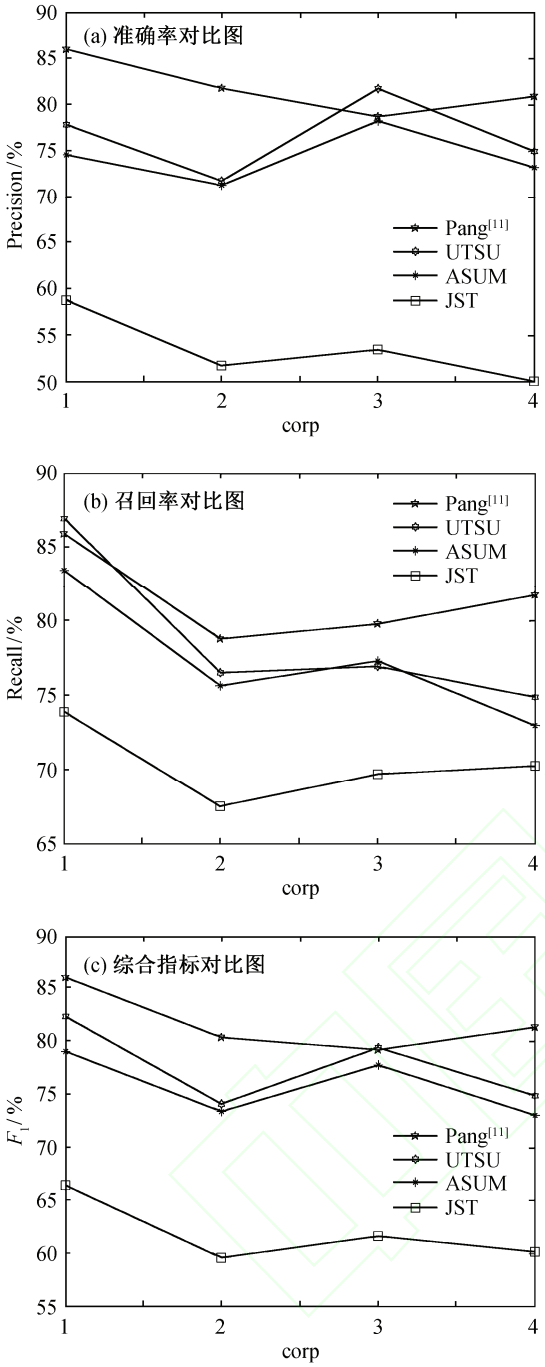

图 2　情感分类效果对比图

Fig. 2　Sentiment classification performances

采样情感标签，对每个词采样主题标签的主题情感混合模型在情感分类上的有效性。由于 JST 模型每次采样情感标签时，对每个词进行采样，不符合自然语言的情感表达，故其情感分类效果最差，这也是 JST 模型与 UTSU 模型和 ASUM 模型最大的区别。

从图 2 中可以看出，4 种方法在 4 个数据集上的情感分类准确率、召回率不同。根据综合指标在不同数据集上从高到低进行排序，依次为：快递>酒店>计算机>烧烤。通过对数据进行分析，我们得到了以下 4 点原因。

1) 用户对快递进行观点表述时，表达比较单一，基本上只有快递时间和服务态度。

2) 用户对酒店关注的主题比快递更为分散，包括床、房间、环境、位置、服务、价格等。

3) 由于计算机含有不同的型号，如联想、惠普等，不同的零件和属性，如屏幕、键盘、蓝牙、重量、电池等，以及计算机的效能，如散热、配置、无线信号等，使得计算机数据集的情感分类更难。

4) 烧烤类数据集涵盖的观点与酒店类很相似，包括口味、房间、环境、位置、服务、价格等，但由于许多观点针对的是不同的烧烤项目，如鸡翅、肉筋、肉串等，且对各种烧烤项目的评价不同，这使得以文本为单位进行情感分类时，对特征集的依赖性较强，这也是不同的方法在烧烤类数据集情感分类中性能最不稳定的原因。

总体来说本文构建的 UTSU 模型情感分类的性能比有监督情感分类方法稍差，但在无监督的情感分类方法中效果最好，比 ASUM 模型提高了 2%，比 JST 模型提高了 16%。

4　结语

本文重点从无监督机器学习和文本表示模型的角度对文本情感分类进行了研究。在 LDA 模型的基础上，提出无监督的主题情感混合模型 UTSU 模型。UTSU 模型对每个句子采样情感标签，对每个词采样主题标签，这种采样方式既符合语言的情感表达，又不会缩小词之间的主题联系，克服了 ASUM 模型和 JST 模型在主题标签和情感标签在同一层盘子中的缺陷。主题–情感词发现实验表明 UTSU 能够获得较准确的主题–情感词。情感分类实验对比表明 UTSU 模型的性能比其他混合模型的效果好，比有监督情感分类方法稍差，但在无监督的情感分类方法中效果最好，比 ASUM 模型提高了 2%，比 JST 模型提高了 16%。

本文只研究了主题–情感词的发现和情感分类，如何将 UTSU 模型应用到情感分析的其他任务中以及结合种子情感词提高情感分类精度是下一步研究方向。

参考文献

[1] Liu B, Zhang L. A survey on opinion mining and

sentiment analysis. Mining text data. New York: Springer, 2012: 415–463

[2] Taboada M, Brooke J, Tofiloski M, et al. Lexicon-based methods for sentiment analysis. Computational linguistics, 2011, 37(2): 267–307

[3] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of Machine Leanrning Research, 2003(3): 993–1022

[4] Titov I, McDonald R. Modeling online reviews with multi-grain topic models // Proceeding of WWW'08. New York: ACM, 2008: 111–120

[5] Titov I, McDonald R. A joint model of text and aspect ratings for sentiment summarization // Proceedings of ACL-08: HLT. Stroudsburg: ACL, 2008: 308–316

[6] Zhao X, Jiang J, Yan H F, et al. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2010: 56–65

[7] Brody S, Elhadad N. An unsupervised aspect-sentiment model for online reviews // Proceedings of the 2010 Annual Conference of the North American Chapter of the ACL. Stroudsburg: ACL, 2010: 804–812

[8] Jo Y, Oh A. Aspect and sentiment unification mode for online review analysis // Proceedings of the 4th ACM International conference on Web search and data mining. New York: ACM, 2011: 815–824

[9] Lin C H, He Y L. Joint sentiment/topic model for sentiment analysis // Proceeding of the 18th ACM conference on Information and knowledge management. New York: ACM, 2009: 375–384

[10] Mei Q Z, Ling X, Wondra M, et al. Topic sentiment mixture: modeling facets and opinions in weblogs // Proceeding of WWW'07. New York: ACM, 2007: 171–180

[11] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques // Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics .Philadelphia, PA, 2002: 79–86