无监督的主题情感混合模型研究

孙艳，周学广，付伟
（海军工程大学电子工程学院，430033，武汉）

摘要：提出了一种基于LDA-Col模型的无监督主题情感混合(UTSU)模型。采用词序流对文本进行表示，对每个句子采样情感标签，对每个词采样主题标签，得到文本的主题情感分布。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模型在同一层盘子中采样主题标签和情感标签的缺陷。实验表明，UTSU模型的情感分类性能比有监督的情感分类方法稍差，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3%，比JST模型提高了17%。

关键词：文本情感分类；无监督学习；混合模型
中图分类号：TP18 **文献标志码**：A **文章编号**：0253-987X(2013)01-0120-06

An Unsupervised Topic and Sentiment Unification Model

SUN Yan, ZHOU Xueguang, FU Wei
(Electronics Engineering College, Naval University of Engineering, Wuhan 430033, China)

Abstract: An unsupervised topic and sentiment unification model, UTSU model for short, is proposed based on the LDA-Col model. Unlike the ASUM model and the JST model that sample sentiments and topics from the same plate, the UTSU model imposes the constraint that all words in a sentence are generated from one sentiment and each word in the sentence is generated from one topic. The constraint accords with the sentiment expression of language and will not limit the topic relation of words. The experiments of sentiment classification show that the result of the UTSU model is close to the results of supervised classification methods and outperforms other topic and sentiment unification models. Comparisons with the ASUM and JST models show that the UTSU model improves the F1 value of sentiment classification by about 3% and 17%, respectively.

Keywords: text sentiment classification; unsupervised learning; unification model

随着Web2.0的蓬勃发展，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。随着这类评论信息迅速膨胀，仅靠人工方法难以应对网上海量信息的收集和处理，因此，如何有效地管理和使用这些评价信息成为当前的紧迫课题，从而促进了自动文本情感分析技术的发展。

文本情感分析又称意见挖掘，简单而言，是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。近年来，文本情感分析已经成为自然语言处理领域的一个热点问题，特别是针对博客、Twitter、产品评论等的情感分析。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。情感分析方法目前主要有基于规则的方法和基于统计的方法。基于规则的方法通常需要基于一系列的语言分析与预处理过程，如词性标注、命名实体识别、句法分析等。新词的不断出现、表达方式的变化以及

收稿日期：2012-04-22。 作者简介：孙艳(1983—)，女，博士生；周学广(通信作者)，男，教授，博士生导师。 基金项目：国家自然科学基金资助项目(611100042)。
网络出版时间：2012-10-22 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20121022.0854.001.html

复杂的语言处理，使得基于规则的情感分析的可扩展性差，而且人工编写工作量大，成本较高。

基于统计的情感分析方法主要包括有监督、半监督和无监督情感分析。在有监督和半监督的机器学习方法中，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，而无监督的机器学习则无需标注的训练样本。

本文构建了一种无监督主题情感混合模型(Unsupervised Topic-Sentiment Unification model based on LDA-Col，简称 UTSU)，应用统计语言建模技术，捕捉语言的情感表达规律，在主题模型的基础上增加了情感模型。UTSU 模型采用词序流对文本进行表示，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

文本信息处理和文本挖掘领域的一个重要研究内容是文本表示。长期以来，文本表示的主要方法停留在向量空间模型(Vector Space Model，VSM)上。VSM 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。一些新的研究力图通过统计的途径对文本表示进行拓展，以挖掘文本的潜在语义，典型代表有潜在语义分析(Latent Semantic Analysis，LSA)[1]。LSA 将 TF-IDF 矩阵通过奇异值分解，分解为 3 个规模较小的矩阵的乘积，通过控制中间过渡矩阵的维数来控制保存空间的大小。LSA 方法在文本信息处理中得到了深入研究，其特征降维作用比较明显。由于奇异值分解算法复杂度高且难以并行化，Hofmann 从生成模型的角度提出了 pLSA 主题模型[2]。pLSA 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。

针对 pLSA 存在的问题，Blei 等人借鉴概率图模型理论和方法，提出了潜在狄里克雷分配(Latent Dirichlet Allocation，LDA)模型[3]。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。LDA 模型采用的是词袋(bag of words)表示方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为易于建模的数字信息，但是词袋表示方法没有考虑词与词之间的顺序，表达的只是一个粗略的上下文语义信息，如“高”和“铁”表达不出“高铁”的含义。为使得主题模型能够表达更精确的语义信息，Griffiths 等人在 LDA 模型的基础上提出了 LDA-Col(LDA-Collocation)模型[4]。该模型采用的是词序流(stream of words)表示方法，能够更准确地表达语义信息。

近年来，随着统计主题模型的逐渐兴起，很多学者将其应用到情感分析领域。Titov 等人提出了一种多粒度 LDA 模型(Multi-Grain LDA，MG-LDA)，并将其应用于基于主题的情感摘要生成中[5]，但是，MG-LDA 需要对已标注好的训练集进行训练，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 Zhao 等人提出的 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[6]。Brody 等人直接将句子作为一个文档，建立“句子～主题～单词”关系[7]。这种方法将 LDA 模型没有考虑文档和文档之间关系的缺点进一步放大，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。Mei 等人提出的 TSM 模型将情感与主题作为分开的语言模型，一个词或是情感词，或是主题词，只能二选一[8]。这种方法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。Jo 等人提出的 ASUM 模型[9]和 Lin 等人提出的 JST 模型[10]的区别是，ASUM 模型对每个句子采集情感标签和主题标签，而 JST 模型对每个词采集情感标签和主题标签。

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。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。

2　UTSU 模型

2.1　UTSU 模型的生成过程

UTSU 模型是在 LDA-Col 模型的基础上添加情感模型而构建的。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)，故 UTSU 模型假设一个句子的所有词由一种情感产生。在运行 UTSU 模型前，需先对文本进行预处理，将转折句从转折处分为 2 句。

假设语料库中有 M 个文档，记为 $D=\{d_1,\cdots,d_M\}$；共有 K 个主题，记为 $T=\{z_1,\cdots,z_K\}$；共有 L 种情感，记为 $S=\{m_1,\cdots,m_L\}$；文档 d 由 R 个句子组成，每个句子由 N 个词汇组成。记语料库中去重后的词汇表中的词汇数量为 V。UTSU 模型的框图如图 1 所示。

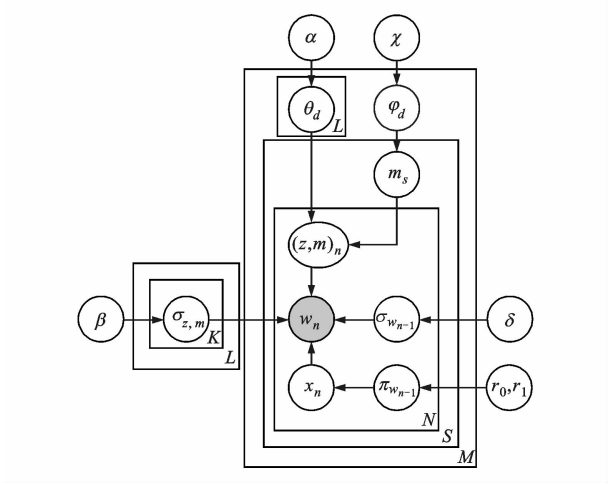

图 1 UTSU 模型框图

在图 1 中：圆圈表示随机变量（实心圆圈表示可观察到的变量，空心圆圈表示潜在变量）；方框表示重复抽样，重复次数在方框的右下角表示；箭头表示概率依存关系；θ_d 为文档 d 的主题分布；φ_d 为文档 d 的情感分布；$\sigma_{z,m}$ 为主题情感～词汇分布；$\sigma_{w_{n-1}}$ 为连词分布；$\pi_{w_{n-1}}$ 为词搭配分布；α、β、χ、δ 为 Dirichlet 分布参数；γ_0 和 γ_1 为 Beta 分布参数；w_n 表示词汇；m_s 表示句子 s 的情感；$(z,m)_n$ 表示 w_n 的主题和情感；x_n 表示词搭配分配，其取值如下

$$x_n=\begin{cases}1, & \text{if } w_n \text{ follows } w_{n-1}\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

式(1)表示若词 w_n 与 w_{n-1} 相连，则 x_n 等于 1，否则为 0。

UTSU 模型是一个 4 层盘子模型，其产生过程如下。

(1)对于每对主题 z 和情感 m：

从参数为 β 的 Dirichlet 分布中，抽取 Multinomial 分布 $\sigma_{z,m}$，即采样 $\sigma_{z,m}\sim\mathrm{Dir}(\beta)$。

(2)对于每篇文档 d：

①从参数为 χ 的 Dirichlet 分布中，抽取 Multinomial 分布 φ_d，即采样 $\varphi_d\sim\mathrm{Dir}(\chi)$；

②对于每种情感 j，从参数为 α 的 Dirichlet 分布中抽取 Multinomial 分布 θ_{dj}，即采样 $\theta_{dj}\sim\mathrm{Dir}(\alpha)$；

③对于文档 d 中的每个句子 s，

(A)从 φ_d 中抽取一个 Multinomial 变量情感 m_s，即采样 $m_s\sim\mathrm{Multi}(\varphi_d)$，

(B)对于句子 s 中的每个单词 w_n，

(a)从参数为 γ_0 和 γ_1 的 Beta 分布中抽取二项分布 $\pi_{w_{n-1}}$，即采样 $\pi_{w_{n-1}}\sim\mathrm{Beta}(r_0,r_1)$，

(b)从 $\pi_{w_{n-1}}$ 中抽取变量 x_n，若 $x_n=0$，则

(i)从 θ_{dm} 中抽取一个 Multinomial 变量主题 $z_{s,n}$，即采样 $z_{s,n}\sim\mathrm{Multi}(\theta_{dm})$，

(ii)从 $\sigma_{z,m}$ 中抽取一个 Multinomial 变量单词 w_n，即采样 $w_n\sim\mathrm{Multi}(\sigma_{z,m})$，

(c)若 $x_n=1$，

(i)从参数为 δ 的 Dirichlet 分布中抽取 $\sigma_{w_{n-1}}$，即采样 $\sigma_{w_{n-1}}\sim\mathrm{Dir}(\delta)$，

(ii)从 $\sigma_{w_{n-1}}$ 中抽取一个 Multinomial 变量单词 w_{n-1}，即采样 $w_{n-1}\sim\mathrm{Multi}(\sigma_{w_{n-1}})$。

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。$w=\{w_i=t,w_{-i}\}$，$z=\{z_i=k,z_{-i}\}$，$m=\{m_{s_i}=j,m_{-s_i}\}$。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 的概率可通过下式得到

$$p(z_i=k,m_{s_i}=j\,|\,z_{-i},m_{-s_i},w)\propto$$
$$\begin{cases}\dfrac{n_{k,j,-i}^{(t)}+\beta}{\sum_{t=1}^{V}(n_{k,j,-i}^{(t)}+\beta)}\dfrac{n_{d,-s_i}^{(j)}+\chi}{\sum_{j=1}^{L}(n_{d,-s_i}^{(j)}+\chi)}\dfrac{n_{d,j,-i}^{(k)}+\alpha}{\sum_{k=1}^{K}(n_{d,j,-i}^{(k)}+\alpha)}, & \text{if}\quad x_i=0\\ \dfrac{n_{d,-s_i}^{(j)}+\chi}{\sum_{j=1}^{L}(n_{d,-s_i}^{(j)}+\chi)}\dfrac{n_{d,j,-i}^{(k)}+\alpha}{\sum_{k=1}^{K}(n_{d,j,-i}^{(k)}+\alpha)}, & \text{if}\quad x_i=1\end{cases} \tag{2}$$

式中：$n_{k,j,-i}^{(t)}$ 表示除当前词汇记号 w_i 外，其他与 w_i 内容相同且主题和情感分别为 k 和 j 的词汇记号数量；$n_{d,-s_i}^{(j)}$ 表示除当前词汇记号所在句子外，文档 d 中情感为 j 的句子数量；$n_{d,j,-i}^{(k)}$ 表示除当前词汇记号外，文档 d 中情感为 j 主题为 k 的词汇记号数量。需要注意的是，所有数据统计都是在 x_i 值的基础上进行的，例如：当 $x_i=0$ 时，$n_{d,j,-i}^{(k)}$ 表示除当前词汇记号外，文档 d 中情感为 j 主题为 k 且满足 $x_i=0$ 的

词汇记号数量；当 $x_i=1$ 时，$n_{d,j,-i}^{(k)}$ 表示除当前词汇记号外，文档 d 中情感为 j 主题为 k 且满足 $x_i=1$ 的词汇记号数量。

x_i 从下式表示的分布中进行采样

$$p(x_i \mid x_{-i}, w, j) \propto \begin{cases} \dfrac{n_{k,j,-i}^{(t)}+\beta}{\sum_{t=1}^{V}(n_{k,j,-i}^{(t)}+\beta)} \dfrac{n_{w_{i-1}}^{0}+\gamma_0}{\sum_{u=0}^{1} n_{w_{i-1}}^{u}+\gamma_0+\gamma_1}, & x_i=0 \\ \dfrac{n_{w_{i-1}}^{(t)}+\delta}{\sum_{t=1}^{V}(n_{w_{i-1}}^{(t)}+\delta)} \dfrac{n_{w_{i-1}}^{1}+\gamma_1}{\sum_{u=0}^{1} n_{w_{i-1}}^{(u)}+\gamma_0+\gamma_1}. & x_i=1 \end{cases} \tag{3}$$

式中：$n_{k,j,-i}^{(t)}$ 表示除词汇记号 w_i 外，其他与 w_i 内容相同且主题和情感分别为 k 和 j 的词汇数量；$n_{w_{i-1}}^{0}$ 表示除词汇记号 w_{i-1} 外，其他与 w_{i-1} 内容相同且不构成词搭配的词汇数量；$n_{w_{i-1}}^{1}$ 表示除词汇记号 w_{i-1} 外，其他与 w_{i-1} 内容相同且构成词搭配的词汇数量；$n_{w_{i-1}}^{(t)}$ 表示除词汇记号 w_i 外，w_{i-1}、w_i 词汇内容同时出现的次数。

文档～主题分布 θ、文档～情感分布 φ 和主题情感～词分布 σ 的概率估计如下

$$\left.\begin{aligned} \hat{\theta}_{d,j,k} &= \frac{n_{d,j}^{(k)}+\alpha}{\sum_{k=1}^{K}(n_{d,j}^{(k)}+\alpha)} \\ \hat{\varphi}_{d,j} &= \frac{n_{d}^{(j)}+\chi}{\sum_{j=1}^{L}(n_{d}^{(j)}+\chi)} \\ \hat{\sigma}_{k,j,w} &= \frac{n_{k,j}^{(w)}+\beta}{\sum_{t=1}^{V}(n_{k,j}^{(w)}+\beta)} \end{aligned}\right\} \tag{4}$$

式中：$\hat{\theta}_{d,j,k}$ 表示主题 k 在文档 d 的主题分布中情感为 j 的概率估计；$\hat{\varphi}_{d,j}$ 表示情感 j 在文档 d 的情感分布中的概率估计；$\hat{\sigma}_{k,j,w}$ 表示词汇 w 分配的主题和情感分别为 k 和 j 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 中分配为主题为 k 情感为 j 的词的数量；$n_{k,j}^{(w)}$ 表示 w 分配在主题 k 情感 j 上的次数。

连词分布 σ_w 和词搭配分布 π_w 的概率估计如下

$$\hat{\sigma}_{w_1}^{(w_2)}=\frac{n_{w_1}^{(w_2)}+\delta_{w_2}}{\sum_{w_i=1}^{V}(n_{w_1}^{(w_i)}+\delta_{w_i})};\ \hat{\pi}_{w_1}=\frac{n_{w_1}^{1}+\gamma_1}{\sum_{u=0}^{1} n_{w_1}^{(u)}+\gamma_0+\gamma_1} \tag{5}$$

式中：$\hat{\sigma}_{w_1}^{(w_2)}$ 表示 w_1w_2 连词概率估计；$\hat{\pi}_{w_1}$ 表示给定 w_1、$x_2=1$ 的概率估计；$n_{w_1}^{(w_2)}$ 表示 w_1w_2 连词出现的次数；$n_{w_1}^{1}$ 表示 w_1 构成词搭配的出现次数。

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、烧烤的评论网页和中科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 9 180 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、负情感分布见表 1。

表 1　数据集

类别	文本数量			
	快递	烧烤	酒店	计算机
正类	1 150	910	1 130	1 270
负类	1 140	1 230	1 200	1 150
共计	2 290	2 140	2 330	2 420

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 2 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科院的 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　文本主题发现

为对比 LDA 模型和 LDA-Col 模型的效果，利用 Gibbs 采样算法对这 2 种模型进行采样，参数设置如下：$\alpha=1$，$\beta=0.01$，$\gamma_0=0.1$，$\gamma_1=0.1$，$\delta=0.1$，均为经验最优值；迭代次数 $N=1\ 000$；为方便看出词汇的类别，取主题数 $K=4$，在显示词汇时设置词的最大搭配长度 $C_{\max}=3$。表 2 为快递类的主题发现词汇，只选取了概率排名前 39 的词汇。

从表 2 可以看出，采用词序流表示文本的 LDA-Col 模型的主题发现词汇确实比采用词袋表示文本的 LDA 模型的语义信息更精确，特别是发现了一些带情感的搭配词，如“不会”“不知道”“比较差”“太差”“比较多”等，这些搭配词对情感分类有很大的帮助。

3.3　情感分类

本文实验的情感只考虑褒义和贬义 2 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：$\alpha=1$，$\chi=1$，$\beta=0.01$，$\gamma_0=0.1$，$\gamma_1=0.1$，$\delta=0.1$，$L=2$，$K=4$，$N=1\ 000$。在显示词汇

时，设置词的最大搭配长度 $C_{\max}=3$，得到的主题～情感词汇见表 3。

表 2 LDA 和 LDA-Col 快递类的主题发现词汇对比表

主题发现词汇					
LDA 模型			LDA-Col 模型		
不	快	北京	申_通	快	不_会
快递	要	比较	就	服务_态度	能
就	公司	挺	不	差	快递_员
都	速度	才	都	态度	一直
通	差	没有	说	要	时候
态度	打电话	货	送	服务	货
送	能	取	快递	公司	找
申	知道	一直	电话	打_电话	问
没	客	时候	东西	不_知道	家
说	服	家	申_通_快递	北京	已经
服务	发	接	没	发	结果
电话	申通	结果	客_服	快递_公司	取
东西	已经	买	速度	申通	寄

表 3 主题情感发现词汇表

主题～情感词汇					
烧烤(负类)			烧烤(正类)		
说	环境	不能	鸡翅	微辣	有点
没	没有	不_知道	家	觉得	哈_哈
要	烤翅	太	味道	烤	心
真	不_会	找	好吃	烤_馒头_片	挺
差	钱	真是	不错	原味	排
多	东西	气	点	变态_辣	早
一般	菜	出	多	口味	甜
点	地方	带	辣	特别	超
大_家	不好	才	感觉	说	香
服务	结果	串	非常	没有	全
大	会	失望	烤翅	火	可能
等	想	时候	喜欢	能	开
服务员	桌	服务_态度	朋友	嫩	推荐

从表 3 中可以看出，正、负情感词在主题～情感发现中分得比较明显，如左边表示贬义的负类情感词“差、一般、没有、不好、失望”等，右边表示褒义的正类情感词“好吃、不错、喜欢、火、嫩、甜”等。利用本文提出的 UTSU 模型的 $\hat{\varphi}_{d,j}$，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，即取 $m_d=\arg\max_j\{\hat{\varphi}_{d,j}\mid j\in[1,\cdots,L]\}$作为文档 d 的情感。

将 UTSU 模型与其他 3 种主题情感混合模型和有监督学习的 Pang 方法[11]在情感分类性能上进行了对比。其他 3 种主题情感混合模型包括 ASUM 模型、JST 模型以及在 LDA 模型基础上采用本文提出的采样方式构建的 UTSU 模型，记为 UTSU(LDA)模型。区别于 UTSU(LDA)模型，将本文在 LDA-Col 模型基础上提出的 UTSU 模型记为 UTSU(LDA-Col)模型。论述 ASUM 模型和 JST 模型的文献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，故本文统一采用无先验知识。Pang 方法中使用信息增益选取了 2 000 个特征，分类器采用 SVM，分类时采用十重交叉验证。上述 5 种方法在 4 个数据集上的情感分类准确率、召回率和 F1 综合指标如图 2～图 4 所示。

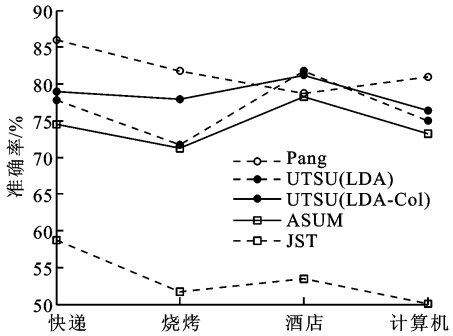

图 2 情感分类准确率对比图

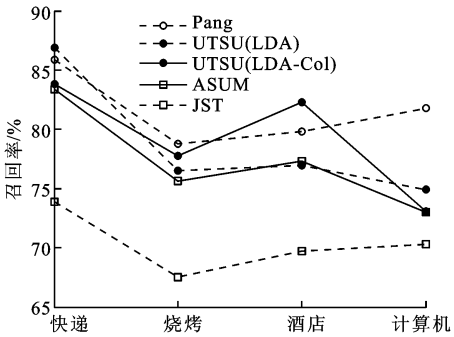

图 3 情感分类召回率对比图

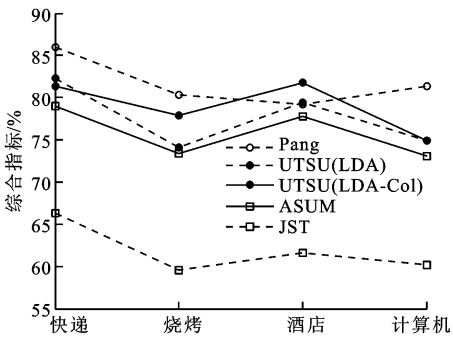

图 4 情感分类 F1 综合指标对比图

从图 2～图 4 中可以看出，在 5 种方法中，情感分类效果最好的是 Pang 方法，在 4 个数据集上的 F1 综合指标平均值为 81.71%。在 4 种无监督的主题情感

模型中，UTSU(LDA-Col)模型的情感分类效果最好，F1 综合指标平均值为 78.95%；其次为 UTSU(LDA)模型和 ASUM 模型，F1 综合指标平均值分别为 77.68%和75.81%；效果最差的是 JST 模型，在 4 个数据集上的 F1 综合指标平均值为 61.91%。

虽然在 5 种方法中 Pang 方法效果最好，但是 Pang 方法属于有监督学习方法，需要对标注好的样本进行训练后才能测试。对于训练集外的新的未标注数据集，其他 4 种主题情感混合模型都可以直接测试，但是 Pang 方法却不可以。UTSU(LDA)模型比 ASUM 模型的效果稍好，F1 综合指标平均值高出约 1.9%，这证明了本文提出的对每个句子采样情感标签、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。UTSU(LDA-Col)模型比 UTSU(LDA)模型的综合指标平均值提高了 1.3%，说明对词语进行搭配合并能够提高情感分类的准确率。由于 JST 模型每次采样情感标签时对每个词都进行采样，不符合自然语言的情感表达，因此效果最差。5 种方法在 4 个数据集上的情感分类效果不同，还与数据集正、负样本的区分度相关。总体来说，本文构建的 UTSU(LDA-Col)模型的情感分类性能比有监督的情感分类方法稍差，但在无监督的情感分类方法中效果最好，F1 综合指标平均值比 ASUM 模型提高了 3%，比 JST 模型提高了 17%。

4　总　结

本文主要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-Col 主题模型的基础上构建了一种无监督的主题情感混合模型——UTSU 模型。UTSU 模型以 LDA-Col 为基本主题模型，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 模型在同一层盘子中采样主题标签和情感标签的缺陷。通过比较实验，证明 UTSU 模型可以取得更好的分类效果，其中在酒店数据集中取得的 F1 综合指标最大值为 81.74%。UTSU 模型的分类性能仍有很大的提升空间，后续我们将考虑引入先验知识来进一步提升文本情感分类效果。

参考文献：

[1] LANDAUER T K, FOLTZ P W, LAHAM D. An introduction to latent semantic analysis [J]. Discourse Processes, 1998, 25(2/3): 259-284.

[2] HOFMANN T. Unsupervised learning by probabilistic latent semantic analysis [J]. Machine Learning, 2001, 42(1/2): 177-196.

[3] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(4/5): 993-1022.

[4] GRIFFITHS T L, STEYVERS M, TENEN-BAUM J B. Topic in semantic representation [J]. Psychological Review, 2007, 114(2): 211-244.

[5] TITOV I, MCDONALD R. A joint model of text and aspect ratings for sentiment summarization [C]// Proceedings of 2008 Association for Computational Linguistics with the Human Language Technology Conference. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008: 308-316.

[6] ZHAO Xin, JIANG Jing, YAN Hongfei, et al. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid [C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010: 56-65.

[7] BRODY S, ELHADAD N. An unsupervised aspect-sentiment model for online reviews [C]// Proceedings of the 2010 Annual Conference of the North American Chapter of the ACL. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010: 804-812.

[8] MEI Qiaozhu, LING Xu, WONDRA M, et al. Topic sentiment mixture: modeling facets and opinions in weblogs [C]// Proceeding of the 16th International Conference on World Wide Web. New York, USA: ACM, 2007: 171-180.

[9] JO Y, OH A. Aspect and sentiment unification model for online review analysis [C]// Proceedings of the 4th ACM International Conference on Web Search and Data Mining. New York, USA: ACM, 2011: 815-824.

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

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

（编辑　葛赵青　武红江）