文章编号: 1009 - 2552(2019)02 - 0043 - 06 DOI: 10.13274/j.cnki.hdzj.2019.02.010

面向多源社交网络舆情的情感分析算法研究

彭 浩[1]，朱望鹏[2]，赵丹丹[1]，吴松洋[3]

(1. 浙江师范大学计算机科学与工程系，浙江 金华 321004;
2. 上海大学计算机工程与科学学院，上海 201900; 3. 公安部第三研究所，上海 201204)

摘 要: 随着互联网技术的快速发展，社交媒体的多元化也应运而生，因此如何有效分析多源社交网络舆情成为当前研究的热点。文中结合舆情信息的抓取、分词、过滤停用词等三个核心处理模块，基于舆情的情感及趋向性分析，提出了一种面向多源社交网络舆情的情感分析算法。仿真结果表明，该算法在多源社交网络舆情的分析处理中检测效果良好，说明该算法有效。文中的算法研究，可为该领域的进一步研究提供有价值的参考。

关键词: 社交网络; 舆情分析; 情感发现

中图分类号: [TN915.03] **文献标识码**: A

Research on sentiment analysis algorithm for public opinion of multi-source social networks

PENG Hao[1], ZHU Wang-peng[2], ZHAO Dan-dan[1], WU Song-yang[3]

(1. Department of computer science and Engineering, Zhejiang Normal University, Jinhua 321004, Zhejiang Province, China;
2. School of Computer Engineering and Science, Shanghai 201900, China;
3. The Third Research Institute of Ministry of Public Security, Shanghai 201204, China)

Abstract: With the rapid development of Internet technology, the diversification of social media has also emerged. Therefore, how to effectively analyze the multi-source social network public opinion has become a hot topic of current research. This paper combines three core processing modules of lyric information, such as crawling, word segmentation and filtering stop words, proposes a sentiment analysis algorithm for multi-source social networks based on lyric emotion and trend analysisi. The simulation results show that the proposed algorithm performs well in the analysis and processing of multi-source social network public opinion, which indicates that the algorithm is effective, and can provide valuable reference for further research in this field.

Key words: social network; public opinion analysis; emotional discovery

0 引言

随着国内以微信、新浪微博、QQ 空间等为代表的社交网络不断呈现[1]，如何挖掘社交网络背后的舆情信息，越来越多的受到研究人员的广泛关注。许多学者在网络舆情分析方面做了一些研究: 李岩等[2-3]提出短文本聚类及用户评论的情感分析算法，在一定程度上解决了由于关键词的稀疏等带来文本相似度漂移等问题; 卢桃坚等[4]提出了一种基于图的半监督学习优化算法，为人工情感标注的短文本和原始未标注间的短文本构建了联系; Yu 等[5]提出了一种有向树算法，该算法在识别垃圾邮件文本分析方面比较有效。曾振东等[6]提出了一

收稿日期: 2018 - 11 - 15
基金项目: 国家自然科学基金项目(61602418); 教育部人文社科研究项目(15YJCZH125); 浙江省公益技术研究社会发展项目(2016C33168); 浙江省自然科学基金(LQ16F02-0002); 信息网络安全公安部重点实验室开放课题(C15610); 上海市信息安全综合管理技术研究重点实验室开放课题(AGK2018001)
作者简介: 彭浩(1982 -)，男，副教授，研究方向为社会舆情分析、分布式系统安全等。通讯作者: 吴松洋。

种基于灰色支持向量基的网络舆情分析算法，该算法主要面向单一社交网络面，缺乏对多源社交网络舆情分析算法的研究。莫靖杰等[7-9]提出基于多源社交网络信息融合的舆情分析算法，但缺乏对用户情感的有效性验证与分析。可以看出，上述舆情分析算法中，结合多源社交网络的舆情分析尤其是情感分析的研究工作较少。鉴于此，本文在现有情感分析算法基础上，提出了一种面向多源社交网络舆情分析的算法模型。同时利用已知的社交网络数据源，对本文提出的情感算法模型进行仿真实验，从而对本文提出的情感分析算法进行有效验证。

1 多源社交网络的情感分析算法

1.1 算法分析

作为表达观点和倾向的一种最主要的方式，情感分析在人们生活的各个方面都扮演着重要的角色。大到政府机关对于社会舆情的考察检测，小到社会交往的察言观色，对于情感的分析始终是一个绕不开的话题。随着社交网络和计算技术的不断发展，利用计算机处理网络上的数据，分析和预测情感倾向成为近年来的研究热点，在该领域内新的方法层出不穷，其中文本情感分析是一种十分有效的技术。为了得到准确可靠的情感分析结果，本文设计了面向多源社交网络舆情的情感分析算法，该算法主要依托情感词库和针对文本信息，具体分析算法包含以下三个方面的内容：

①文本抓取算法：微博网络、微信公众号、门户网站，拥有用户基数大，数据量巨大的特点，所以要实现爬虫自动抓取的功能，主要根据关键词搜索的方式获取各个源头的文本数据；

②文本预处理算法：通过爬虫获取的信息，可能包含文字，表情，标点，标签等，内容多样而复杂，所有抓取到的文本要通过一系列的预处理，提取利于情感分析的中文文本，并对处理后的文本进行分词处理；

③情感分析算法：对分词后的文本和词库中的情感词、程度词进行匹配，判断词语是正向、负向还是程度词，进行文本的情感分析。

1.2 算法设计

社交网络以传播广度为主，拥有大量的用户群体，信息传播快，群体用户表达的情感丰富多样。结合社交网络的这些特点，本文给出算法设计框架，如图1所示。该算法包括社交平台的信息抓取(本文主要针对微博、门户网站、微信公众号推文三个源头)、文本预处理、情感分析等主要模块，其中情感分析是本文研究的主要模块，包含了情感词库的建立，情感词、程度词的匹配，情感倾向权重计算、得出情感分析等方面。

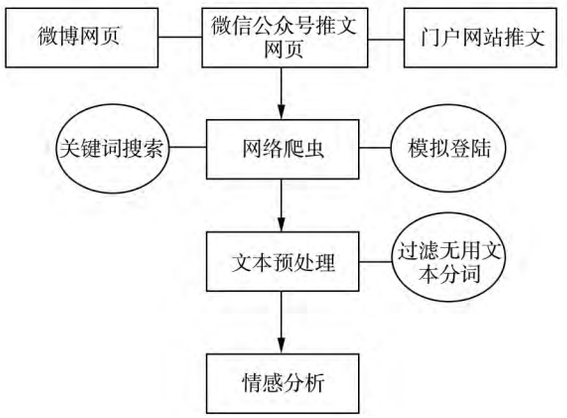

图1 多源社交网络情感分析算法

2 基于多源社交网络舆情的情感分析

2.1 文本获取

通过网络爬虫实现对微博、微信公众号推文、门户网站的文本内容获取。网络爬虫是一种按照一定的规则自地爬取万维网信息的程序或者脚本，传统的爬虫在爬取了初始 url 以后，会从页面中不断的寻找链接，建立一个 url 队列，一直抓取网络上的信息，直到满足用户需求为主。本文结合多源情感分析的需求，从分析更具有针对性，分析效率高，分析结果准确等方面考虑，制定了根据关键词搜索的爬虫，实现对热点话题的准确定位和分析。同时改进后的爬虫能够有效应对网站的反爬技术和网站需要登陆的情况。

以下是文本获取算法的伪代码：

```
1. Begin
//根据关键词搜索
2. construction url with keyword
//设置抓取页面数
3. set min and max page
//如果遇到要先登录的网站先模拟登陆
4. if need login
5. do login
//抓取网页内容
6. spider start
7. getText
//若不用登录网页则直接抓取网页内容
8. else
9. spider start
10. getText
11. End
```

2.2 文本特征提取

通过爬虫获取到的数据，往往夹杂着很多对情感分析无用的数据，比如网络表情、图片、符号等等，所以在通过爬虫得到数据以后，要对文本进行预处理，过滤对情感分析无用的数据，同时还需要对过滤以后的数据进行分词。

以下是文本特征提取算法实现的伪代码:

```
1. Begin
   //读取爬虫抓取的文本内容
2. data = read data spider get
   //对文本进行预处理
3. datafter = handle data
   //对预处理过的文本进行分词
4. cut the datafter
5. End
```

2.3 多源社交网络舆情的情感分析

情感分析，就是根据文本表达的含义和情感信息将文本分为褒扬或者贬义的两种或多种类型，本文主要分为正向和负向两类。情感分析是特殊的文本分类问题，既有一般文本分类的共性问题，也有其特殊性。例如情感信息的表达的隐蔽性，多义性和极性不明显等。

针对上述情况，本文首先建立了完善的情感词库，通过匹配文本和情感词，判断情感的正负性。情感词典包含以下几个部分:

①正向情感词: 该词库收集整理了社交网络中常见的正向情感词，主要针对中文分析;

②负向情感词: 该词库收集整理了社交网络中常用的负向情感词汇，主要针对中文分析;

③程度词词汇: 该词库收集整理了社交网络中常用的程度副词、修饰词等，主要针对中文分析，并根据程度的强弱进行逐级划分，以便更加准确的定位情感。

其次，在拥有了完备的情感词库以后，读取预处理以后的文本，也即经过分词以后的文本，和情感词汇进行比对，判断文本经过分词以后得到的一个个词是属于正负中的哪一面。在判断了词的正负以后，再判断是否有修饰的程度词，并判断程度词的强弱程度。根据强弱把程度词分为 6 个小类别，并分别给情感乘以不同的权重计入情感分析的结果，分别是 2.0，1.5，1.25，0.5，0.25，-1 六个权重，得出最终的情感分析结果。情感分析算法表达式:

单个词的情感权重:

$$poscount = poscount * value \quad (1)$$

$$negcount = negcount * value \quad (2)$$

情感权重数字之和:

$$possum += poscount \quad (3)$$

$$negsum += negcount \quad (4)$$

综合情感权重:

$$totalsum = 1/n * y1 + 1/n * y2 + \cdots + 1/n * ym \quad (5)$$

其中，$y1$、$y2$、$\cdots ym$ 是各个源的情感权重，n 是数据源的数量。

以下是情感分析算法的伪代码:

```
1. Begin
   //读取分析以后的数据
2. get handledata after cut as word
   //判断词汇是否属于积极情感词
3. if word in postivedic
   //如果词汇是积极的，再判断是否含有程度副词
4. than
   //如果有程度副词则按权重计算该词的情感值
5. has degree words or not
6. if has
7. pos_count* value
   //计算总的情感值
8. sum_pos = pos_count + sum_pos
9. elseif word in negitivedic
10. than
11. has degree words or not
12. if has
13. neg_count* value
14. sum_neg = neg_count + sum_neg
15. End
```

3 仿真与分析

3.1 热点事件舆情仿真

本仿真实验，硬件平台是 2.9GHz Intel Core i5 的双核处理器，8GB 2133 MHz LPDDR3 和 64 位 macOS High Sierra 操作系统的 PC。软件平台中，集成的 IDE 环境是 Pycharm，在此基础上基于 Python 语言实现多源社交网络的情感分析。仿真环境中，会对新浪微博、微信公众号推文、门户网站进行网页抓取，并对其内容进行预处理得到情感分析的原始数据，通过情感分析算法得出情感分析结果。

本实验的数据主要通过自行编写的 Python 爬虫获得，爬虫效率可观，能为情感分析提供大量的文本数据; 分词主要依托 Python 中的 jieba 分词库; 情感分析主要依托情感词典和 Python 编写的情感词

的处理算法。为了保证实验的真实、有效和准确性，本文选取了浙江高考改革、王宝强妻子马蓉出轨、鹿晗公布和关晓彤恋情三个事件做仿真，分别抓取了新浪微博、微信公众号 100 篇文章、搜狐新闻做情感分析，同时结合实际，得到以下两类分析结果：

第一类，浙江高考改革，鹿晗公布和关晓彤恋情为第一类，在该类别中，各个源头的数据客观符合实际，分析结果也切合实际。

第二类，王宝强妻子马蓉出轨事件分析为第二类。在该类别的分析中，能够发现个别源头的数据和实际会有所出入，但是整体的分析结果符合事实。在分析该事件的过程中，来自微信公众号的 100 篇文章分析结果为，正向情感略多于负向情感，这显然不符合事件的事实；进一步分析发现，因为王宝强妻子马蓉出轨事件发生的时间在 2016 年，而抓取的微信公众号文章数据大多是 17 年的，也就是在事件发生以后较长一段时间内的文章，这些文章大多是针对马蓉出轨事件做的分析，内容客观，同时情感的指向性不够明确，所以导致在该单源数据下的分析结果和实际有所出入。但是在综合微信、微博和新闻多源的情况下，分析的结果符合客观事实，也充分说明了多源分析的必要性和准确性，同时也论证了本文情感分析算法具有一定的准确性和实用性，各个事件的分析汇总表如表 1 所示。

表 1　各事件分析汇总表

舆情事件	微博		微信		搜狐新闻		综合情感
马蓉出轨事件	1000 条微博	正：2044.5 负：3873.2	100 篇推文	正：3742.6 负：3233.7	200 条新闻	正：1518.4 负：1673.1	正：2435.5 负：2926.7
鹿晗关晓彤公布恋情事件	1000 条微博	正：1716.9 负：1190.2	100 篇推文	正：4109.2 负：2274.6	200 条新闻	正：4466.2 负：2182.7	正：3431.1 负：1882.5
浙江高考改革事件	1000 条微博	正：2110.8 负：729.5	100 篇推文	正：13468.8 负：3842.3	200 条新闻	正：9977.4 负：2586.0	正：8519.3 负：2386.0

上述各个事件的分析结果分别如图 2－4 所示。从图 2 分析结果中可以看出，虽然近年来的浙江高考改革，取消了原本的文理分科制度，同时高考也不再是一次高考，但社会对于此改革的评价是趋向于肯定和正面的态度。另外鹿晗的公布和关晓彤恋情事件，曾一度挤爆新浪微博的服务器，可见其影响之大。从图 3 中可以看出，更多的人对这对恋情持有正向积极的态度，分析的整体情感趋势也是趋于正向，这也符事实；同时也说明了本文设计的面向多源社交网络舆情情感分析算法的准确性。

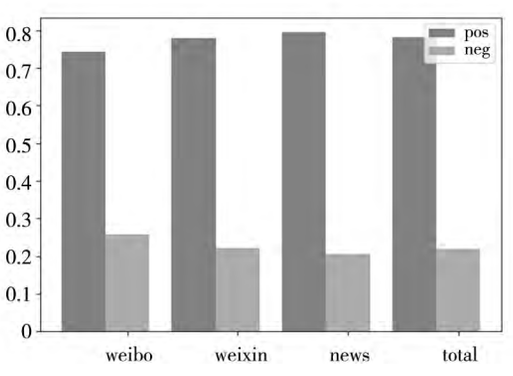

图 2　浙江高考改革分析图

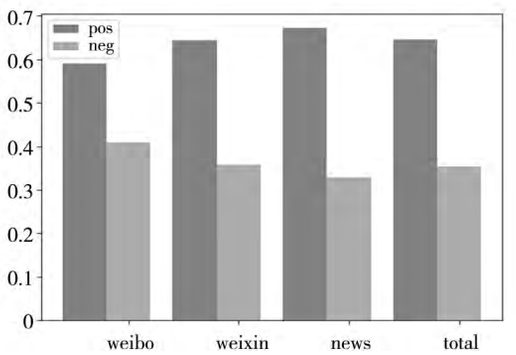

图 3　鹿晗公布和关晓彤恋情事件分析图

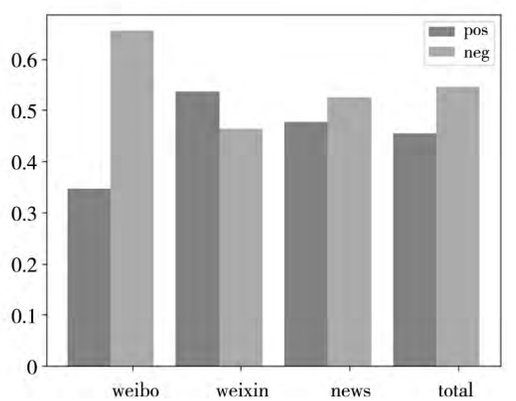

图 4　王宝强妻子马蓉出轨事件分析图

从图 4 和表 2 中可以看出，微博和新闻数据的分析均是偏向于负面情感，这符合客观事实和广大群众的情感指向，但是微信公众号 100 篇文章的结果分析却是正面情感略高于负面情感。针对这一偏差，本文在仿真部分开头做了说明，这也是对结果分

类的一个标准。因为马蓉出轨事件发生的时间距本文做该热点话题的分析较长，文章的实效性比较差，多是事后的客观分析和评价，这是造成单源条件下分析出现偏差的原因，同时也是情感分析算法进一步需解决的问题和改进的一个方向。

表2　王宝强妻子马蓉出轨事件分析表

数据源	数据量	正向情感值	负向情感值
微博	1000	2044.5	3873.2
微信	100	3742.6	3233.7
搜狐新闻	100	1518.4	1673.2
综合情感值	正:7306.5	负:8780.1	

王宝强妻子马蓉出轨的各单源数据下的分析结果分别如图5－7所示。从各单源分析结果来看，虽然有些数据源的数据因为实效等问题在结果上会和实际的情感取向会有所偏差，但是在多源综合分析的结果来看，情感取向符合实际且较准确，这也验证了本文设计的面向多源社交网络舆情的情感分析算法的必要性和该算法的准确性。

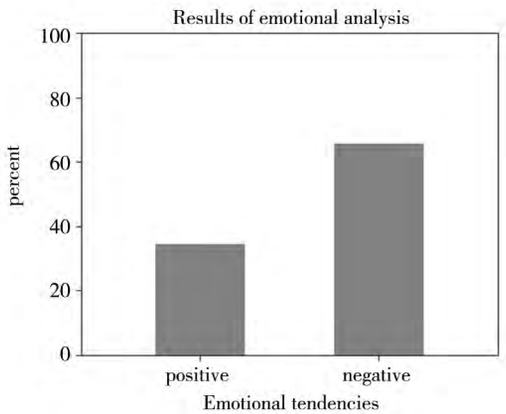

图5　新浪微博文本分析结果图

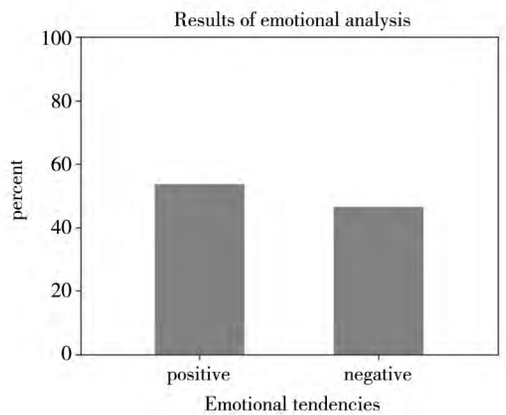

图6　微信公众号100篇文章分析结果图

3.2　理论图解和实际仿真结果

在固定微博1000条数据，正向情感值2110.8，

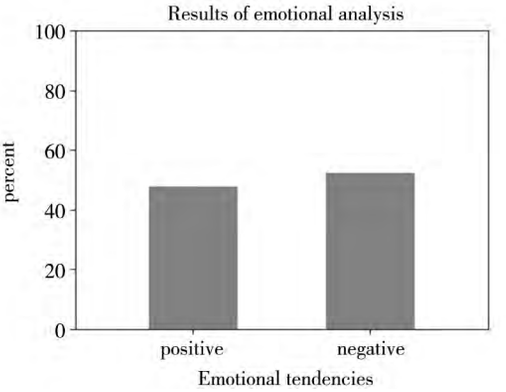

图7　搜狐新闻文本分析结果图

负向情感值为729.5的前提下，变化其它两个源的数据量，研究本实验的理论图解，其结果分别如图8－11所示。

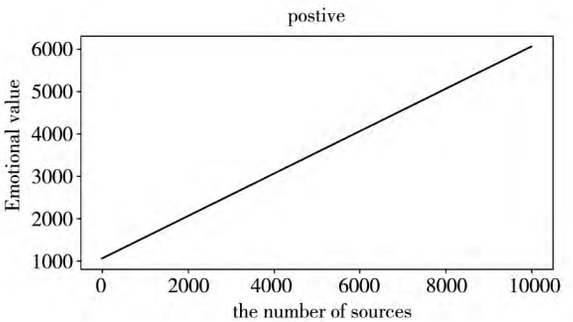

图8　二源情感分析图像(正向)

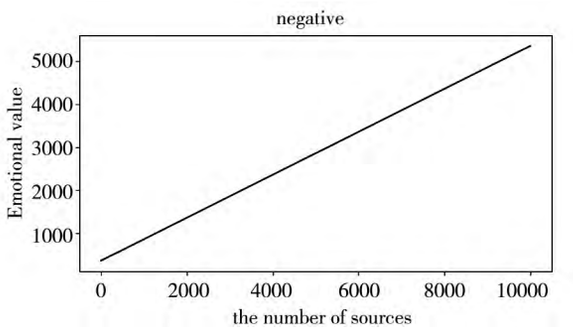

图9　二源情感分析图像(反向)

以浙江高考为例，基于多源数据的情感分析算法可以得到，其正向情感值为2110.8，负向情感值为729.5；在微博、微信、搜狐新闻三源情况下的综合情感值，正向为8519.3，负向情感值为2386.0，综合图像分析理论值为正向8519，负向为2385.93，算法结果与实际结果基本符吻合。因此，基于上述图10－11的分析结果可以表明，面向多源社交网络舆情的情感分析算法，能有效挖掘热门事件背后的隐含舆情，并提高多源社交网络场景下舆情分析结果的新颖性和准确性。

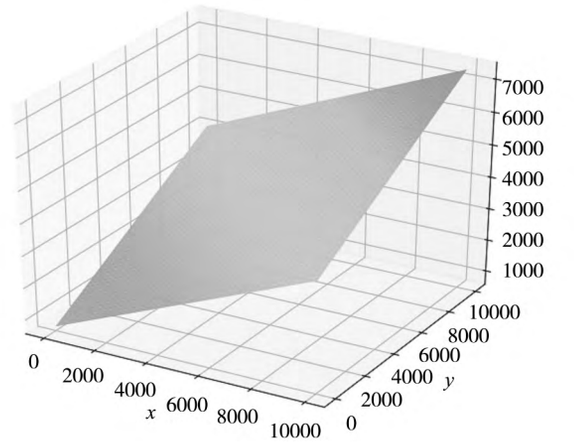

图 10　三源情感分析图像(正向)

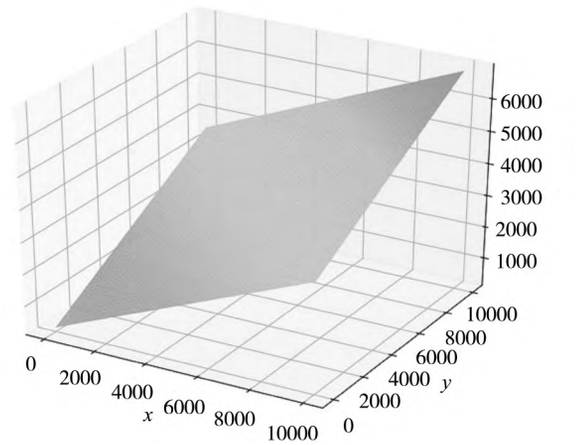

图 11　三源情感分析图像(反向)

4　结束语

本文针对多源社交网络舆情的情感分析算法，进行了一定的分析和研究。该算法基于多源社交网络舆情的文本信息提取，进行舆情的情感分析，并给出相应的舆情发展趋向性分析，从而对多源社交网络的综合舆情情感趋向进行预测。同时本文对多源社交网络中的舆情分析算法给出了实验结果分析和验证，后续的工作将对多源社交网络舆情中情感分析算法的优化进行进一步的研究。

参 考 文 献:

[1] 何凌云，洪良怡，周洁，等. 社交网络隐私安全研究综述 [J]. 信息技术，2018，42(5)：153－159.

[2] 李岩，韩斌，赵剑. 基于短文本及情感分析的微博舆情分析 [J]. 计算机应用与软件，2013(12)：240－243.

[3] 章讯，龙华，周芝民. 基于网络结构改进社交网络好友推荐算法研究 [J]. 信息技术，2017，41(8)：152－156.

[4] 卢桃坚. 社交网络中的短文本情感分析 [D]. 上海: 上海交通大学，2015.

[5] Yu M，Yang W，Wang W，et al. Information Diffusion and Influence Measurement Based on Interaction in Microblogging [M]. Social Media Processing. Springer Berlin Heidelberg，2014：129－140.

[6] 曾振东. 基于灰色支持向量机的网络舆情预测算法 [J]. 计算机应用与软件，2014，31(2)：300－302.

[7] 莫靖杰，屠晨阳，彭佳，等. 基于多源信息融合的社交网络挖掘 [J]. 信息网络安全，2017(9)：73－76.

[8] 张继东，李鹏程. 融合多源数据的移动社交网络用户行为感知研究 [J]. 情报科学，2016(12)：17－21.

[9] 李敏毓，陈兴蜀，尹雅丽，等. 面向多源社交网络的社团结构特征研究 [J]. 工程科学与技术，2017(S2)：195－202.

责任编辑: 丁玥

(上接第 42 页)

[2] CHO Y S，JUNG H K，CHEON C，et al. Adaptive back-projection algorithm based on climb method for microwave imaging [J]. IEEE Transactions on Magnetics，2016，52(3)：9400704.

[3] LIN Jian-he，LV Xiao-lei，LI Rui. Automatic registered back-projection approach based on object orientation for airborne repeat-track interferometric SAR [J]. IEEE Radar，Sonar and Navigation，2018，12(9)：1066－1076.

[4] TULGAR O，ERGIN A A. Improved pencil back-projection method with image segmentation for far-field/near-field SAR imaging and RCS extraction [J]. IEEE Transactions on Antennas and Propagation，2015，63(6)：2572－2584.

[5] 师君. 双基地 SAR 与线阵 SAR 原理及成像技术研究 [D]. 成都: 电子科技大学，2009.

[6] ZHANG Peng，ZHANG Xiao-juan，FANG Guang-you. Comparison of the imaging resolutions of time reversal and back-projection algorithms in EM inverse scattering [J]. IEEE Geoscience and remote sensing letters，2013，10(2)：357－361.

[7] TONG Xiao-hua，YE Zhen，LI Ling-yun，et al. Detection and Estimation of Along-Track Attitude Jitter From Ziyuan-3 Three-Line-Array Images Based on Back-Projection Residuals [J]. IEEE Transactions on Geoscience and Remote Sensing，2017，55(8)：4272－4284.

[8] SHAO Yun-feng，WANG Robert，DENG Yun-kai，et al. LOFFELD Otmar. Fast backprojection algorithm for bistatic SAR imaging [J]. IEEE Geoscience and Remote Sensing Letters，2013，10(5)：1080－1084.

[9] ZHOU Lin，HUANG Chun-lin，SU Yi. A fast back-projection algorithm based on cross correlation for GPR imaging [J]. IEEE Geoscience and Remote Sensing Letters，2012，9(2)：228－232.

[10] ZHANG Guang-min，SONG Yue. Novel imaging method based on cross-correlation function for suppressing the interference of noise [C]. IEEE International Conference on Signal and Image Processing，Beijing，China，2016：251－255.

[11] FO Senglee，Kashyap S. Cross-correlated back projection for UWB radar imaging [C]. IEEE Antennas and Propagation Society Symposium，Monterey，CA，2004：1275－1278.

[12] BROCEA I，PAPANICOLAOU G，TSOGKA C. Theory and applicationsof time reversal and interferometric imaging [J]. Inverse Problems，2003，19(6)：S139－S164.

责任编辑: 丁玥