软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(3):694−707 [doi: 10.13328/j.cnki.jos.005157]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于社交关系的微博主题情感挖掘*

黄发良[1,2,4], 于 戈[1,3], 张继连[5], 李超雄[2], 元昌安[7], 卢景丽[6]

[1](东北大学 计算机科学与工程学院,辽宁 沈阳 110819)
[2](福建师范大学 软件学院,福建 福州 350108)
[3](医学影像计算教育部重点实验室(东北大学),辽宁 沈阳 110819)
[4](福建省公共服务大数据挖掘与应用工程技术研究中心,福建 福州 350108)
[5](广西财经学院 信息与统计学院,广西 南宁 530003)
[6](Grasslands Research Centre, Palmerston North, 11008, New Zealand)
[7](广西师范学院 计算机与信息工程学院,广西 南宁 530023)
通讯作者: 于戈, E-mail: yuge@mail.neu.edu.cn

摘 要: 微博情感分析是社交媒体挖掘中的重要任务之一,在个性化推荐、舆情分析等方面具有重要的理论和应用价值.挖掘性能良好且可同步进行文档主题分析与情感分析的主题情感模型,近年来在以微博为代表的社交媒体情感分析中备受关注.然而,绝大多数现有主题情感模型都只简单地假设不同微博的情感极性是互相独立的,这与微博生态的现实状况不相一致,从而导致这些模型无法对用户的真实情感进行有效建模.基于此,综合考虑了微博用户相互关联的事实,提出了基于 LDA 和微博用户关系的主题情感模型 SRTSM(social relation topic sentiment model).该模型在 LDA 中加入情感层与微博用户关系参数,利用微博用户关系与微博主题学习微博的情感极性.针对新浪微博真实数据集上的大量实验结果表明:与代表性算法 JST,Sentiment-LDA 及 DPLDA 相比较,SRTSM 模型能够对用户真实情感与讨论主题进行更加有效的分析建模.

关键词: 情感分析;微博情感分析;主题情感模型;社交关系;社会媒体处理
中图法分类号: TP311

中文引用格式: 黄发良,于戈,张继连,李超雄,元昌安,卢景丽.基于社交关系的微博主题情感挖掘.软件学报,2017,28(3):694−707. http://www.jos.org.cn/1000-9825/5157.htm
英文引用格式: Huang FL, Yu G, Zhang JL, Li CX, Yuan CA, Lu JL. Mining topic sentiment in micro-blogging based on micro-blogger social relation. Ruan Jian Xue Bao/Journal of Software, 2017,28(3):694−707 (in Chinese). http://www.jos.org.cn/1000-9825/5157.htm

Mining Topic Sentiment in Micro-Blogging Based on Micro-Blogger Social Relation

HUANG Fa-Liang[1,2,4], YU Ge[1,3], ZHANG Ji-Lian[5], LI Chao-Xiong[2], YUAN Chang-An[7], LU Jing-Li[6]

[1](School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China)
[2](Faculty of Software, Fujian Normal University, Fuzhou 350108, China)
[3](Key Laboratory of Medical Image Computing of Ministry of Education (Northeastern University), Shenyang 110819, China)
[4](Fujian Engineering Research Center of Public Service Big Data Mining and Application, Fuzhou 350108, China)
[5](School of Information and Statistics, Guangxi University of Finance and Economics, Nanning 530003, China)
[6](Grasslands Research Centre, Palmerston North, 11008, New Zealand)
[7](College of Computer and Information Engineering, Guangxi Teachers Education University, Nanning 530023, China)

* 基金项目: 国家重点基础研究发展计划(973)(2012CB316201); 国家自然科学基金(61433008, 61363009, 61363037); 福建省教育厅 K 类科技项目(JK2016007)
Foundation item: National Basic Research Program of China (973) (2012CB316201); National Natural Science Foundation of China (61433008, 61363009, 61363037); Foundation of Fujian Educational Committee (JK2016007)
收稿时间: 2016-07-22; 修改时间: 2016-09-14; 采用时间: 2016-11-01; jos 在线出版时间: 2016-11-29
CNKI 网络优先出版: 2016-11-29 13:34:58, http://www.cnki.net/kcms/detail/11.2560.TP.20161129.1334.004.html

Abstract: Sentiment analysis in micro-blogging is an important task in mining social media, and has important theoretical and application value in personalized recommendation and public opinion analysis. Topic sentiment models have attracted much attention due to their good performance and ability of synchronized topic and the sentiment analysis in micro-blogs. However, most existing models simply assume that topic sentiment distributions of different micro-blogs are independent, which is contrary to the realistic status in micro-blogging and thus further leads to unsatisfactory modeling of micro-blogger's true sentiment. To address the issues, a probabilistic model, SRTSM (social relation topic sentiment model) is proposed. The new model introduces sentiment and micro-blogger social relation into LDA inference framework and achieves synchronized detection of sentiment and topic in micro-blogging. Extensive experiments on Sina Weibo show that SRTSM outperforms state-of-the-art unsupervised approaches including JST, SLDA and DPLDA significantly in terms of sentiment classification accuracy.

Key words: sentiment analysis; microblog sentiment analysis; topic sentiment model; social relation; social media processing

微博是 Web 2.0 时代兴起的一种集成化、开放化的互联网社交服务,它让用户能够向公众发布简短的文本消息.由于其简便的特点,日益受到互联网用户的青睐.目前,新浪微博用户规模已经超过 3 亿,每天都有大量的微博消息发布.在这些海量的微博消息中,有许多包含个人情感的资源,例如,不同读者对于同一条新闻事件持有不同的看法,不同用户对于某款手机有着其个性化的用户体验,不同影视爱好者对于同一部电影会留下不同的观影评论,等等.研究如何高效挖掘隐藏于这些鱼目混杂的微博消息中的观点与情感,即文本情感挖掘,有助于各级政府机构、企业组织与理性个体的管理决策.例如,政府机构可以对网络舆论进行实时监测与导向,网上商家能够根据用户反馈意见及时调整生产服务实现利润最大化,个体网民可以敏捷获取目标信息,等等.

以微博情感分析为代表的网络短文本情感挖掘正在吸引着来自人工智能、数据挖掘、自然语言处理等不同领域研究者的广泛关注[1-5],涌现出的各种算法大致可归纳为 3 类:有监督的情感挖掘、无监督的情感挖掘与半监督的情感挖掘.有(半)监督的情感挖掘方法不同程度地利用训练语料来训练生成文本情感分类器,一般具有较高的分类准确率,但获取训练样本的昂贵代价极大地限制此类方法的应用性.因此,以 JST[6],Sentiment-LDA[7]与 ASUM[8]等为代表的无监督情感分类方法近年来备受青睐,此类方法有效地避免了传统无监督情感分类方法具有的情感词典依赖性缺点,能够达到较好的情感识别效果.然而,现有的这些 LDA 情感主题模型也许还不能完全捕获网络短评用户的真实情感,下面以例 1 说明.

例 1:作为 NBA 球星库里的粉丝,用户 *A* 与用户 *B* 在微博上是相互关注的,二者针对 NBA 球星各自发了一条微博.

- 用户 *A*:"库里的三分球真是太准了,库里太厉害了,很崇拜他!"
- 用户 *B*:"库里太变态了,简直不是人!"

从用户 *A* 的微博可以看出:其对库里表达的是钦佩之情,情感是积极的,且现有的 LDA 主题情感模型可以正确地分析出该微博的情感极性,由于其包含"厉害""崇拜"等积极情感词;而对于用户 *B* 发表的微博,现有 LDA 主题情感模型往往将其归属到消极情感类中,由于该微博包括"变态""不是人"等贬义词.然而,联系到 *A* 与 *B* 互相关注的事实,这在一定程度上可以表明他们的兴趣爱好相似.用户 *A* 的微博总体情感极性为积极,那在判断用户 *B* 所发微博的时候,应该认为用户 *B* 的微博为积极情感极性的概率更大,但是现有 LDA 主题情感模型假设不同微博的情感极性是互相独立的,从而无法准确识别用户 *B* 所发微博的情感极性.事实上,社会心理学研究早在 20 多年前就已经得出人们在社会交互的过程中表现出情绪感染特性[32],亦即,朋友间交互表现出的情感比非朋友间交互表现出的情感更可能相似.近期社会媒体分析领域[24]发现,情绪感染特性也存在于网民的微博交互过程中.

从上面的分析可以看出:虽然以 JST,Sentiment-LDA 与 DPLDA 为代表的 LDA 情感主题模型可以获取单条微博的主题与情感极性,但这些模型都假设不同微博的情感极性与主题偏好是相互独立的,这一方面极大地损害模型的识别性能,另一方面,该假设是与微博生态的现实状况不相一致的,由于微博用户之间通过"粉""评论"与"转发"等行为可以建立起不同程度的关联.

针对以上不足,本文提出了基于 LDA 和微博用户关系的用户关系主题情感模型 SRTSM(social-relation-based topic sentiment model),该模型在 LDA 中加入情感层与微博用户关系参数,当判断用户 *A* 的微博

时,将参考与用户 A 相互关注的其他用户所发微博的总体情感极性,适当修改积极或消极情感极性先验参数.该模型与现有微博情感分析方法的主要区别是:SRTSM 考虑了微博用户关系对微博情感分析准确率的影响,在建模微博单词生成过程时加入用户关系参数,用以更好地刻画微博间的主题情感关系.

本文的贡献如下:

1) 综合考虑了微博用户相互关联的事实,基于 LDA 模型提出了适合于微博主题情感分析的新模型 SRTSM;
2) 利用吉布斯采样对 SRTSM 模型进行求解,实现情感与主题挖掘;
3) 在真实的微博数据集上对模型进行实验,表明 SRTSM 模型能够较好地对微博进行情感与主题挖掘.

本文第 1 节简要介绍相关工作.第 2 节将提出我们的模型.第 3 节将对我们的模型与其他模型进行实验,对它们的性能进行比较与分析.第 4 节将对本文进行总结.

1 相关工作

1.1 基于LDA的无监督情感挖掘

基于主题模型的无监督情感挖掘主要是通过应用主题建模技术对主观性文本进行学习来实现隐含情感知识的发现,作为完全生成模型的 LDA(latent dirichlet allocation)[9]主题模型,由于其具有良好的数学基础和灵活拓展性而被广为使用.

Mei 等人[10]提出了主题情感模型 TSM 进行主题及其相关情感的演化分析.TSM 一方面存在着类似 pLSI 所有的学习过度问题,另一方面需要相关后处理操作才能完成文档情感的预测.Titov 等人[11]应用 MG-LDA 提取评论对象中的各个被评价,然后提出 MAS 模型对情感进行总结,MAS 模型要求评论对象的每个方面至少在部分评论中被评价过,然而,这对真实评论文本数据集来说是不实际的.Dasgupta 等人[12]提出一种基于用户反馈的谱聚类技术进行网络文本的无监督情感分类,聚类分析过程涉及数据特征都是具有情感倾向的主题,然而在该分析过程中,需要人为指定最重要的特征维.Lin 等人[6]提出一种基于 LDA 模型的 JST 模型,该模型将文本情感标签加入 LDA,形成了包含词、主题、情感和文档的 4 层贝叶斯概率模型.电影评论数据集上的实验表明, JST 模型的分类效果要优于 Pang 等人的有监督分类,但由于该模型是基于 BOW(bag of words)模型的文本特征之间相互独立的假设而没有考虑词的语境,会导致 not good movie 被分为积极情感与 not bad movie 被分为消极情感的错误.观测到 JST 模型中的 Gibbs 采样推理过程中出现大量“1”的现象,He[13]对 LDA 模型的目标函数进行修改,即:在建立情感先验分布时,应用广义期望标准来表达情感词的情感期望.Jo 等人[8]提出与 JST 类似的情感分类主题模型 ASUM,将 JST 中的主题替换为方面(aspect).为了克服 JST 的不足,Li 等人[7]提出与 JST 类似的 4 层贝叶斯概率模型 Dependency-Sentiment-LDA,引入一个转移变量来刻画单词之间的情感关联性. Brody 等人[14]对主题词进行了情感识别,然而没有建立文档或句子的情感模型.基于产品评分是与产品某个方面质量的优劣相互依赖的,Moghadda 等人[15]提出 ILDA 模型,通过增加相关参数来改进 LDA,依据产品的文本评论同时实现产品属性方面的提取与评分.孙艳等人[16]提出一种主题情感混合模型 UTSU,通过对每个句子与词分别进行采样情感标签与主题标签来得到各个主题的主题情感词,进而实现文本的情感分类.Samaneh 等人[17]提出 D-PLDA 模型,假设文本为 bag-of-phrases 模型,基于 bag-of-phrases 对文本提取主题词与情感词. Mukherjee 等人[18]提出 SAS 模型,假设我们已有待建模语料的种子词集,然后利用这些种子词集对 aspect 词语进行簇分析,进而得到文本的 aspect 词语与情感词语.欧阳继红等人[19]提出两个多粒度主题情感混合模型:文档级 MGR-JST 与局部 MG-JST.Rao 等人[20]提出有监督的多标签主题模型 MSTM 和隐含情感主题模型 SLTM 对社交情绪分类.Li 等人[21]提出了基于文本主题与用户-商品潜在因子的有监督情感分析模型 SUIT.Yang 等人[22]提出了用户感知的主题情感模型 USTM,该模型把评论者的人口统计学信息纳入到主题建模过程中.黄发良等人[33]提出一种新的基于 LDA 和互联网短评行为理论的主题情感混合模型 TSCM.

与 JST 等上述模型相同,本文提出的 SRTSM 对每个词进行情感与主题采样.不同的是:SRTSM 在基于采样的情感推理过程会根据微博用户的关系分布对微博的情感极性进行调整,而上述模型却将微博用户关系直接

忽略.

1.2 基于微博行为分析的情感识别

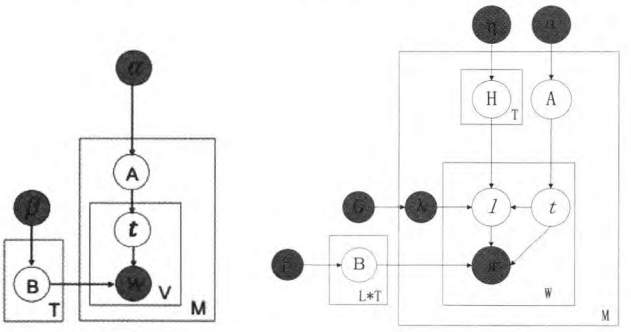
在微博社交平台上,微博用户之间通过关注、粉丝、互相关注等行为实现信息分享与传播,目前已有不少学者尝试利用微博用户关系对微博情感分析展开研究.Zhou 等人[23]提出:一个社区不仅由社区中活跃度较高的人组成,而且由这些活跃度较高的人所讨论的话题组成.基于此,提出 COCOMP(collaborator community profiling)模型来挖掘社区中参与度较高的人与这些人讨论的热门话题.COCOMP 通过如下方法产生文本集:首先产生文本 d 的社区 Cd,然后对于每个人 p,决定 p 是否属于社区 Cd,最后生成文本 d 的单词.通过真实 Twitter 数据集证明该模型可以挖掘不同社区中活跃度较高的人与社区中的热门话题.Hu 等人[24]提出一种社会学方法来分析 Twitter 的情感极性,该方法将情感一致性与情感传染理论融入有监督学习,并且利用稀疏学习来处理微博中的噪声.通过两个真实 Twitter 数据集上的实验证明,该方法具有较好的微博情感分析性能.West 等人[25]利用微博用户交互文本的情感值构造正负值加权的用户关系网络,进而预测用户彼此间的看法与观点.Wu 等人[26]提出结构化微博情感分类框架 SMSC,根据两种不同社交关系(不同微博用户间的链接与同一用户不同微博消息间的链接)将情感分类问题转化为图优化问题,从而实现微博情感分类.Tan 等人[27]将微博用户的关注、粉丝与"@"关系加入到一个半监督学习框架中,提出新的模型来提高微博情感分析准确率.该模型假设有着某种关联的用户有较大可能拥有相同的性格,相比只用词作为特征训练的 SVM 分类器,该模型具有更高的微博情感分析准确率.Lu 等人[28]使用微博关系构造了基于图的半监督学习分类器 SSA-ST 来分析微博情感,该方法综合微博用户关系与微博文本相似性来建立微博关系.Speriosu 等人[29]把微博粉丝融入微博情感分析,利用标签传播算法将最大熵分类器从噪声数据里训练的标签、词典里单词类型知识与微博粉丝图结合,用来分析微博情感.为了克服微博文本较短且难以对意见进行分析的不足,Fu 等人[30]分别构建了用户关注图与用户粉丝图,并将其融入贝叶斯与 SVM 分类器,以此提高贝叶斯与 SVM 分类器对微博情感分析的准确率.

上述模型虽然都使用微博用户关系对微博进行分析处理,但是这些模型都存在着一些缺陷:(1) 虽然文献[23]对微博用户与单词的生成过程进行建模,但是只能发现社区中活跃度较高的人与社区的热门话题,无法对话题进行情感分析;(2) 虽然文献[24–30]利用微博用户关系提高了微博情感分析准确率,但是这些模型都是基于有(半)监督学习的,没有基于无监督 LDA 模型来构建微博单词的生成过程.

2 基于社交关系的微博主题情感模型

2.1 模型描述

LDA 模型是 Blei 等人于 2003 年提出的"文档-主题-单词"三层贝叶斯模型(如图 1 所示,该图中的符号说明见表 1),通过运用概率推导方法来寻找数据集的语义结构,从而得到文本的主题.



(a) LDA (b) SRTSM

Fig.1 Graphical models LDA and SRTSM

图 1 图模型 LDA 与 SRTSM

Table 1 Symbols and notations

表 1 相关符号

符号	说明	符号	说明
α	微博-主题分布的 Dir 参数	η	(微博,主题)-情感分布的 Dir 参数
β	(主题,情感)-词语分布的 Dir 参数	λ	用户关系参数
A	微博-主题分布	B	(主题,情感)-词语分布
H	(微博,主题)-情感分布	G	用户关系分布
t	主题	w	词语
l	情感	M	微博数
W	微博中词语数	L	情感数
T	主题数	V	微博词库的词语数

该模型是建立在如下假设之上的:文档是由不同主题组成的,而一个主题是单词集合的概率分布.在此假设下,文档单词的产生步骤可以分为两个阶段:首先,从文档-主题分布中选择一个主题;然后,根据随机选择的主题从主题-单词分布中选择一个单词.

牢固的数学基础与良好的扩展性,使得完全生成模型 LDA 在文本主题挖掘研究中广为使用,但情感层的缺失使得 LDA 无法完成文档情感的分析.基于此,我们对 LDA 进行改造,通过在 LDA 中嵌入情感层,并在情感层中添加用户关系分布 G 和用户关系参数 λ.在 SRTSM 中,情感标签不仅与文档、主题相关联,而且还与微博用户社交关系分布相关联.

对于微博集 $C=\{d_1,d_2,\ldots,d_M\}$,其中,M 为微博集的微博数,与微博集 C 对应的词典大小为 V,微博 d_m 由 W_m 个单词组成,即,$d_m=\{w_1,w_2,\ldots,w_{Wm}\}$.SRTSM 产生微博集 C 的过程可简单归结为如下两个步骤.

1) 初始化 SRTSM 模型的分布 $\Theta=\{A,B,H\}$.具体地,A,B 与 H 分别服从狄利克雷分布 $Dir(\alpha)$,$Dir(\beta)$与 $Dir(\eta)$,其中,β是指单词在微博集 C 中出现的先验次数,η是指情感标签 l 在微博 d_m 中出现的先验次数,α是指主题 t 在微博 d_m 中出现的先验次数;

2) 生成微博集 C 中的单词,此生成过程可简单描述如下:首先,从微博-主题分布 A 中选出一个主题 t,t 服从 $Mul(A)$分布($Mul(*)$表示多项分布);接着,根据产生的主题 t,从(微博,主题)-情感分布 H 中选出一个情感标签 l,l 服从 $Mul(H)$分布并且受 λ 的影响,λ受用户关系分布 G 影响,G 为已知的微博用户关系矩阵.假设当前微博作者为用户 X,若用户 X 与 Y 互相关注,则 $G_{X,Y}=1$,否则 $G_{X,Y}=0$,当 $G_{X,Y}$ 为 1 时,计算用户 Y 的情感极性值,通过所有与用户 X 互相关注的用户的情感极性值确定用户关系参数 λ;最后,根据选出的主题 t 和情感并且 l,从(主题,情感)-词语分布 B 中选择一个单词 w,w 服从 $Mul(B)$分布.该生成过程的形式化描述见算法 1.

算法 1. SRTSM 词生成过程.

```
for each t∈{1,2,...,T}
    for each l∈{1,2,...,L}
for each v∈{1,2,...,V}
    choose B_{t,l,v}~Dir(β)
for each document d_m
    choose A_{m,t}~Dir(α)
    for each t∈{1,2,...,T}
        choose H_{m,t,l}~Dir(η)
for each document d_m
    for each word in d_m
        choose t~Mul(A_m)
        choose l~Mul(H_{m,t})
        choose w~Mul(B_{t,l})
```

2.2 模型推导

SRTSM 模型的推导采用吉布斯采样(Gibbs sampling)的方法,吉布斯采样是一种快速高效的 MCMC (Markov chain Monte Carlo)采样方法,它是通过迭代的采样方式对复杂的概率分布进行推导.为了得到参数分布 A,B 与 H,我们需要计算联合分布 $p(t_i=t,l_i=l|t_{-i},l_{-i},w)$,其中,$t_{-i}$ 与 l_{-i} 分别是指除微博 d_m 中第 i 个词以外的其他词的主题与情感标签.联合分布可以拆分为如下项:

$$P(w,t,l)=P(w|t,l)P(l|t)P(t) \tag{1}$$

通过对公式(1)进行展开可得:

$$P(w|t,l)=\int P(w,|t,l,B)P(B|\beta)\mathrm{d}B=\left(\frac{\Gamma(V\beta)}{[\Gamma(\beta)]^V}\right)^{T\times L}\times\prod_{l=1}^{L}\prod_{t=1}^{T}\frac{\prod_{w=1}^{V}\Gamma(n_{t,l,w}+\beta)}{\Gamma(n_{t,l}+W\beta)} \tag{2}$$

其中,$n_{t,l,w}$ 表示单词 w 同时属于主题 t、情感标签 l 的频数,$n_{t,l}$ 表示所有同时属于主题 t、情感标签 l 的单词总频数.$\Gamma(*)$表示伽马函数.

$$P(l|t)=\int P(l|t,\lambda,H)P(H|\eta)\mathrm{d}H=\left(\frac{\Gamma(L\alpha)}{[\Gamma(\alpha)]^L}\right)^{M\times T}\times\prod_{m=1}^{M}\prod_{t=1}^{T}\frac{\prod_{l=1}^{L}\Gamma(n_{m,t,l}+\eta+\lambda)}{\Gamma(n_{m,t}+L\eta)} \tag{3}$$

其中,$n_{m,t,l}$ 表示微博 d_m 的中情感标签为 l 的词语属于主题 t 的频数,$n_{m,t}$ 表示微博 d_m 中属于的主题 t 的词语总频数.λ为用户关系参数,λ由与当前微博作者相互关注用户的情感极性决定,每个微博用户的λ取值不同.

$$P(t)=\int P(t|A)P(A|\alpha)\mathrm{d}A=\left(\frac{\Gamma(T\eta)}{[\Gamma(\eta)]^T}\right)^{M}\times\prod_{m=1}^{M}\frac{\prod_{t=1}^{T}\Gamma(n_{m,t}+\alpha)}{\Gamma(n_m+T\alpha)} \tag{4}$$

其中,$n_{m,t}$ 表示微博 d_m 中主题 t 出现的频数,n_m 表示微博 d_m 总单词数.

有了公式(2)~公式(4)后,就可以计算吉布斯采样的联合概率.

$$\begin{aligned}p(t_i=t,l_i=l|t_{-i},l_{-i},w)&=\frac{P(w|t,l)P(l|t)P(t)}{P(w)P(w_{-i}|t_{-i},l_{-i})P(l_i|t_{-i})P(t_{-i})}\\&\propto\frac{\{n_{t,l}^{w}\}_{-i}+\beta}{\{n_{t,l}\}_{-i}+W\beta}\times\frac{\{n_{m,t}^{l}\}_{-i}+\eta+\lambda}{\{n_{m,t}\}_{-i}+L\eta}\times\frac{\{n_{m}^{t}\}_{-i}+\alpha}{\{n_{m}\}_{-i}+T\alpha}\end{aligned} \tag{5}$$

其中,$\{n_{t,l}^{w}\}_{-i}$ 表示除了当前单词,所有微博中单词 w 同时属于主题 t 和情感标签 l 的频数;$\{n_{t,l}\}_{-i}$ 表示除了当前单词,所有微博中属于主题 t 和情感标签 l 的单词总频数;$\{n_{m,t}^{l}\}_{-i}$ 表示微博 d_m 中,除了当前单词,情感标签 l 属于主题 t 的频数;$\{n_{m,t}\}_{-i}$ 表示微博 d_m 中,除了当前单词,属于主题 t 的情感标签总频数;$\{n_{m}^{t}\}_{-i}$ 表示除了当前单词,微博 d_m 中主题 t 的频数;$\{n_m\}_{-i}$ 表示除了当前单词,微博 d_m 的单词总数.

进一步利用最大似然估计方法对参数$\Pi=\{A,B,H\}$进行估计,其可形式化为公式(6)~公式(8).

$$B_{t,l}^{w}=\frac{n_{t,l}^{w}+\beta}{n_{t,l}+W\beta} \tag{6}$$

$$A_{m}^{t}=\frac{n_{m}^{t}+\alpha}{n_m+T\alpha} \tag{7}$$

$$H_{m,t}^{l}=\frac{n_{m,t}^{l}+\eta+\lambda_l}{n_{m,t}+L\eta} \tag{8}$$

其中,$B_{t,l}^{w}$ 表示所有微博中词语 w 同时属于主题 t 和情感标签 l 的概率,A_{m}^{t} 表示微博 d_m 中,主题 t 出现的概率,$H_{m,t}^{l}$ 表示微博 d_m 中情感标签 l 属于主题 t 的概率.

2.3 情感挖掘

在推导出 SRTSM 模型求解需要用到的公式后,我们就可以通过 SRTSM 模型来判断文档的情感极性.为方

便叙述,构造临时参数向量 $TmpVec=(n_{m,t,l},n_{m,t},n_m,n_{t,l,w},n_{t,l})$.

对于微博集 $C=\{d_1,d_2,\ldots,d_{|C|}\}$,

- 首先,对每条微博 $d=\{w_1,w_2,\ldots,w_n\}$中的每个单词 w 随机分配情感极性 l 与主题 t,并且更新向量 $TmpVec$,直到微博集 C 中每个微博 d 的所有单词都已被分配情感极性 l 与主题 t;
- 然后,循环执行如下过程 MAX 次(MAX 是指定的循环控制参数):对每篇微博 d 中的每个单词 w,计算 $p(z_i=z,l_i=l|z_{-i},l_{-i},w)$,并且更新向量 $TmpVec$,若当前迭代次数大于某一个指定值 X(本文取 X=1000),则每 Y 次(Y=10)更新分布 A,B 与 H;
- 最后根据 $H^l_{m,t}$ 计算微博的情感极性,若微博属于积极情感的概率大于微博属于消极情感的概率(即 $H^0_{m,t}>H^1_{m,t}$,其中,0 为积极情感,1 为消极情感),则判定该微博的情感极性为积极;反之,则判定该微博的情感极性为消极.

我们将基于 SRTSM 模型的过程形式化为算法 SRTSM_Miner.

算法 2. SRTSM_Miner.

输入:微博集 $C,G,\alpha,\beta,\eta,L,T$;

输出:微博情感极性.

```
1:  初始化分布 A,B 和 H,并对微博集 C 中的词语进行话题与情感随机初始化
2:  Count=1;
3:  while Count<=MAX
4:    for each d_m∈C do
5:      for each word w in d_m do
6:        从 TmpVec 中除去当前词语 w 所属的情感标签与主题;
7:        利用公式(4)重新给 w 赋一个情感标签和主题,情感标签受到用户关系参数λ的影响,通过用户关
          系分布 G 查找与当前微博作者相互关注的用户并以此更新λ;
8:        更新变量 TmpVec;
9:      end for
10:   end for
11:   若 Count>1000,每 10 次循环根据公式(6)~公式(8)用 TmpVec 更新分布 A,B,H;
12:   Count=Count+1;
13: end while
14: for each d_m∈C do
15:   if H^0_{m,t} > H^1_{m,t} then d_m.l=0 else d_m.l=1;
16: end for
```

3 实验与分析

为了定量地分析和比较不同模型的性能,我们在 3 个不同的真实微博数据集进行实验,然后分别从情感分类准确率、用户关系对准确率的影响、主题提取与时间效率 4 个方面进行分析和比较.实验环境为:CPU 为 Intel Core i7-2600M@3.4GHz,内存 8G,OS 为 Windows 7.

3.1 数据集

由于微博主题情感分析研究目前还处于萌芽状态,再加上诸如新浪、Twitter 之类的微博平台处于隐私安全需要对其提供的微博服务加以各种不同限制条件,从而导致在科研上很少有用于实验比较的标准数据集.虽然有少部分公开的文本情感分析实验数据集,诸如电子商务评论数据(电影评论、Amazon 商品评论)与社交媒体数据(Sentiment140),但这些数据仅包含文本数据而缺乏用户之间的链接关系数据,从而无法满足本实验的要

求.基于此,我们通过调用新浪微博 API 接口编写网络爬虫来构造实验数据集(见表 2).

Table 2 Experimental data sets

表 2 实验数据集

数据集	用户数	文档数	正向情感	负向情感
Data1	121	10 000	5 000	5 000
Data2	98	10 000	5 000	5 000
Data3	128	10 000	5 000	5 000

对于采集到的微博数据,我们通过聘请第三方人员对数据集的情感极性进行人工标注.在微博情感人工标注的过程中,我们请 3 个微博情感标注者对采集微博数据进行情感极性标注,并对标注结果的一致性进行 Kappa 检验,检测结果见表 3.

Table 3 Consistency examination of labeling microblog sentiment polarity

表 3 微博情感标注一致性检测

数据集	标注者对(1-2)	标注者对(1-3)	标注者对(2-3)
Data1	0.783	0.755	0.777
Data2	0.785	0.817	0.796
Data3	0.754	0.749	0.761

对于情感标注不一致的微博,我们根据 high-voting 的投票原则来确定其情感极性归属.

从表 3 可以看出:与数据集 Data1 和 Data2 相比较,Data3 的情感人工标注结果一致性较低,这说明 Data3 中的微博情感模糊性相对较强,可能会给微博情感自动分析提出更大的挑战.经过分词与去停用词等相关预处理的文本数据见表 4.

Table 4 Comparison of data sets

表 4 数据集预处理前后对比

数据集	单词表个数(处理前)	单词表个数(处理后)	文本平均长度(处理前)	文本平均长度(处理后)
Data1	24 088	20 777	73.46	42.76
Data2	24 801	21 381	80.20	44.29
Data3	23 962	20 593	75.38	41.12

3.2 情感分类准确率

为了评价 SRTSM 的微博情感识别能力,考虑到 SRTSM 模型学习的无监督性,我们将其与当前最具代表性的无监督情感学习模型(JST[6],Sentiment-LDA[7]和 DPLDA[8])、半监督情感模型 SSA-ST[28]、基于分词特征(1-gram+2-gram)的有监督算法 libSVM 进行微博情感分类正确率 ACC 的比较.通过独立同分布的随机抽样,对数据集(Data1,Data2 与 Data3)分别构造 8 组实验数据集.实验结果分别见表 5~表 7.

Table 5 Sentiment classification accuracy in Data1 (%)

表 5 Data1 情感分类准确率 (%)

	libSVM	SSA-ST	JST	Sentiment-LDA	DPLDA	SRTSM
1	71.66	69.23	62.62	59.92	60.71	**66.26**
2	70.29	68.72	63.34	60.99	61.83	**67.73**
3	69.54	64.34	61.41	58.28	55.24	**64.95**
4	68.84	65.58	62.27	59.83	57.78	**69.33**
5	72.38	68.83	60.67	58.70	55.45	**64.76**
6	71.66	66.65	63.41	62.11	58.73	**65.49**
7	71.71	67.74	64.43	61.39	56.53	**66.30**
8	70.42	65.27	65.34	60.35	59.73	**68.90**
Average	70.81	67.05	62.94	60.20	58.25	**66.72**

Table 6 Sentiment classification accuracy in Data2 (%)

表 6 Data2 情感分类准确率 (%)

	libSVM	SSA-ST	JST	Sentiment-LDA	DPLDA	SRTSM
1	71.25	67.22	63.15	61.49	61.55	**66.09**
2	73.51	65.45	62.03	61.83	59.74	**65.13**
3	70.62	67.73	65.07	62.31	61.03	**66.41**
4	71.47	70.08	69.61	64.62	61.91	**71.73**
5	69.93	70.25	66.67	65.42	59.13	**69.54**
6	72.16	71.16	68.08	63.4	61.78	**70.73**
7	73.09	69.25	64.1	68.23	63.44	**69.5**
8	71.24	68.87	67.18	64.96	59.58	**68.12**
Average	71.66	68.75	65.74	64.03	61.02	**68.41**

Table 7 Sentiment classification accuracy in Data3 (%)

表 7 Data3 情感分类准确率 (%)

	libSVM	SSA-ST	JST	Sentiment-LDA	DPLDA	SRTSM
1	70.71	67.87	60.52	63.7	57.72	**67.51**
2	69.41	62.58	58.29	62.86	59.29	**63.67**
3	68.87	66.03	62.08	58.84	61.47	**65.64**
4	69.25	65.19	61.72	53.45	55.71	**64.6**
5	67.38	67.54	56.78	63.94	58.65	**66.32**
6	69.16	64.88	63.4	56.68	61.74	**65.19**
7	65.55	64.25	58.75	55.99	60.75	**63.39**
8	66.69	65.21	58.35	57.5	61.81	**64.91**
Average	68.38	65.44	59.99	59.12	59.64	**65.15**

由上述各表可以看出:(1) SRTSM 的微博情感分类正确率远高于其他 3 种非监督情感分类算法 JST,Sentiment- LDA 和 DPLDA,尽管其在不同数据集上有不同的性能表现(在数据集 Data2、Data1 与 Data3 上的情感分类表现分别为最好(68.41%)、最差(66.72%)与次差(65.15%));(2) 与 SSA-ST 相比较,SRTSM 的微博情感分类正确率略低,但这劣势非常微弱;(3) 检测 3 个表中的最大 ACC 可以发现,所有最大的 ACC 都出现在 SVM 中,这说明 SVM 具有比其他非监督情感分类算法高的情感识别能力,而这正好与 Pang 等人[31]的实验结论相吻合.尽管与 SVM 相比较,SRTSM 的情感分类能力存在着差距,但考虑到获取带标签微博情感训练数据的高昂代价,其差距还是可以接受的.值得指出的是,与半监督情感分析方法 SSA-ST 和有监督情感分析方法 SVM 相比较,SRTSM 的情感分类正确率略低一些,但是 SRTSM 无需消耗任何带情感标签的微博消息数据,这可以为算法应用节约巨大的成本,而这些代价成本是 SSA-ST 与 SVM 无法回避的.

为进一步分析 SRTSM 的微博情感识别能力,我们利用 4 个不同指标(正例召回率 Recall、正例命中率 PV+、反例召回率 Specificity 与反例命中率 PV−)对 SRTSM 在 3 个数据集上的平均表现进行评价,实验结果见表 8.

Table 8 Sentiment detection performance of SRTSM in experimental data sets

表 8 SRTSM 的情感检测性能表现

Dataset	Recall (%)	PV+ (%)	Specificity (%)	PV− (%)
Data1	82.95	39.38	59.01	68.71
Data2	80.81	45.11	63.40	66.66
Data3	82.62	53.70	68.33	71.88

正例召回率 Recall、正例命中率 PV+、反例召回率 Specificity 与反例命中率 PV−可分别形式化为公式(9)~公式(12).

$$\text{Recall}=\text{正确分类的正例数}/\text{实际正例总数} \tag{9}$$

$$\text{PV+}=\text{正确分类的正例数}/\text{预测正例总数} \tag{10}$$

$$\text{Specificity}=\text{正确分类的负例个数}/\text{实际负例总数} \tag{11}$$

$$\text{PV−}=\text{正确分类的负例数}/\text{预测负例总数} \tag{12}$$

从表 8 可以发现一些有趣的现象:SRTSM 在 3 个实验数据集上的正例召回率 Recall 均高于其对应的反例召回率 Specificity,而正例命中率 PV+都高于反例命中率 PV−,虽然高出的程度在不同数据集上存在着差异.对

此可以做出这样的解释:由于微博社交媒体中,正面的消息更易于在朋友间传播,特别是在舆情监控的环境中,朋友之间在做出转发、评论或发布负面消息的决策时是非常谨慎的;微博用户在对一些诸如天怒人怨的事件往往表现非常强烈的消极情感倾向,进而使得消极情感消息更易于识别,即有 SRTSM 命中率更高.

3.3 用户关系对准确率的影响

为了评价用户关系对情感识别准确率的影响程度,我们首先对 3 个数据集的微博用户集分别构造有向关注网络,即形成 3 个表示关注关系的布尔矩阵 $Mat1_{121\times121}$,$Mat2_{98\times98}$ 与 $Mat3_{128\times128}$;然后,根据给定关系比例(0,10%,20%,30%,40%,50%,60%,70%,80%,90%,100%)随机选取用户关系,形成具有不同用户关系比例的微博数据集,实验结果如图 2 所示.

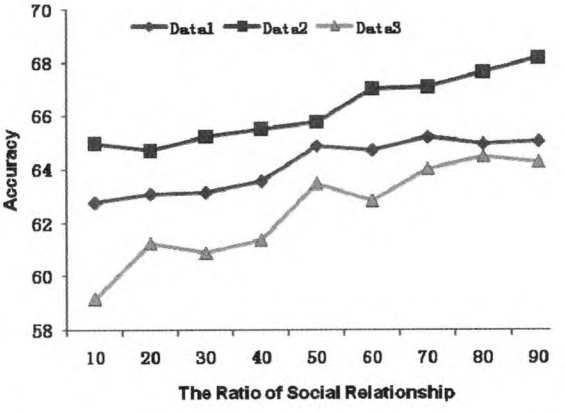


Fig.2 Impact of micro-blogger relationship ratio on accuracy

图 2 微博用户关系比例对准确率的影响

从图 2 可以看出:虽然随着互相关注用户比例的增加,微博情感分析准确率有提升也有降低,但是总体趋势是提升的.从 Data1 的曲线可以看出:除了在用户比例 50%~60%和 70%~80%处准确率呈下降趋势外,随着比例的提高,微博情感分析准确率呈上升趋势,在 40%~50%处提升最大,并且在 50%处达到准确率最大值.对于 Data2 来说,微博集在 10%~20%处出现了降低趋势,其余都是呈上升趋势,在 50%~60%处上升最多,在 90%处达到最大值.类似地,微博集 Data3 在用户比例为 20%~30%,50%~60%和 80%~90%处准确率呈降低趋势,其余准确率都呈上升趋势,在 10%~20%与 40%~50%处准确率提高较大,在 80%处达到最大值.

从上述分析可以得出:微博用户关系对准确率的影响较大,当互相关注的用户比例较大时,微博情感分析准确率较高,所以微博用户关系有助于提高微博情感分析的准确率.

3.4 主题提取

本实验用 SRTSM 分别对 Data1,Data2 与 Data3 进行主题提取,并且列出积极情感与消极情感出现概率最高的 15 个主题词,结果见表 9.

Table 9 Topic words of data sets

表 9 数据集主题词

	积极情感主题词	消极情感主题词
Data1	留言、泡面、链接、宝宝、喜欢、抽奖、好看、手机、颜色、安利、外套、可爱、分享、好玩、什么	天气、可能、转发、最近、孩子、苦恼、数据、一起、反转、包括、接受、温度、烦躁、衣服、不会
Data2	今天、谢谢、猎人、有趣、湖南卫视、期待、节目、游戏、支持、可以、加油、主意、电视剧、搞笑、微访谈	付费、音乐、表情、帮扩、难过、记得、时候、麻烦、希望、可以、不爽、看到、体谅、事情、流行
Data3	直播、恭喜、想念、控制、生日、谢谢、礼物、聚会、平台、快乐、唱歌、喜欢、女神、特别、漂亮	面包、猫打架、头发、需要、害怕、没有、知道、放空、难受、好像、燃烧、拒绝、肚子、听说、行动

Data1 的积极情感主题词中包含“喜欢”“好看”“可爱”“好玩”等积极情感极性较强的词语,从“泡面”“抽奖”“手机”等可以看出:这可能是一个讨论抽奖的话题,用户应该是抽中了手机、泡面等奖品,因此表示自己喜悦的心情.Data1 的消极情感主题词中出现的“苦恼”“烦躁”等词较明显地展示了用户消极的情感,“天气”“孩子”等词表达了用户对天气不好而造成孩子生病的抱怨.

从 Data2 的积极情感主题词可以看出:“谢谢”“有趣”“支持”等词具有较强的积极情感色彩,“湖南卫视”“节目”“游戏”等词可能说明用户们正在谈论湖南卫视的一档节目,对该节目某个环节的游戏或其他表示了支持,“电视剧”“搞笑”等词表示用户可能正在讨论一部搞笑的电视剧.而 Data2 消极情感主题词中的“难过”“麻烦”、“不爽”具有较强的消极情感色彩,其中,“付费”“音乐”“流行”等词也许是在谈论某些流行歌曲需要付费,用户对此表达了自己的不满.

Data3 积极情感主题词中出现了“恭喜”“谢谢”“快乐”“喜欢”“漂亮”,这些词具有较强的积极情感极性,“生日”“礼物”“聚会”“唱歌”等向我们展示了一幅庆祝生日的场景,表示用户们可能在谈论一场生日会或者是帮某人过生日.Data3 消极情感主题词中“害怕”“难受”具有较强消极情感极性,从“好像”“燃烧”“肚子”中可以看出,该主题应该是对于肚子疼或其他类似主题的探讨.而这样的主题通常是消极情感的.

从上述分析可以得出:SRTSM 模型可以较好地提取出微博的主题词,让我们大致了解微博的主题,为微博主题分析提供很好的帮助.

为了更好地分析 SRTSM 的主题提取性能,我们进一步引入 KL 散度对 SRTSM 与 LDA 进行实验比较分析(如图 2 所示).

$$KL=\frac{1}{k}\sum_{P,Q\in DT}\sum_{w\in W}P(w)\log\frac{P(w)}{Q(w)} \tag{13}$$

其中,k 为指定主题数,DT 为算法从数据集中提取的主题集合,$P(w)$与 $Q(w)$指不同主题中单词 w 的出现概率.KL 值越大,表示所提取主题之间的相似度越低.

从图 3 中可以看出:与 LDA 相比较,SRTSM 提取的主题集合具有更高 KL 值.这说明由 SRTSM 从同一个数据集中提取的不同主题具有更大的区分度,从而能更好地为微博话题发现服务.当然,SRTSM 在不同数据集上表现出的这种优势存在着差异,这种优势可以从 SRTSM 的主题提取过程得到解释,即:SRTSM 的主题提取与微博情感识别是协同进行的.

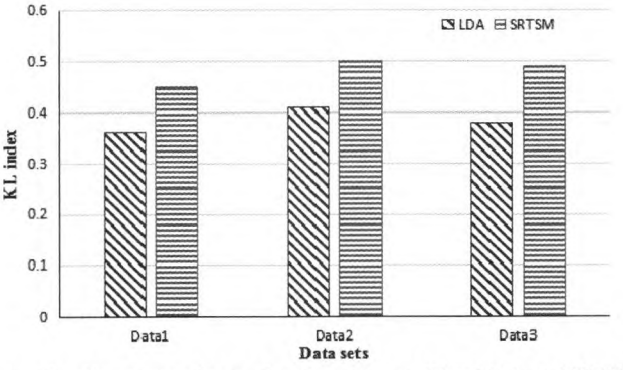


Fig.3 Quality comparison of topics extracted by LDA and SRTSM

图 3 提取主题的 LDA 和 SRTSM 质量比较

3.5 时间效率分析

为了评价 SRTSM 模型的时间性能,我们进一步将 SRTSM 与上述 3 个无监督情感分析算法(JST,SLDA 和 DPLDA)进行比较分析,实验结果如图 4 所示.

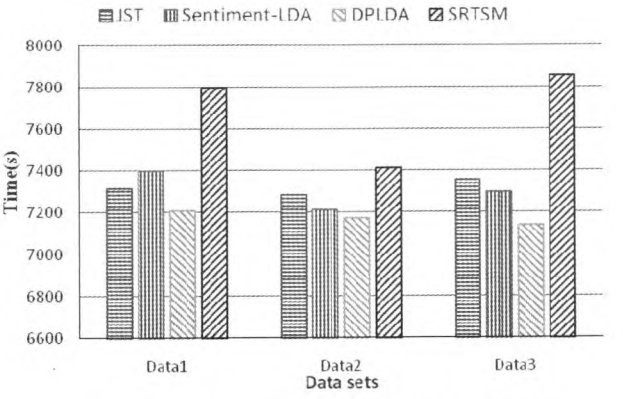


Fig.4 Time performance analysis of SRTSM

图 4 SRTSM 的时间性能分析

从图 4 可以看出:在 3 个数据集上,SRTSM 的耗时均高于其他 3 种算法.结合 SRTSM 的算法流程可以不难解释这种统计现象:由于我们在比较相关算法时都是以迭代次数作为算法循环终止条件,且 4 种算法的迭代次数相等(均为 1 000),从而算法耗时总量可以归结到算法每一次迭代的耗时;而与 JST,SLDA 和 DPLDA 比较,在对微博情感推理过程中,SRTSM 不仅考虑微博文本信息,而且还将微博用户间的链接关系纳入其内,这必然会导致其时间代价的增加.进一步比较 SRTSM 在 3 个不同数据集的时间性能表现可以发现,SRTSM 的耗时从 Data2 到 Data1 再到 Data3 是逐步减少的.而这恰好与表 2 中的微博用户数是相一致的,从而进一步验证了上面的解释.鉴于当前 SRTSM 的时间性能表现欠佳,我们将改进情感推理机制以提高计算效率作为未来的研究工作之一.

4 结束语

随着微博服务的广泛普及,人们可以在微博平台上针对现实各种事件进行所见分享与所感交流,挖掘隐藏在海量微博消息中的主题情感能够有效辅助用户个体、企业组织与政府机构等的决策行为.针对传统无监督的主题情感分析模型的不足,本文提出了一个新的基于微博用户社交关系的主题情感分析模型 SRTSM.实验结果表明:SRTSM 不仅能实现微博消息的主题情感同步检测与分析,而且具有比现有典型无监督主题情感模型(JST, Sentiment-LDA 和 DPLDA)更优的情感分类能力.

将来的工作将在以下几个方面进行:首先,我们将微博用户的档案属性信息纳入微博消息情感极性与主题推理以提升情感分类的准确率;其次,微博消息具有实时特征,我们将对微博消息的情感主题动态演化模式进行分析;最后,改进 SRTSM 的推理机制以提升算法时间效率.

References:

[1] Tang H, Tan S, Cheng X. A survey on sentiment detection of reviews. Expert System with Applications, 2009,36(7):10760−10773.

[2] Missen MMS, Boughanem M, Cabanac G. Opinion mining: Reviewed from word to document level. Social Network Analysis and Mining, 2013,3(1):107−125.

[3] Liu B. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 2012,5(1):1−167.

[4] Zhao YY, Qin B, Liu T. Sentiment analysis. Ruan Jian Xue Bao/Journal of Software, 2010,21(8):1834−1848 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3832.htm [doi: 10.3724/SP.J.1001.2010.03832]

[5] Zhang L, Qian GQ, Fan WG, Kun H, Li Z. Sentiment analysis based on light reviews. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2790−2807 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4728.htm [doi: 10.13328/j.cnki.jos.004728]

[6] Lin C, He Y, Everson R, Ruger S. Weakly supervised joint sentiment-topic detection from text. IEEE Trans. on Knowledge and Data Engineering, 2012,24(6):1134−1145.

[7] Li F, Huang M, Zhu X. Sentiment analysis with global topics and local dependency. In: Maria F, ed. Proc. of the 24th AAAI Conf. on Artificial Intelligence. Atlanta: AAAI Press, 2010. 1371−1376.

[8] Jo Y, Oh AH. Aspect and sentiment unification model for online review analysis. In: Li H, ed. Proc. of the 4th ACM Int'l Conf. on Web Search and Data Mining. New York: ACM Press, 2011. 815−824.

[9] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003,3:993−1022.

[10] Mei QZ, Ling X, Wondra M, Su H, Zhai CX. Topic sentiment mixture: modeling facets and opinions in Weblogs. In: Patel-Schneider P, ed. Proc. of the 16th Int'l Conf. on World Wide Web. New York: ACM Press, 2007. 171−180.

[11] Titov I, McDonald R. Modeling online reviews with multi-grain topic models. In: Ma WY, *et al.*, eds. Proc. of the 17th Int'l Conf. on World Wide Web. New York: ACM Press, 2008. 111−120.

[12] Dasgupta S, Ng V. Topic-Wise, sentimentwise, or otherwise? Identifying the hidden dimension for unsupervised text classification. In: Koehn P, *et al.*, eds. Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2009. 580−589.

[13] He Y. Latent sentiment model for weakly-supervised cross-lingual sentiment classification. Advances in Information Retrieval, 2011,6611:214−225.

[14] Brody S, Elhadad N. An unsupervised aspect-sentiment model for online reviews. In: Kaplan RM, ed. Proc. of the Human Language Technologies: The 2010 Annual Conf. of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010. 804−812.

[15] Moghaddam S, Ester M. ILDA: Interdependent LDA model for learning latent aspects and their ratings from online product reviews. In: Baeza-Yates R, *et al.*, eds. Proc. of the 34th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2011. 665−674.

[16] Sun Y, Zhou X, Fu W. Unsupervised topic and sentiment unification model for sentiment analysis. Acta Scientiarum Naturalium Universitatis Pekinensis, 2013,49(1):102−108 (in Chinese with English abstract).

[17] Moghaddam S, Ester M. On the design of LDA models for aspect-based opinion mining. In: Lebanon G, ed. Proc. of the 21st ACM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2012. 803−812.

[18] Mukherjee A, Liu B. Aspect extraction through semi-supervised modeling. In: Proc. of the ACL. 2012. 339−348.

[19] Ouyang JH, Liu YH, Li XM, Zhou XT. Multi-Grain sentiment/topic model based on LDA. Acta Electronic Sinica, 2015(9): 1875−1880 (in Chinese with English abstract).

[20] Rao Y, Li Q, Mao X, Liu W. Sentiment topic models for social emotion mining. Information Sciences, 2014,266(5):90−100.

[21] Li F, Wang S, Liu S, Zhang M. Suit: A supervised user-item based topic model for sentiment analysis. In: Proc. of the 28th AAAI Conf. on Artificial Intelligence. AAAI Press, 2014. 1636−1642.

[22] Yang Z, Kotov A, Mohan A, Lu S. Parametric and non-parametric user-aware sentiment topic models. In: Lalmas M, *et al.*, eds. Proc. of the 38th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2015. 413−422.

[23] Zhou W, Jin H, Liu Y. Community discovery and profiling with social messages. In: Agarwal D, *et al.*, eds. Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2012. 388−396.

[24] Hu X, Tang L, Tang J, Liu H. Exploiting social relations for sentiment analysis in microblogging. In: Ferragina P, ed. Proc. of the 6th ACM Int'l Conf. on Web Search and Data Mining. New York: ACM Press, 2013. 537−546.

[25] West R, Paskov HS, Leskovec J, Potts C. Exploiting social network structure for person-to-person sentiment analysis. arXiv preprint arXiv: 1409.2450, 2014.

[26] Wu F, Huang Y, Song Y. Structured microblog sentiment classification via social context regularization. Neurocomputing, 2016, 175:599−609.

[27] Tan C, Lee L, Tang J, Jiang L, Zhou M, Li P. User-Level sentiment analysis incorporating social networks. In: Ghosh J, *et al.*, eds. Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2011. 1397−1405.

[28] Lu T. Semi-Supervised microblog sentiment analysis using social relation and text similarity. In: Proc. of the 2015 Int'l Conf. on Big Data and Smart Computing (BigComp). 2015. 194–201.

[29] Speriosu M, Sudan N, Upadhyay S, Baldridge J. Twitter polarity classification with label propagation over lexical links and the follower graph. In: Proc. of the 1st Workshop on Unsupervised Learning in NLP. Stroudsburg: Association for Computational Linguistics, 2011. 53–63.

[30] Fu MH, Chen LY, Lee KR, Kuo YH. A novel opinion analysis scheme using social relationships on microblog. In: Proc. of the Future Information Technology, Application, and Service. Springer-Verlag, 2012. 687–695.

[31] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: Hajic J, Matsumoto Y, eds. Proc. of the Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2002. 79–86.

[32] Hatfreld E, Cacioppo JT, Rapson RL. Emotional contagion. Current Directions in Psychological Science, 1993,2(3):96–99.

[33] Huang F, Li C, Yuan C, Wang Y, Yao Z. Mining sentiment for Web short texts based on TSCM model. Acta Electronic Sinica, 2016,44(8):1887–1891 (in Chinese with English abstract).

附中文参考文献:

[4] 赵妍妍,秦兵,刘挺.文本情感分析.软件学报,2010,21(8):1834–1848. http://www.jos.org.cn/1000-9825/3832.htm [doi: 10.3724/SP.J.1001.2010.03832]

[5] 张林,钱冠群,樊卫国,华琨,张莉.轻型评论的情感分析研究.软件学报,2014,25(12):2790–2807. http://www.jos.org.cn/1000-9825/4728.htm [doi: 10.13328/j.cnki.jos.004728]

[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.北京大学学报:自然科学版,2013,49(1):102–108.

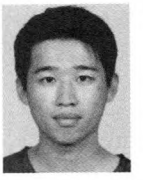
[19] 欧阳继红,刘燕辉,李熙铭,周晓堂.基于 LDA 的多粒度主题情感混合模型.电子学报,2015,43(9):1875–1880.

[33] 黄发良,李超雄,元昌安,汪焱,姚志强.基于 TSCM 模型的网络短文本情感挖掘.电子学报,2016,44(8):1887–1891.

黄发良(1975－),男,湖南永州人,博士,副教授,主要研究领域为数据挖掘,智能计算.

于戈(1962－),男,博士,教授,博士生导师,CCF 会士,主要研究领域为数据库理论和技术,分布与并行系统.

张继连(1978－),男,博士,主要研究领域为空间数据库,信息检索.



李超雄(1991－),男,硕士,主要研究领域为社会媒体处理.

元昌安(1964－),男,博士,教授,主要研究领域为数据挖掘.

卢景丽(1979－),女,博士,主要研究领域为数据挖掘.