ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2017/11(01)-0001-23
doi: 10.3778/j.issn.1673-9418.1601037

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-89056056

社会媒体大数据分析研究综述*

杜治娟+,王 硕,王秋月,孟小峰
中国人民大学 信息学院,北京 100872

Survey on Social Media Big Data Analytics*

DU Zhijuan+, WANG Shuo, WANG Qiuyue, MENG Xiaofeng
School of Information, Renmin University of China, Beijing 100872, China
+ Corresponding author: E-mail: nmg-duzhijuan@163.com

DU Zhijuan, WANG Shuo, WANG Qiuyue, et al. Survey on social media big data analytics Journal of Frontiers of Computer Science and Technology, 2017, 11(1): 1-23.

Abstract: Social media, which consists of a large number of meaningful information, is an important way for people to propagate information and express themselves. In recent years, it has become one of the most representative sources of big data. Mining and analyzing the information has profound impact on social development. According to the elements of social media, the current researches are divided into three categories, including analysis based on users, analysis based on relationships and analysis based on interactive contents. Firstly, analyzing user-centered data from user identification based multi-source heterogeneous network, community detection and user influence computing. Secondly, analyzing user relationship strength calculation, information diffusion and influence maximization issues based on interactive relationship-center. Thirdly, discussing feature extraction and selection, the topic or event mining, multimedia data analysis and sentiment analysis issues based on user interactive content analyzing interactive content-centric. Finally, this paper elaborates challenges of mining big data of social media and points out the future work from information diffusion, influence computing, feature extraction and selection, news mining based on Microblog, social media big data fusion and cross-lingual sentiment analysis 6 aspects.

* The National Natural Science Foundation of China under Grant Nos. 61379050, 61532010, 91224008, 61532016 (国家自然科学基金); the National Key R&D Program of China under Grant Nos. 2016YFB1000602, 2016YFB1000603 (国家重点研发计划); the Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant No. 20130004130001 (高等学校博士学科点专项科研基金); the Research Funds of Renmin University under Grant No. 11XNL010 (中国人民大学科学研究基金).
Received 2016-01, Accepted 2016-09.
CNKI网络优先出版: 2016-09-08, http://www.cnki.net/kcms/detail/11.5602.TP.20160908.1045.002.html

Key words: social media; big data; user behavior; interactive relationship; interactive content

摘 要:社会媒体作为人们传播信息和表达观点的重要渠道,包含大量丰富的有用信息,近年来已成为大数据最具代表性的数据来源之一,挖掘与分析这些信息对社会发展影响深远。按照社交媒体的构成要素将目前研究划分为3类,即从基于用户的分析、基于关系的分析和基于交互内容的分析三方面进行总结分析。首先,从多源异构网络中识别用户身份,发现社群并计算用户影响力来分析基于用户的数据;其次,从用户关系强度计算、信息传播和影响力最大化3个角度探讨了基于交互关系为中心的数据分析;然后,基于用户交互内容探讨了特征提取与选择、话题事件挖掘、多媒体数据分析以及情感分析4个问题。最后,从信息传播、影响力计算、特征提取与选择、微博新闻挖掘、社会媒体大数据融合和跨语言情感分析6个方面指出了现有研究的挑战性和未来研究的新视角。

关键词:社交媒体;大数据;用户行为;交互关系;交互内容

文献标志码:A **中图分类号:**TP393

1 引言

近十几年来,在线社会网络越来越流行,如博客,以照片共享为主要功能的Flickr、Facebook、Google+、LinkedIn以及具有强媒体性质的微博等。它们快速增长并允许用户连接、互动、共享和合作,创建了一个新的强大的通信媒体和信息发现、共享平台[1-2]。平均而言[3],Facebook的用户每人每月花7.75小时与朋友进行交流,每天发帖32亿,而Twitter每天发帖3.4亿,Flickr每分钟上传3 000多张照片,博客每年发帖量也超过1.53亿。

社交网络的快速、深度发展使其自身变得越来越庞杂。当前社交网络用户过亿,社交图谱异常庞大,如RenRen社交图谱[4]有75.33万条边、2.74万个可见交互图、24.1万个隐性交互图;用户在不同的社交媒体中持续交互;各种信息在多种社交网络中快速传播。这些特点给社交网络的研究带来巨大挑战。虽然社交网络形形色色,但它们都由用户、关系和内容组成。因此,本文从用户、关系和内容三方面分析现有研究,如图1所示。

从用户层面上看,活跃用户是社交网络的核心,主导整个社交网络的交互。社会媒体中的用户可分为博主、关注对象和粉丝,可以进行发布、关注、转发(RT)、提及(@)、回复和评论操作,并且同一个用户可以参与多个社交网络的互动。因此,以用户为中心的研究主要集中在:(1)从多源异构网络中识别用户身份,判断用户角色,可以借助URL、提及等分析。例如利用URL判断与其他社会网络连接情况[5],使用@提及属性的出入度判定不同角色的用户[6-7]等,对于用户信息的融合非常有用。(2)人以类聚,物以群分,当社交网络中用户在某段时间内互动形成具有稳定群体结构、一致行为特征和统一意识形态后他们就会形成社群[8]。这对于研究人的群体特征、行为规律等非常有用。(3)各行各业都有具有影响力的人物,社交网络中也不例外,用户影响力计算[9]、意见领袖发现[10]在推荐系统、病毒式营销、广告投放、信息传播、专家发现等多个领域广泛应用[11]。

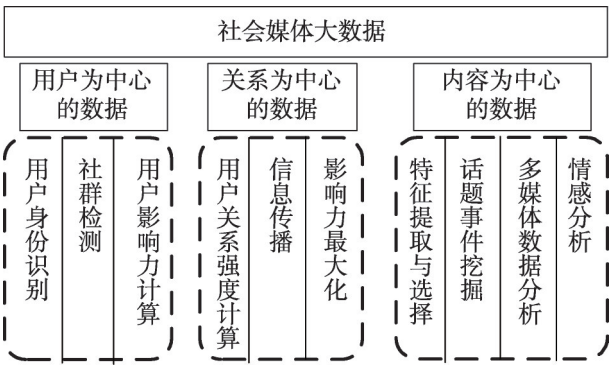


Fig.1 Typical characteristics for social media big data

图1 社会媒体大数据典型特征

从交互关系的层面看,用户之间存在关注关系、传播关系和互惠关系。其中,关注关系由粉丝行为引起,可用于影响力分析[12],关注关系引发了用户的

网络弱关系性和聚类性[13]；传播关系由转播、提及和内嵌的URL引起，具有更强的话题关联性[14]；互惠关系由评论、回复引起，是传播关系的特殊情况。这些研究的基本依据是信息学的传播，它们的价值更多地体现在商业价值和政治价值，比如研究用户及用户群体的传播能力和权威性，可以选取出有传播力、影响力的用户组成初始种子集合，使信息得到最大化的传播；与此同时，各方的利益也将不同程度地得到最大化，利益双方可以从社会网络关系的广度和深度采取不同措施制约对方发展或提升自身利益[15-16]。

从用户交互内容看，用户交互的内容不仅有文本信息，还会包含大量的地理位置、图像和视频等多媒体信息，并且在这些信息中还会包含情感信息。因此，社会媒体的价值体现在：(1)利用位置信息、社会媒体的动态性和时效性分析多媒体数据。(2)从交互内容中分析情感有助于提取不同领域的公众情绪和意见，可以确定民意调查的影响[17]，有效解释和描述政治事件[18]，预测股票趋势[19]等。但是微博讨论的话题不拘泥于任何方式，可变性大，这种互动引发公众情绪的不断变化，挑战性变大。(3)碎片信息的关联与整合，由于海量的不同文化背景的各种思维在交互中相互交融，使原本碎片状的信息以话题事件的方式相关联，进而汇聚为思想流。这种思想流看问题的角度各异，也更能显现出事情的本来面目。但是微博的短文本、多语言背景[20]，以及口语化、错误拼写和缩写、使用特殊符号等对内容的理解造成很大挑战。#标签、转播、提及、URL等可以辅助分析内容[21]。比如利用#标签收集特定话题和事件的信息[5-6]，提高检索性能和进行语义分析[14]等。使用转播估计话题兴趣度或博文重要度[7,20]，提及查找具有特定兴趣的个人或特定话题的视图[22]，使用URL计数度量事件流行度[14]等。

由此可见，社会媒体大数据中潜藏着大量有价值的信息，挖掘过程面临很多挑战。因此，本文第2、3、4章分别基于用户、交互关系和交互内容三方面综述现有研究工作；第5章指出面临的挑战和新问题。

2 基于用户的分析

社会网络中基于用户的研究包括多源异构网络中用户身份识别、社群发现和用户影响力计算。

2.1 用户身份识别

在线社会网络可看做异构信息网络，其中的信息通常包括时间、地点、人物、事件等，而用户往往同时存在于多个不同的社会网络中。由于异构的特点，导致同一个人在不同的网络中会呈现一定的差异，如何在此种情况下识别这个人的身份成为近年来异构社会网络研究的一个热点。文献[23]提出了跨异构社会网络的用户身份识别方法，如图2所示。

用户身份识别主要思想是用户匹配的推理策略，在一对一匹配条件约束下，通过扩展Jaccards系数和扩展Adar度量来对文本内容、空间分布、时间分布等多个特征进行分析。类似的，也可以采用协同分割模型[24]来解决在多个大规模社会网络上处于不同网络中的相同身份的辨识问题。该方法主要利用图论知识，对一个社会网络的拓扑进行平衡化分割，从而在不同的网络中发现相同的分割规律，进而实现身份对齐。文献[25]受力的相互作用和能量守恒原理的启发，提出了基于能量方程的COSNET模型，采用的方法分别是无监督成对网络对齐和传递集成网络对齐的方法，分别从局部一致性和全局一致性两方面来分析异构网络环境下的用户匹配问题。

以上这些都是针对非匿名网络的，实际的匿名网络中用户的身份识别问题也很重要，因此，文献[26]针对匿名社会网络设计了一个无监督的多网络对齐模型，能够解决匿名网络中用户信息和锚链接缺失的问题。总之，以上方法考虑到异构网络的特点，挖掘同一身份在不同网络中的共性，从而完成身份识别。

2.2 社群发现

社群是指用户在某段时间内互动形成的具有稳定群体结构、一致行为特征和统一意识形态的个体和社会关系的集合。社群内部用户关系强度强，聚合强度大，而社群之间用户关系强度弱，离散程度大[27]。社群挖掘的目的在于从用户的行为、群体结构和关系模式中发现潜在的规律。

社群结构按照用户社会关系和对文本内容的兴趣度划分为两种[27]：(1)以用户个体为中心的社群结构。由微博主、粉丝、好友及具有相同#标签或兴趣

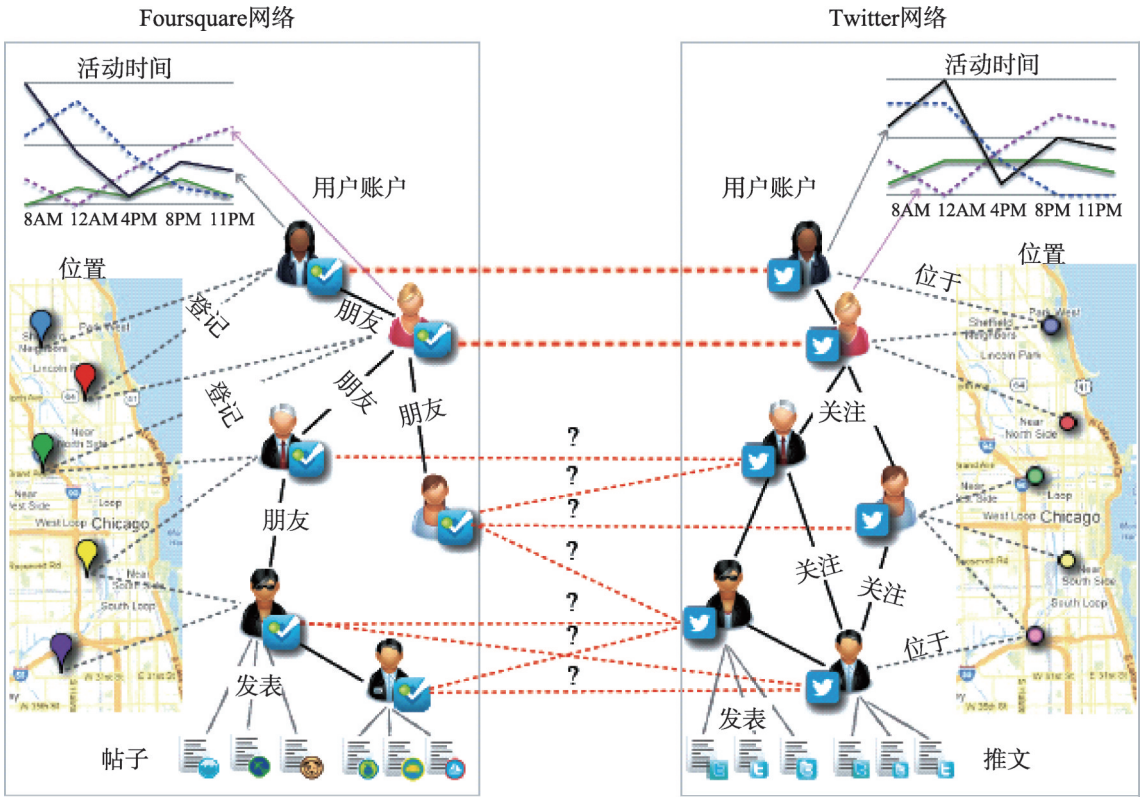


Fig.2 Identification schematic

图2 身份识别示意图

度的用户组成,其主体微博主一般影响力较大,充当意见领袖的角色,其他用户对微博主的某条博文进行评论、转发,这种结构随着微博主的威望或博文热度的降低而减弱。(2)以话题为中心的社群结构,以话题内容为中心,聚合大部分兴趣爱好相同或具有相同#标签的用户,他们讨论的主题大多以时效性较强、关注度较高的热点话题为主,社群成员地位平等,分布均匀,这种结构随着话题的结束而消失。

早期社群划分以静态划分为主,采用基于图聚类的方法和基于相似度计算的方法。基于图聚类的方法采用图建模复杂网络,通过计算节点相似度,按照子网内部节点相似度高,不同子网中节点的连接数最少的原则划分网络,每个子网记为一个社群。大部分算法采用迭代二分的方式寻找二分网络各自的最优化分解以获得满足条件的子图。比较著名的有Kernighan-Lin算法[28]和基于图的Laplace矩阵特征向量的谱二分法[29]。基于相似度计算的方法是根据网络中节点间的相似性或者连接的强弱来决定是否保留或删除边,实现网络群体的重构。GN算法[30]、Newman的快速算法[31]等都是这类方法的代表。此外,用户个体同一时间可能以不同身份出现在不同的社群,因此出现了重叠社群发现[32],后来演变出了动态社群发现[33]。它根据信息资源和网络结构进行动态稳定的变化规律划分,如从分组群和个体两个层次进行动态规划,或者依据当前社群结构约束、历史演变模式和特定时刻单节点的多社群属性进行划分。

2.3 用户影响力计算

影响力计算对单个用户的影响力进行衡量,通常采用节点权重进行表征。目前主要从网络拓扑结构、个体及其关系特征和信息传播结构3个角度来研究,其中,从网络拓扑结构出发的方法如表1所示。

根据网络拓扑结构可以将影响力分为节点的影响力和边的影响力。表1中节点度分为节点的入度[34]、出度[35]和度中心度[36]。其中出入度是有方向的,表示信息传播的方向,出度度量邻居节点对当前节点的影响力,入度反之,而度中心度度量的是当前对邻居

Table 1 User influence measures from network topology

表1 从网络拓扑结构角度度量用户影响力

度量方法	度量指标	影响方式	计算代价	影响范围
节点度[34-36]	当前节点与邻居节点之间的影响力	直接	较小	局部
接近中心度[37]	当前节点对其他节点的影响速度	间接	较大	全局
中介中心度[36]	当前节点对信息传播的重要程度	直接	较大	局部
PageRank算法[38]	当前节点的影响力排名	间接	迭代计算	全局
HITS算法[39]	节点的权威度与中心度	直接	迭代计算	局部
聚集系数[40]	当前节点的邻居节点相互影响的程度	间接	较大	局部

节点的平均影响力。接近中心度[37]表示信息从当前节点传播到其他节点的距离，可以度量当前节点对其他节点的间接影响，也可以度量当前节点自身的关系强度。因此，接近中心度越大，当前节点影响其他节点的速度越快。中介中心度[36]表示信息流经当前节点的数量，值越大，该节点在网络中越重要。PageRank算法[38]计算的是当前节点在网络中的影响力排名，当前节点影响力受其他节点影响力影响时，它们之间呈正相关关系。HITS（hyperlink induced topic search）算法[39]综合考虑了节点的权威度与中心度，但没有考虑节点影响力的划分。聚集系数[40]表示当前与邻居节点产生联系的可能性，具有传递性，可以用来预测形成社区的可能性。

也有文献表明信息传播与网络拓扑结构没有必然联系，单纯基于网络拓扑结构计算影响力不够准确[41-42]。用户行为及其关系特征可以增强计算的准确性[43-45]。如文献[43-44]采用转发、提及等用户行为的传播频率和执行范围有效度量用户发布信息的影响力。信息传播结构主要用到信息传播树，从其规模、深度、广度等方面进行研究。例如，基于用户和信息传播树的话题相似性的影响力排序算法[13]，简单易行，但客观性和准确性欠佳；采用粉丝数、转发数、用户被提及数和PageRank值等来衡量用户影响力[38]；利用产生式图模型来度量Twitter异构网络中的话题影响力[46]。这些研究表明基于不同角度得出的度量结果差别较大，但基于同类度量角度的度量结果较为相似。此外，研究也表明时间因素对影响力计算有重要作用[47-48]。

3 基于关系的分析

社会网络通常采用图表示，节点代表用户集合，有激活和未激活两类状态，节点的权重代表用户影响力；边代表用户关系的集合，边的权重代表用户关系强度。目前主要是根据用户关系研究信息传播，它在调控舆情、产品推广等方面有实用价值。

3.1 用户关系强度计算

用户关系强度用于表征用户之间交互的概率，在微博网络中用边的权重表示。研究表明影响用户关系强度的因素较多，如用户类型和行为[49]、网络结构[50]、微博博文特征和语法特征[51]等。目前典型的计算方法如表2所示。

Table 2 User relationship strength calculation methods

表2 用户关系强度计算方法

度量方法	度量指标	网络结构	影响方式
相似度计算[52-53]	两节点的邻居重叠度	依赖	直接
边介数[30]	经过当前边的最短路径的总和	依赖	直接
影响力图[38]	弧的重数	依赖	直接
文献[54]	期望最大化	依赖	直接
隐含变量模型[55]	描述内容的相似度与用户间的交互关系	依赖	直接
时间模型[44]	指数衰减模型	依赖	直接
HF-NMF[56]	历史交互信息	无关	间接
转移熵[47]	交互信息的演化过程	无关	间接

单纯从网络链接分析的角度看，用户关系强度计算可以分为相似度计算、边介数、影响力图3种。其中，相似度计算[52-53]通常采用Jaccard相似度、Cosine相似度和Overlap相似度；边介数[30]类似于中介中心度，只不过面向的对象是边；影响力图[38]采用有向带权图表示社交网络，弧的方向表示影响力来源，弧的权重表示影响力强度，与弧的重数呈正相关。

从是否考虑时间因素的角度看，用户关系强度

计算可以分为静态模型和时间模型。其中静态模型包括文献[54]所提出的方法、隐含变量模型[55]等。文献[54]根据独立级联模型和真实传播数据将研究问题建模为一个似然函数最大化问题,然后利用期望最大化进行求解。隐含变量模型是根据用户描述内容的相似度与用户间的交互关系计算关系强度。这两种方法计算代价高,不适用于大规模数据集。时间型方法[44]增加了理论时间与实际时间的关联关系,通常采用连续型或离散型指数衰减模型。连续型用户关系强度具有时间动态性,但只能非增量式地计算用户的联合影响力,不适用于大规模数据,为此出现了用离散时间函数近似表示的用户关系强度,它可以增量式地计算用户的联合影响力。

研究表明,即便网络结构上不相关联,只要交互内容上有影响关系,那么这些用户之间就存在间接影响关系。为此文献[56]提出了基于历史交互信息的HF-NMF方法,其中交互信息包括信息条目、用户与信息的关系。文献[47]利用转移熵量化交互信息的演化过程,从而计算用户之间的间接影响力。

3.2 信息传播

信息传播模型研究社会网络中用户对信息的转播和采纳。例如Twitter中,当一个用户转发一条信息,他首先要与信息本身交互,因此初始消息的广播创建了一个新的通知和帖子的级联,这些对象被称为信息级联[57]。传播模型分为意见动态(opinion-dynamic,OD)模型、博弈论(game-theoretical,GT)模型。模型的对比如表3所示。

意见动态模型包括级联模型、阈值模型和传染病模型。级联模型认为只要未激活邻居节点中任意节点 v 以概率 $p_{u,v}$ 激活 u 成功,u 将被激活;否则,v 从此不能再激活 u。$p_{u,v}$ 的取值与 u、v 节点无关,是独立的。线性阈值模型认为当 v 中所有节点的激活能力之和 $\sum_{u\in\Gamma_i(v)}p_{u,v}$ 大于 u 的被激活阈值 θ_v 时,u 将被激活。级联模型、阈值模型中已激活节点不可以向未激活状态转换。也有一些扩展模型,如文献[58]用增量函数 $p_v(u,F_v)$ 代替独立级联模型中的 $p_{u,v}$,F_v 表示被 u 激活失败的邻居节点集合;并且文献[63]发现 $p_v(u,F_v)$ 值随着被激活失败次数的增加而递减。文献[58]还采用阈值函数 $f_u(A_v)$ 代替 $\sum_{u\in\Gamma_i(v)}p_{u,v}$,其中 A_v 表示前一时刻被激活的邻居节点集合。

病毒传播模型认为只要 v 不为空,u 就会以固定概率 p 被感染(激活),并且在一段时间之后,u 可以重新回归易感染(未激活状态)。除此之外,处于免疫状态的节点不会被感染,也不会去感染其他节点。其中 p 取固定值,一般与用户关系无关,只与信息本身有关。例如,在Twitter中传播谣言,如果谣言传播给易感染者,则易感染者会以概率 α 变成已感染者;如果谣言传播给已感染者,则已感染者会以概率 β 变成免疫者;否则邻居节点不再发送该谣言给它时,它以 $1-\beta$ 的概率继续传播该谣言。

意见动态模型的最大特点是需要预定义简单的规则和行为,这样模型失去了动态性和灵活性,且考虑因素单一。因此出现了博弈论模型,它认为节点 u 应该同时考虑所有邻居节点和信息内容,使其自身利益最大化。例如,动态随机最优反应(stochastic

Table 3 Information propagation models

表3 信息传播模型

类型	模型	激活对象	激活条件	邻居因素	信息因素	特点
OD	独立级联模型[58]	邻居节点 v	$p_{u,v}$	单	否	不寻求最大化目标函数,需要预定义简单的规则和行为,仅考虑单一因素
	线性阈值模型[59]	当前节点 u	$\sum_{u\in\Gamma_i(v)}p_{u,v}\geqslant\theta_v$	多	否	
	病毒传播模型[60]	未感染节点	p	无	是	
DT	动态随机最优反应[61]	当前节点 u	$\dfrac{1}{\arg\max\limits_{w\in W}\sum_{t\in V_s}G(w,\eta(t))}$	多	是	自身可获取利益最大化,考虑多种因素
	创新传播[62]	当前节点 u	$p_{i,\beta}(y_i\mid x_{N(i)})$	多	是	

注:u 表示 t 时刻未被激活的当前节点;v 表示 t 时刻未被激活的 u 的邻居节点。

best-response dynamics)模型根据每个动作未来效用的概率分布选择动作[30]；而文献[62]的未来效用 $p_{i,\beta}(y_i|x_{N(i)})$ 根据当前节点与邻居节点的相互作用 $e^{\beta y_i(h_i+\sum_{j\in N(i)}x_j)}\Big/(e^{\beta(h_i+\sum_{j\in N(i)}x_j)}+e^{-\beta(h_i+\sum_{j\in N(i)}x_j)})$ 产生。它与阈值模型的区别在于用户通过衡量自身可选策略的利益大小使其自身利益最大化来选择动作，而不是基于阈值，灵活性强。

上述两大类模型都属于理论型传播模型，它们单纯从理论上模拟信息传播，模型中的时刻都是理论上的时间间隔，并非真实的时间。为此，出现了用户关系强度的计算源于实际数据的传播模型，它们采用信息本身特性、用户关系、微博网络外部因素等多方面对信息传播进程建模，预测信息传播动态以及用户个体的传播行为。主要有两条研究主线：(1)从整体出发，预测信息的扩散速度、范围、广度和深度等[7,42,64]；(2)从个体出发，预测用户个体传播某条信息的概率，进而研究整个社会网络的信息传播情况[7,65]。

3.3 影响力最大化

影响力计算是针对单个用户节点而言的，而影响力最大化问题[66]涉及网络中的多个用户，考量集体的联合影响力，它利用信息传播模型聚集用户，使用户集合可以最大程度地影响其他用户，从而使信息最大程度地扩散。它是在线社交网络的重要研究问题，主要研究可分为传统影响力最大化问题和新型影响力最大化问题。

传统影响力最大化是针对单条信息而言的，主要研究方法包括基于信息传播模型的近似贪心算法、启发式算法和混合算法[58,67-69]，以及这些算法在扩展性上的改进算法[70-73]。贪心算法中基于独立级联模型和线性阈值模型可以避免多个信息同时扩散到某一节点的现象；基于节点度中心度和距离中心度的方法，将信息扩散仅限制在一个局部团体内，无法扩散到整个网络。启发式算法只考虑了网络结构，而没有考虑到信息在网络中扩散的动态性。混合方法，例如采用级联和阈值模型计算影响力，采用贪婪近似启发式方法选择 k 个最优的初始种子达到影响力传播最大化。

新型影响力最大化问题包括竞争性影响力最大化、最低成本影响力最大化和自适应影响力最大化问题。竞争性影响力最大化是针对同时传播的多条相互影响的信息而言的，比如不同品牌或厂家的新品信息、关于某一事件的谣言信息和可信信息等。对于其中每条信息，如何从自身的角度选择初始节点集合使得该信息影响力得到最大化，这个问题称为竞争性信息影响力最大化问题。最早解决这个问题的是文献[61,74-75]，它们证明了以竞争对手的初始种子作为先验知识实现竞争群影响最大化是一个NP难问题和次模(submodular)问题，并设计了两个爬山算法。文献[76]研究了类似的问题，但是其采用的是线性阈值模型和通用阈值模型。文献[15-16]研究了社交网络中广告活动的影响力最大化问题。这些研究两个共同的缺点是假设了两个不太现实的情况：(1)假设当前用户已经选择了种子，就不再感知新竞争对手的存在；(2)假设当前用户感知竞争对手的策略，并且从已经提供免费样品的目标用户中选定种子。针对这种情况，文献[76]提出了基于博弈论的文献[77]新框架，给出了3个更为实际的假设：(1)在竞争网络中给定 r 个影响力已经最大化的分组，每个分组在相同的策略下独立选择 k 个种子；(2)假设每个分组可以感知竞争对手的存在，但不感知他们所采用的策略；(3)假设在影响力传播过程中，一旦当前用户受到一些分组的影响，则他不再受任何其他分组的影响。

最低成本影响力最大化的目的是确定种子用户的最小数目，这些用户能够触发宽级联的信息传播[78]。早期研究限定在单个网络中，但是只考虑单个网络的信息传播会影响计算的准确度，因为一个用户可以处于Twitter、Facebook等多种社交网络，并传播相同的信息。最近，出现了跨多个社交网络的影响力最大化的研究[78]，它采用无损耦合和有耦合模式将多个网络映射到单个网络。无损耦合方案保留原有网络的所有属性，提供高质量的解决方案，而有损耦合方案考虑了运行时间和内存消耗因素。

以上研究都采用了非自适应设置，即营销人员应该选择所有种子用户，给予免费样品等。这样营

销人员被迫仅依赖于传播模型选择所有的种子。如果某些选定的种子表现得不好,就没有机会选择正确的了。为此,文献[79]提出了自适应影响力最大化方法,并给出了两个自适应离线策略MaxSpread和MinTss。MaxSpread给定种子数预算和时间范围,使其最大限度地发挥影响;MinTss给出一个时间范围和受影响的目标用户的预期数量,最小化所需的种子数量。

4 基于内容的分析

文本是社会媒体数据的核心[80],其研究包括文本特征提取与选择、话题挖掘、事件和新闻检测。

4.1 文本特征提取与选择

收集到的原始文本组织松散,直接用于文本分析会影响分析的准确性[81-82]。预处理就是采用特征抽取和特征选择的方法将文档组织成固定数目的预定义类别,典型处理技术如图3所示。

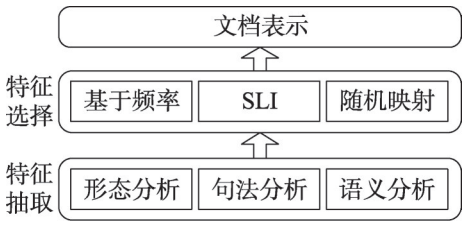


Fig.3 Document preprocessing technology

图3 文档预处理技术

4.1.1 特征抽取

特征抽取方法大概分为3类[81]:形态分析(morphological analysis)、句法分析、语义分析。

形态分析主要是将文档转化为词序列(去除标点符号),包括词语切分(tokenization)、去除停用词、词干还原。词语切分是指将文档去除标点符号并切分成词的序列[83];去除停用词是指去除如“the”、“a”、“or”这种词,主要是为了削减文档包含词的数量来提高文本处理的效率和效果[84];词干还原是指将词还原为词根的形式,如“talking”→“talk”,典型的词根还原算法如Brute-force、Suffix-stripping、Affix-removal和*n*-gram[81]。

句法分析用于分析句子的逻辑语义,典型方法包括词性标注(part-of-speech tagging,POS)和解析法。词性标注就是根据单词在句子中的上下文语法知识为单词添加词汇分类,以便进行语言分析。词性标注的典型技术可分为基于规则的形态分析和随机模型,如隐马尔科夫模型(hidden Markov model,HMM)[85]。HMM是一种随机标记技术,主要用来从输入词序列中发现最类似的POS标记。解析[86]用于检测句子的语法结构,通常采用解析树分析句子的语序。

语义分析就是理解句子的含义,包括关键词识别技术和语义网技术。关键词识别技术用于从文本信息中提取有用内容,通常基于语义词典,如WordNet-Affect,可用于情感分析,但是它依赖于文本中的显示词汇。比如多人在飞机失事中遇难,表达悲伤情绪,但文中没有出现“悲伤”,因此,它检测不出悲伤这种情绪。为了弥补这种缺陷,出现了语义网技术,用于表示概念、事件,以及它们之间的关系[86-87],这种技术利用的是词语的背景信息而非明显的关键字。

4.1.2 特征选择

特征选择是为了消除目标文本中无关和冗余的信息,主要是根据词在文档中的重要性得分选择重要特征[88]。主要分为基于频率的方法、潜在语义索引(latent semantic indexing,LSI)和随机映射,其中最常见的度量方法是基于频率的技术,如TF/IDF,定义如下:

$$TFIDF(t)=TF(d,t)\times IDF(t) \tag{1}$$

$$IDF(t)=\lg\frac{|D|}{DF(t)} \tag{2}$$

其中,D是一组文档;文档频率$DF(t)$代表出现t的文档数。*IDF*用于缩减词条权重,降低词条频繁出现的影响,但是它不适合分析微博数据[89],原因有4个:(1)博文的低冗长使得词频通常接近于0、1,不能体现博文之间的区别[89],并且词频权重仅表示词条在集合中的重要性,而不是它相对于时间的重要性,不足以衡量时间敏感的话题。(2)噪声和稀有词的*IDF*得分更高,而话题词在多个博文中出现说明其更重要,因此减少此类话题词的重要性可能会导致性能下降[90]。(3)博文的高通量性使得计算整体权重不切实际。(4)词频技术不捕获词条的顺序,致使处理过程中信

息丢失。

LSI[91]倾向于提高词汇匹配，随机映射则是通过大的文档集创建映射图。图中任何选定的区域可以用于提取类似主题的新文档。

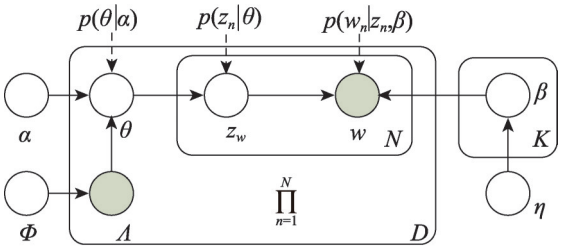
4.2 话题事件挖掘

事件是指在特定的时间和地点下发生的有前因和后果的事情，而话题是指由所有直接相关事件构成的大事件[92]。话题挖掘的主要任务是话题检测与跟踪（topic detection and tracking，TDT），采用历史事件追溯检测和在线新事件自动识别方法[93]，已有大量研究，尤其针对完整新闻报导[93-94]和博客[95]的话题检测已取得了一些成绩。然而，由于微博格式复杂，内容简短，用语不规范等特点，TDT技术不能简单应用到微博[20]。下面从话题模型、话题摘要、话题的检测与跟踪三方面进行介绍。

4.2.1 话题模型

话题模型用于识别文本内容的潜在语义，典型的静态话题模型有：(1)向量空间模型，用向量表示词，计算方便，但缺乏信息的语义关联，并且新词、多义词、别义词对基于第三方词典或者语言资料的词汇链模型挑战性很大。(2)图模型，充分考虑上下文的语义关系，弥补了传统话题模型语义信息缺失的不足，但是在实际应用中存在着计算代价高，存储容量大等问题。(3)概率模型，典型的模型如LDA（latent Dirichlet allocation）[96]，它采用三层贝叶斯的形式表示潜在的话题，具有较好的泛化性，但也不太适合稀疏数据和短文本。因此演变出了针对微博中单一话题的L-LDA（labeled-LDA）模型[97]和Twitter-LDA模型[98]。L-LDA主要考虑了标签（Hashtag），在建模推文排名、为用户推荐任务[97]方面有应用。L-LDA、LDA发现潜在话题的处理流程如图4所示。

Twitter-LDA模型[98]发现单个推文通常是关于单一话题的。此外，各种研究表明推文的短文本特性导致LDA不太适合Twitter。克服这种问题的一个想法是将推文聚合在一起来提供更多的背景知识，可以根据词条按内容[99]、话题[13]或author-topic（AT）模型[100]对推文进行分组。然而研究表明，相比简单的基于词条的方法，直接应用AT模式不会产生显著改



K：话题集；D：文档个数；Λ：标签集
Θ：主题分布；θ：主题分布；w：关键词
η：对称 Dirichlet 先验值；α、β：给定的参数

$$p(\theta, z, w|\alpha, \beta)=p(\theta|\alpha)\prod_{n=1}^{N}p(z_n|\theta)p(w_n|z_n, \beta)$$

Fig.4 Labeled-LDA and its generation process

图4 标记LDA及其生成过程

进，内容聚合的性能比AT模型的聚合更好。

微博话题数据具有流数据特征，随时间不断动态演化，因而出现了话题动态演化方法[101]，主要有：(1)将文档的时间信息作为话题特征的一个指标维度，并基于传统空间向量构建具有动态演变性的话题模型。(2)基于概率话题模型在强度和内容上挖掘话题演化，主要是计算时间信息与话题、文档[102]、词项[103]的后验概率分布。(3)基于已有词项特征与新增词项特征的演变特性挖掘话题演化。方法(1)、(2)在动态更新问题上具有明显不足，并且仅采用均值泛化的思想去增量扩充演化中的话题特征，影响了计算准确度；方法(3)提高话题关联的正确率，有效地解决了话题演化的偏斜问题。

4.2.2 话题摘要

话题摘要旨在从多条博文中自动为相同话题生成摘要，以辅助话题核心语义的理解。微博话题摘要的研究大概分为两类：一类是针对话题事件的摘要；另一类是针对信息检索的摘要。主要研究有：(1)基于抽取的自动摘要，如根据相关性、最大边缘相关度或相关系数和覆盖内容最大化[104]自动摘要，这种方法的缺点是高效的算法摘要的质量差，质量高的算法计算量大。(2)基于理解的自动摘要，这种方法与第一种方法的区别在于摘要内容不完全出自原文，根据语义理解得到摘要意义表示进而生成摘要，如使用隐马尔科夫模型发现Twitter事件的隐状态摘要话题的所有博文[105]。(3)基于关键词选取与序

列化的自动摘要,如提取命名实体、时间、事件短语和类型作为摘要[106],采用图模型序列化词条作为摘要[107]。这些摘要方法主要针对事件话题,针对用户查询的摘要方法并不多见。

目前微博的摘要方法主要针对事件话题,缺乏针对用户查询的摘要方法,这种方法需要根据用户查询和微博的特征选择和组织摘要内容。并且已有成果中摘要大都是词、短语、句子或消息的简单陈列,缺乏对博文之间内在关联关系的考虑,摘要的组织形式和呈现方式欠佳,需要探索合适的微博摘要组织策略和呈现方式。

4.2.3 话题检测与跟踪

话题检测与跟踪包括在线新事件的检测和历史事件追溯[93,108]。在线新事件检测的任务是实时从媒体反馈中识别事件[93];历史事件追溯的任务是从历史积累的文档中识别以前未知的事件[108]。

对于新事件检测研究已久,早期研究采用传统事件检测方法,目前大多是基于特征的方法,如基于突发性、趋势分析等检测新事件。已有文献的对比如表4所示。

文献[109]针对经典事件检测方法在处理海量数据上速度和效率方面的局限性,提出了恒定时间和恒定空间的办法来解决这个问题。主要优势在于,它结合了位置敏感哈希(locality sensitive hashing,LSH)技术和变分推理策略,这种近似技术可以检测未知事件。文献[110]采用增量在线聚类技术从博文中检测相似话题,通过支持向量机(support vector machine,SVM)模型进行分类,利用了时间、社交、局部特征。其中,事件特征用来刻画此词聚类频率的海量性;局部特征捕捉消息聚类中用户的交互关系,如转发、回复、提及等;话题特征描述聚类的局部连贯性,事件聚类围绕中心话题展开。文献[111]提出统一的事件检测、跟踪和摘要的流程,首先采用话题词聚类检测事件;然后将相关事件的跟踪问题转化为二分图匹配问题;最后将已跟踪的时间链做成方便用户理解的摘要。文献[112-113]将问题转化为图划分问题并基于小波分析检测事件。文献[112]用词构建小波信号,而文献[113]用#标签事件作为小波信号。利用#标签共现原理实现事件检测。文献[114]提出了Twevent,一个面向推文的基于片段的事件检测系统。片段是指推文中的一个或连续的多个字。采用不重叠的 k -最近邻图和基于广义对称条件概率(symmetric conditional probability,SCP)的连续文件段的 n -gram技术。它的独特性在于采用突发性片段

Table 4 New event detection method based on burst detection and trend analysis

表4 基于突发性检测和趋势分析的新事件检测方法

事件类型	文献	关键技术	数据来源	有监督	扩展性	实时性	时空特性
一般性事件	[109]	局部敏感哈希LSH	(a)		√		
	[110]	SVM、增量在线聚类/分类	(a)	√	√	√	
	[111]	共现图、分裂的聚类	(c)				
	[112]	图划分、基于小波变换的聚类	(b)				
	[113]	基于小波变换、隐含狄利克雷分布(LDA)	(a)		√	√	
	[114]	对称条件概率(SCP)的*n*-grams、二项分布检测突发性、k 近邻图聚类	(a, d)				
自然灾害事件	[115]	二项式模型检测突发性	(a)			√	√
	[116]	SVM	(b)	√	√	√	√
犯罪和疾病	[117]	分类	(b)	√			√
重大活动	[106]	命名实体分割、条件随机场的学习和推断、事件潜在变量模型分类事件	(a)	√			
突发新闻	[12]	朴素贝叶斯分类、加权词向量、在线聚类	(b)	√			√
争议事件	[104,118]	梯度提升决策树、回归机器学习	(a)	√	√		

注:(a) Twitter; (a,d) Twitter, Wikipedia streaming API; (b) Twitter Search API; (c) Sina Microblog API。

作为事件片段，而不是依赖于突发性词或者话题。文献[116-117]针对特殊事件进行检测。其中文献[116]采用有监督分类技术检测地震、台风、交通事故等事件。而文献[117]采用预定义规则TEDAS对Twitter中与犯罪、疾病相关的事件进行检测。文献[118]提出3种可选择的基于梯度提升决策树的回归及其学习模型，用于检测有争议的事件。文献[104]对文献[118]进行扩展，允许对实体排序。文献[106]提出了TwiGl，它是面向Twitter的开放领域事件抽取和分类系统，从800条随机选择的推文上训练命名实体标注器来抽取命名实体，用已有的Twitter-tuned part-of-speech tagger工具抽取事件提及，然后根据潜在变量模型LinkLDA分类已抽取的事件。文献[12]提出了TwitterStand系统，它利用位置信息和基于加权词向量的聚类算法从博文中自动获取突发新闻。

对于历史事件追溯的研究也有很多，其对比如表5所示。

文献[119]在消息级别，采用条件随机场技术抽取位置等场值，因子图模型捕捉决策之间的交互、变分推理技术来提升海量消息预测的效率和效果。文献[120]利用词频分析和位置共现技术来提高召回率，文献[121]将此方法扩展到不同的社交媒体网络。文献[124]提出ETree系统，它首先采用 n -gram技术将短消息分组到预以连贯的信息块，然后采用增量层次建模技术构建不同粒度的时间主体结构，最后采用时间分析技术识别信息块之间的内在因果关系。文献[125]通过用户标注的标签从Flickr照片中检测事件分3步完成：(1)事件标签检测，使用标签的事件和位置分布信息发现事件相关的标签，并采用小波转换技术减少噪音。(2)事件产生，检测分布模式的特征，聚类事件相关的标签。(3)事件照片识别，根据每个标签，聚类照片相关的事件。

4.2.4 微博新闻检测

目前，微博对新闻业的影响很大，如重大政治事件和紧急情况中记者使用微博和观众互动，跟踪新闻的发展[126]，并可以通过微博发表个人看法，为新闻报道的发展提供了一个额外的语境和对新闻的额外透明度[127]。此外，Twitter多次表明，它是一个新闻媒体[128]。微博新闻话题的检测不同于新闻集成，也不同于传统的话题和趋势检测。新闻集成如谷歌新闻和雅虎新闻注重新闻文章，新闻文章包含丰富的新闻话题和较少的噪音。话题和趋势检测是从博文中识别并合并话题，并不检测话题与真实事件的相关性。但新闻话题源于话题，因此话题检测是微博新闻话题检测的根本任务。此外突发性[129]、话题趋势[130]、信息监测等技术也是必不可少的。目前，微博信息检测跟踪工具如表6[131]所示。

目前也有一些有影响力的微博新闻检测系统，如Eddi，一个互动的Twitter话题浏览系统[132]，但仅适用于单个用户流，不处理来自公众的tweets流。典型的系统还有TwitterStand[12]，如图5所示。

从架构的角度，TwitterStand和TwitInfo有更完整的特征集，包括推文爬取、事件识别和话题识别。TwitterStand[12]使用词频进行在线聚类，寻找话题并定期合并重复聚类，在收集微博新闻方面效果很好，但不完全自动化，其话题检测性能取决于预先选定的

Table 5 Event tracing methods

表5 事件追溯方法

事件类型	文献	关键技术	数据来源	突发性检测	有监督	扩展性	时空性
重大活动	[119]	因子图模型、条件随机场	(a)	√	√		√
计划事件	[120]	基于规则的分类	(b)		√		
	[121]	面向精度、召回率的策略	(c)	√			
有趣话题	[122]	利用前 K 个词做查询扩展	(b)		√		
	[123]	基于词的时间共现关系做时间查询扩展	(a)				
	[124]	层次聚类	(b)				
周期性事件	[125]	离散小波变换、密度聚类	(d)	√		√	√

注：(a) Twitter streaming API; (b) Twitter Search API; (c) Last.fm, EventBrite, LinkedIn, Facebook; (d) Flickr。

Table 6 Existing information discovery and tracking tool for Twitter

表6 现有Twitter信息发现与跟踪工具

服务类型	服务	描述
趋势分析	Twitter Trends	Twitter.com提供排名前10的热点话题或#标签
	TweetStats	绘制用户趋势图的工具
	Trendistic	基于查询关键词的日线图(不再使用)
关键词检测	Monitter	允许用户检测所选择的关键词(不再使用)
用户排序	WeFollow	基于突出成绩的突出用户目录
	Twellow	基于粉丝数量的不同类别的流行用户列表
	JustTweetIt	使用预定义类别的用户目录
数据采集	TwapperKeeper	从Twitter API爬取tweets(不再使用)
	HootSuite	使用关键字和#标签从Twitter API有偿爬取和收集tweets
搜索	Twitter Search	基于查询反向顺序搜索并返回结果
	Google Real Time Search	在原来的Twitter网页内搜索和合并结果(不再使用)

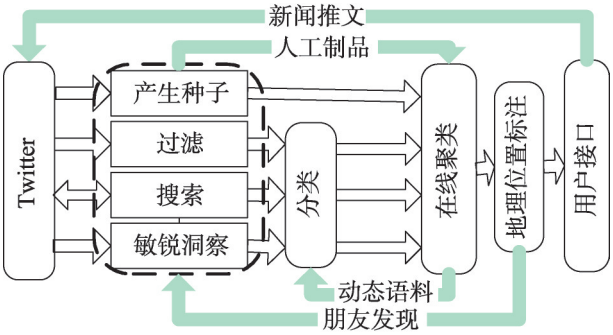


Fig.5 System architecture of TwitterStand

图5 TwitterStand系统架构

种子,并允许用户根据地域浏览新闻,适合于地理特征的应用。TwitInfo[133]旨在帮助用户浏览由用户指定的事件,通过计算时间序列推文频率的峰值检测话题,还可以运用情感分析帮助用户可视化一个事件的关键点。这两种系统都使用TF-IDF权重以减少流行用语的影响。

4.3 多媒体数据的挖掘与分析

多媒体数据源于特定领域的特定问题,主要从位置信息的利用、社会媒体的动态性和时效性以及社会媒体大数据中存在的深层语义三方面进行分析。

4.3.1 地理或位置信息分析

理解和发现人的移动规律在交通管理、城市规划、安全管理等方面尤为重要,近年来出现的GeoSM(地理标记社会媒体)为研究此类问题提供了方便。可以利用社会媒体中的地理标签信息学习人的位置信息,但是人的移动信息往往很分散且移动能力是有限的,为了解决此种标签稀疏的问题,文献[134]采用了以标签信息为参考对用户进行分组的策略,然后通过建立HMMs模型来同时完成分组和移动模型的建立,既可以将移动行为类似的人员分到一组,又能据此预测人的移动趋势,从而兼顾了分组的合理性和地理位置引导的有效性。文献[135]提出了GeoSAGE系统。该系统使用了一个稀疏性附加生成模型来实现空间目标推荐,该模型基于概率混合模型LCA-LDA来建立,能通过收集用户的位置和频度信息得到用户的空间偏好模型;接下来再通过空间金字塔结构把个人的偏好与地理区域的人群偏好结合到一起来挖掘总体的偏好模式,进而达到对用户进行有效推荐的目标。文献[136]针对大气环境检测问题提出利用地理传感器中的时空标签信息来完成时空协同演化模式挖掘的功能。首先对空间传感器的数据通过小波变化进行过滤,而且要在空间信息上附加时间信息,以此得到单个传感器的演化模式;然后通过SCP搜索树这个数据结构来存储收集的这些单个数据;接下来在此结构上通过搜索算法来完成带空间约束的集成协同演化模式发现,从而使人们可以通过大气环境的时空关系变化模式来进行更有效的相关分析和合理的治理。

4.3.2 社会媒体的动态性和时效性分析

最典型的例子就是新闻事件或者说新闻报道,里面不仅会有事件发生的时间和地点,甚至还会产

生很多社会话题和社会影响，这些都会随着时间的推移而发生变化。文献[137]提出了基于在线新闻分析的EKNOT系统，能在给定的时间范围内从Google-News中提取出新闻事件，并从Twitter中找出关于该事件的评论和有关话题，最终可以提供新闻的事件描述、事件发展时间轴、事件所涉及的实体对象和对象间的关系、关于该事件的评论和态度的统计，可以看成一个联动的过程，基本采用自然语言处理的常见方法，但是该系统需要一定的时间来完成整个分析过程。文献[138]将实时分析和地理信息结合到一起，提出了一个GeoBurst系统。它能够通过Tweet流在线地实时发现突发性的本地事件，并提取出带有地理标签的新闻主题；然后对此进行聚类来找到与该主题和地理位置有关的其他突发性话题，从而建立该突发性事件的时间轴来更好地实现舆论分析等任务。该系统还同现有的EvenTweet和Wavelet两个系统进行了比较，实验证明在同样的数据集上GeoBurst系统更能准确地发现突发性事件，换句话说就是对突发性更敏感和准确。文献[139]进一步提出了新闻事件随着时间推移会产生许多不同角度的衍生话题，事件本身的动态性也是多样化的，一个主题下会有很多子主题，因此提出建立一个分层的事件侧面模型来帮助检测事件，并确定事件的主题，当然这主要建立在批处理系统之上。文献[140]也提出了多视角聚类算法，利用两阶段随机游走策略来建立动态的中心主题模型，它能更好地描述主题的变迁过程，而不仅仅是主题的不同侧面。

4.3.3 社会媒体大数据中存在的深层语义挖掘

对新闻、视频、图像等社会媒体进一步挖掘深层语义也越来越受到重视。比如之前的新闻事件或主题发现问题，文献[141]提出了舆论偏见发现模型，不仅仅是发现主题或事件，还要根据该主题或事件的变迁过程来发现新旧状态下是否出现了偏激的倾向。这个问题类似于情感分析，但是需要对情感进行分类，并以此识别出那些不正常或偏激的论点。文献[142]除了采用很多新的自然语言处理技术来完成实体识别和共指消歧之外，还利用了实体和外部知识库来更好地发现用户的偏好，并据此可以估计哪些新闻或事件会更流行。文献[143]也做了同样的事，只不过聚焦的媒介发生了变化，他们更关注电视、电影等包含更多图像的媒体，在浏览数量这个特征之外，提出通过加入额外的诸如图像相关性分析等方法引入深层影响因子分析的机制，进而更好地判断话题的流行程度。文献[144]提出分析人类的行为，在定义了人类行为的一般模型的同时，指出要充分利用各种媒介，包括电影脚本、视频片段、粉丝讨论等多种渠道去挖掘行为模式，从而构建多种人类的行为模式并加以集成构成行为语义框架。

社会媒体中的信息挖掘技术也催生了很多应用，文献[145]利用发信者与信息之间的关系分析来完成垃圾信息检测；文献[146]利用地理位置信息来帮助出租车司机实时规划路线；文献[147]通过YouTube和Twitter相结合来发现病毒式营销模式。总之，研究人员已经开始利用这些社会媒体中的语义信息来帮助人们完善社会服务和构建网络安全，这些应用就不再赘述了。

4.4 情感分析

情感分析也叫意见挖掘，旨在依据意见目标从语料中识别和提取特定主题的属性、要素和隐含的主观信息。意见目标通常称作实体，可以是人物、事件或话题，与要素和子要素相关联，每个要素都有其自己的一套情感属性。微博情感分析可以提取不同领域的公众情绪和意见，可以确定民意调查的影响[17]，有效解释和描述政治事件[18]，预测股票趋势[19]等。各种情感分析技术、高密度的情感承载词和非正式的词（如“coooool”）有助于微博感情的分类[148]。情感分析面临的挑战和已有研究工作在报告[149]和专著[150]中有详细的分析和总结，但是缺乏多维度的情感度量方法，并且微博的多关系特征和话题的演化特性引发了情感的动态演化现象，随着微博数据流的迅速增长，这个问题也需要考虑。

5 研究展望

由上述分析可知，社会媒体已经引起广泛关注，已有一些研究成果，但随着社会的发展，需求的变化，社会媒体大数据挖掘又面临着新的挑战。

(1)信息传播效应捕捉

社会媒体网络中信息传播效应的刻画是一个复杂的问题,它受到信息自身因素、社会因素和网络外部因素的综合影响,并且用户本身的属性与信息本身的属性也相互影响,准确全面地反映信息传播效应已成为关键。这一问题的解决还依赖于影响力、用户关系强度和传播规律。

①用转发数来衡量影响力以及从单个独立的角度研究影响力的方法不能很好地刻画信息传播情况和完全展现用户的影响力,需要将网络的拓扑结构与信息传播树结合使用,不仅要考虑信息传播树的规模,还要着重关注其深度和广度等特征。

②信息传播是一个动态过程,需要捕捉用户关系强度与传播关系的动态规律。目前一般采用理论型传播模型,但是这种模型计算得到的用户关系强度脱离实际,并且存在着理论时间与真实时刻关联的问题。可以考虑从信息传播历史数据挖掘分析用户关系强度,将理论模型和实际数据联通起来体现实际应用价值。并且利用社交媒体数据的群体特征,借助动态社区捕捉信息传播规律。

(2)影响力计算

基于关系分析的一个具有重要商业价值的研究方向是影响力计算和信息传播的最大化问题。其中信息传播的最大化问题的全局最优化被证明是NP难问题,对于大规模的社会网络,目前只能采用一些优化算法获取近似的较优解,并且对于影响力最大化问题目前的最佳解决算法也只处理了百万级规模的社会网络[69]。而目前微博网络节过亿,如何在微博网络中快速计算出固定数量的最有影响力的节点集合还有待进一步探究。

此外,①因为竞争性信息在选择初始节点时有先后顺序,所以不同次序的信息会有不同的选择策略,这也需要考虑。②在线社交网络除了文本数据,还包含大量的图像声音等多媒体信息,它对影响力分析也提出了新挑战。③研究表明,隐式交互图比可见交互图传播信息的速度更快,揭示的关系更重要[4],因此,两种图中的影响力是什么关系,如何量化它们之间的联系有待研究。④话题传播模型多种多样,但用户影响力相对稳定,它们之间如何影响,程度如何还有待探索。⑤对于影响力最大化问题,除了竞争性影响力最大化问题外,最低成本影响力最大化、自适应影响力最大化和多重影响力最大化也是目前有待研究的问题。

(3)特征提取与选择

针对传统数据的特征提取与选择方法已有很多,但是不利于处理低频词和发现新特征,而这种情况在微博数据中大量存在。与词频模型相比,序列模式挖掘保持了词的顺序并可以捕捉潜在的语义,更能解释话题。但是采用模式挖掘的两大挑战是:大量冗余模式的产生和长模式的低支持度问题。冗余模式是任何模式挖掘中不可避免的问题,但是博文中的噪音加剧了这种问题。对于新特征发现问题,尤其针对博文,区分信息新颖性和发现新特征很重要。在信息新颖性区分方面,词性标注、词重叠度和博文语句相似度等方法都发挥着很大作用。此外,目前社交网络中特征提取与选择是针对文本数据而言的,但是社交网络中还包含大量的图像声音等多媒体信息,这些信息又将如何处理也是目前需要考虑的问题,有待进一步研究。

(4)微博新闻挖掘

目前社交网络中新闻检测研究成果很多,但是微博新闻检测仅限于特定的域或事件,仍然缺少针对微博的跨领域新闻话题检测技术和适合微博属性的单独计算模式;另一方面,新闻的第一要义是新,那么如何在线实时处理这种社会化的短文本流?微博新闻信息弥散分布在海量博文中,每个博文仅是大话题的一个小碎片,如何识别新闻话题?如何实时检测新闻事件?新闻话题存在动态演化性,那么如何判断事件的连续性?如何挖掘这种动态的关联演化性?新闻挖掘的核心是话题挖掘,那么如何迅速从海量博文中提取有意义且更容易被理解的微博话题?目前微博用户中移动用户占多数,那么挖掘到的新闻以什么形式呈现?如何设计针对微博的动态新闻集成系统?这些都有待深入研究和探索。另外,传统新闻检测大多针对文本信息,很少考虑多媒体信息对新闻检测的影响,这也有待进一步解决。

(5)社会媒体大数据融合

随着社会网络服务的发展，用户在社交互动中加入了多种服务，并收集了大量的信息。因此，如何整合分布式社会网络，进而对各种社会媒体数据源进行融合，为知识的挖掘提供更好的数据资源已经成为亟待解决的问题。在这个过程中，由于社会媒体的自发性，导致了发布的信息不能保证其真实可靠，这一挑战加大了融合的难度。社会媒体数据的利用价值之一是事件话题挖掘，目前也倾向于采用构建话题知识库方法，将其用作参照物。比如构建缩写的知识库用于缩写词的识别和链接；类似的还可以构建社会媒体常用语知识库，更复杂的可以构建一个话题事件知识库。这也是目前的一个重点研究方面。

(6)跨语言情感分析

挖掘情感是为了体现商业价值，目前大数据向跨语言融合迈进，相应的情感分析也向跨语言情感分析发展。但是，语言的不同体现在语言特征、要素分布的不同，语言间关联的障碍使得跨语言情感分析成为更大的挑战，这是目前亟待解决的问题。

社会媒体大数据有其独特的特性，不仅包含社会关系属性，还包括文本数据、多媒体数据等挖掘价值。研究热点问题很多，本文仅从用户行为、信息传播、文本挖掘、多媒体数据分析4个方面对相关研究成果做了总结、分析和展望。

References:

[1] Khan N, Yaqoob I, Hashem I A T, et al. Big data: survey, technologies, opportunities, and challenges[J]. The Scientific World Journal, 2014: 1-18.

[2] Brewin M W. Media, society, world: social theory and digital media practice[J]. New Media & Society, 2013, 15 (7): 1195-1197.

[3] 216 social media and Internet statistics[EB/OL]. [2015-12-16]. http://thesocialskinny.com/216-social-media-and-internet-statistics-september-2012/.

[4] Saini S, Jin H, Jespersen D, et al. An early performance evaluation of many integrated core architecture based SGI rackable computing system[C]//Proceedings of the 2013 International Conference on High Performance Computing, Networking, Storage and Analysis, Denver, USA, Nov 17-21, 2013. New York: ACM, 2013: 94.

[5] Chang H C. A new perspective on twitter hashtag use: diffusion of innovation theory[J]. Proceedings of the American Society for Information Science and Technology, 2010, 47(1): 1-4.

[6] Bruns A, Burgess J E, Crawford K, et al. # qldfloods and @ QPSMedia: crisis communication on Twitter in the 2011 south east Queensland floods[M]. Brisbane: ARC Centre of Excellence for Creative Industries and Innovation, 2012: 19-23.

[7] Yang Zi, Guo Jingyi, Cai Keke, et al. Understanding retweeting behaviors in social networks[C]//Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Toronto, Canada, Oct 26-30, 2010. New York: ACM, 2010: 1633-1636.

[8] Wang Meng, Wang Chaokun, Yu J X, et al. Community detection in social networks: an in-depth benchmarking study with a procedure-oriented framework[J]. Proceedings of the VLDB Endowment, 2015, 8(10): 998-1009.

[9] Li Dong, Xu Zhiming, Li Sheng, et al. A survey on information diffusion in online social networks[J]. Chinese Journal of Computers, 2014, 37(1): 189-206.

[10] Merton R K. Social theory and social structure[M]. New York: Simon and Schuster, 1968.

[11] Wu Xindong, Li Yi, Li Lei. Influence analysis of online social networks[J]. Chinese Journal of Computers, 2014, 37 (4): 735-752.

[12] Sankaranarayanan J, Samet H, Teitler B E, et al. TwitterStand: news in Tweets[C]//Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, USA, Nov 4-6, 2009. New York: ACM, 2009: 42-51.

[13] Weng Jianshu, Lim E P, Jiang Jing, et al. TwitterRank: finding topic-sensitive influential Twitterers[C]//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, New York, Feb 4-6, 2010. New York: ACM, 2010: 261-270.

[14] Welch M J, Schonfeld U, He D, et al. Topical semantics of twitter link[C]//Proceedings of the 4th International Conference on Web Search and Web Data Mining, Hong Kong, China, Feb 9-12, 2011. New York: ACM, 2011: 327-336.

[15] Budak C, Agrawal D, El Abbadi A. Limiting the spread of misinformation in social networks[C]//Proceedings of the 20th International Conference on World Wide Web, Hyderabad, India, Mar 28-Apr 1, 2011. New York: ACM, 2011: 665-674.

[16] Tsai J, Nguyen T H, Tambe M. Security games for controlling contagion[C]//Proceedings of the 26th AAAI Conference on Artificial Intelligence, Toronto, Canada, Jul 22-26, 2012. Menlo Park, USA: AAAI Press, 2012: 1464-1470.

[17] Tumasjan A, Sprenger T O, Sandner P G, et al. Election forecasts with twitter: how 140 characters reflect the political landscape[J]. Social Science Computer Review, 2011, 29(4): 402-418.

[18] Shamma D A, Kennedy L, Churchill E F. Tweet the debates: understanding community annotation of uncollected sources [C]//Proceedings of the 1st SIGMM Workshop on Social Media, Beijing, Oct 23, 2009. New York: ACM, 2009: 3-10.

[19] Bollen J, Mao Huina, Zeng Xiaojun. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2011, 2(1): 1-8.

[20] Kwak H, Lee C, Park H, et al. What is Twitter, a social network or a news media?[C]//Proceedings of the 19th International Conference on World Wide Web, Raleigh, USA, Apr 26-30, 2010. New York: ACM, 2010: 591-600.

[21] Liao Yang, Moshtaghi M, Han Bo, et al. Mining micro-blogs: opportunities and challenges[M]//Computational Social Networks. London: Springer, 2011: 129-159.

[22] Teevan J, Ramage D, Morris M R. TwitterSearch: a comparison of microblog search and Web search[C]//Proceedings of the 4th ACM International Conference on Web Search and Data Mining, Hong Kong, China, Feb 9-12, 2011. New York: ACM, 2011: 35-44.

[23] Kong Xiangnan, Zhang Jiawei, Yu P S. Inferring anchor links across multiple heterogeneous social networks[C]// Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, San Francisco, USA, Oct 27-Nov 1, 2013. New York: ACM, 2013: 179-188.

[24] Jin Songchang, Zhang Jiawei, Yu P S, et al. Synergistic partitioning in multiple large scale social networks[C]//Proceedings of the 2014 IEEE International Conference on Big Data, Washington, Oct 27-30, 2014. Piscataway, USA: IEEE, 2014: 281-290.

[25] Zhang Yutao, Tang Jie, Yang Zhilin, et al. COSNET: connecting heterogeneous social networks with local and global consistency[C]//Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, Aug 10-13, 2015. New York: ACM, 2015: 1485-1494.

[26] Zhang Jiawei, Yu P S. Multiple anonymized social networks alignment[C]//Proceedings of the 2015 IEEE International Conference on Data Mining, Atlantic, USA, Nov 14-17, 2015. Piscataway, USA: IEEE, 2015: 599-608.

[27] Bhattacharya P, Ghosh S, Kulshrestha J, et al. Deep twitter diving: exploring topical groups in microblogs at scale[C]// Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, Baltimore, USA, Feb 15-19, 2014. New York: ACM, 2014: 197-210.

[28] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs[J]. Bell System Technical Journal, 1970, 49(2): 291-307.

[29] Pothen A, Simon H D, Liou K P. Partitioning sparse matrices with eigenvectors of graphs[J]. SIAM Journal on Matrix Analysis and Applications, 1990, 11(3): 430-452.

[30] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826.

[31] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 2004, 69(6): 66-133.

[32] Jiang Yawen. Community detection in complex networks [D]. Beijing: Beijing Jiaotong University, 2014.

[33] Tantipathananandh C, Berger-Wolf T, Kempe D. A framework for community identification in dynamic social networks[C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, USA, Aug 12-15, 2007. New York: ACM, 2007: 717-726.

[34] Borgatti S P, Everett M G. A graph-theoretic perspective on centrality[J]. Social Networks, 2006, 28(4): 466-484.

[35] Ghosh R, Lerman K. Predicting influential users in online social networks[J]. arXiv:1005.4882, 2010.

[36] Freeman L C. Centrality in social networks conceptual clarification[J]. Social Networks, 1978, 1(3): 215-239.

[37] Sabidussi G. The centrality index of a graph[J]. Psychometrics, 1966, 31(4): 581-603.

[38] Java A, Kolari P, Finin T, et al. Modeling the spread of

influence on the blogosphere[C]//Proceedings of the 15th International World Wide Web Conference, Edinburgh, UK, May 23-26, 2006. New York: ACM, 2006: 22-26.

[39] Awekar A C, Mitra P, Kang J. Selective hypertext induced topic search[C]//Proceedings of the 15th International World Wide Web Conference, Edinburgh, UK, May 23-26, 2006. New York: ACM, 2006: 1023-1024.

[40] Holland P W, Leinhardt S. Transitivity in structural models of small groups[J]. Comparative Group Studies, 1971, 2(2): 107-124.

[41] Kwak H, Lee C, Park H, et al. What is Twitter, a social network or a news media?[C]//Proceedings of the 19th International Conference on World Wide Web, Raleigh, USA, Apr 26-30, 2010. New York: ACM, 2010: 591-600.

[42] Yang J, Leskovec J. Modeling information diffusion in implicit networks[C]//Proceedings of the 2010 IEEE International Conference on Data Mining, Sydney, Australia, Dec 14-17, 2010. Piscataway, USA: IEEE, 2010: 599-608.

[43] Cha M, Haddadi H, Benevenuto F, et al. Measuring user influence in Twitter: the million follower fallacy[C]//Proceedings of the 4th International Conference on Weblogs and Social Media, Washington, May 23-26, 2010: 30.

[44] Goyal A, Bonchi F, Lakshmanan L V S. Learning influence probabilities in social networks[C]//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, New York, Feb 4-6, 2010. New York: ACM, 2010: 241-250.

[45] Tan Chenhao, Tang Jie, Sun Jimeng, et al. Social action tracking via noise tolerant time-varying factor graphs[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, Jul 25-28, 2010. New York: ACM, 2010: 1049-1058.

[46] Liu Lu, Tang Jie, Han Jiawei, et al. Mining topic-level influence in heterogeneous networks[C]//Proceedings of the 19th ACM Conference on Information and Knowledge Management, Toronto, Canada, Oct 26-30, 2010. New York: ACM, 2010: 199-208.

[47] Ver Steeg G, Galstyan A. Information transfer in social media[C]//Proceedings of the 21st International Conference on World Wide Web, Lyon, France, Apr 16-20, 2012. New York: ACM, 2012: 509-518.

[48] Ver Steeg G, Galstyan A. Information-theoretic measures of influence based on content dynamics[C]//Proceedings of the 6th ACM International Conference on Web Search and Data Mining, Rome, Italy, Feb 4-8, 2013. New York: ACM, 2013: 3-12.

[49] Lumezanu F, Klein H. Measuring the tweeting behavior of propagandists[C]//Proceedings of the 6th International AAAI Conference on Weblogs and Social Media, Dublin, Ireland, Jun 4-7, 2012. Menlo Park, USA: AAAI, 2012: 864-864.

[50] Lerman K, Ghosh R. Information contagion: an empirical study of the spread of news on Digg and Twitter social networks[C]//Proceedings of the 4th International Conference on Weblogs and Social Media, Washington, May 23-26, 2010. Menlo Park, USA: AAAI, 2010: 90-97.

[51] Lehmann J, Gonçalves B, Ramasco J J, et al. Dynamical classes of collective attention in Twitter[C]//Proceedings of the 21st International Conference on World Wide Web, Lyon, France, Apr 16-20, 2012. New York: ACM, 2012: 251-260.

[52] Granovetter M S. The strength of weak ties[J]. American Journal of Sociology, 1972, 36(3): 361-366.

[53] Crandall D, Cosley D, Huttenlocher D, et al. Feedback effects between similarity and social influence in online communities[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, USA, Aug 24-27, 2008. New York: ACM, 2008: 160-168.

[54] Saito K, Nakano R, Kimura M. Prediction of information diffusion probabilities for independent cascade model[C]//LNCS 5179: Proceedings of the 2008 International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Zagreb, Croatia, Sep 3-5, 2008. Berlin, Heidelberg: Springer, 2008: 67-75.

[55] Xiang R, Neville J, Rogati M. Modeling relationship strength in online social networks[C]//Proceedings of the 19th International Conference on World Wide Web, Raleigh, USA, Apr 26-30, 2010. New York: ACM, 2010: 981-990.

[56] Cui Peng, Wang Fei, Liu Shaowei, et al. Who should share what?: item-level social influence prediction for users and posts ranking[C]//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, Beijing, Jul 25-29, 2011. New York: ACM, 2011: 185-194.

[57] Easley D, Kleinberg J. Networks, crowds, and markets: reasoning about a highly connected world[M]. Oxford: Cambridge University Press, 2010.

[58] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network[J]. Theory of Computing, 2015, 11(4): 105-147.

[59] Chen Wei, Yuan Yifei, Zhang Li. Scalable influence maximization in social networks under the linear threshold model[C]//Proceedings of the 2010 IEEE International Conference on Data Mining, Sydney, Australia, Dec 14-17, 2010. Piscataway, USA: IEEE, 2010: 88-97.

[60] May R M, Lloyd A L. Infection dynamics on scale-free networks[J]. Physical Review E, 2001, 64(6): 66-112.

[61] Blume L E. The statistical mechanics of strategic interaction [J]. Games and Economic Behavior, 1993, 5(3): 387-424.

[62] Young H P. The dynamics of social innovation[J]. Proceedings of the National Academy of Sciences, 2011, 108(S4): 21285-21291.

[63] Kempe D, Kleinberg J, Tardos É. Influential nodes in a diffusion model for social networks[C]//LNCS 3580: Proceedings of the 32nd International Colloquium on Automata, Languages, and Programming, Lisbon, Portugal, Jul 11-15, 2005. Berlin, Heidelberg: Springer, 2005: 1127-1138.

[64] Yang J, Counts S. Predicting the speed, scale, and range of information diffusion in Twitter[C]//Proceedings of the 4th International Conference on Weblogs and Social Media, Washington, May 23-26, 2010. Menlo Park, USA: AAAI, 2010: 355-358.

[65] Song X, Chi Y, Hino K, et al. Information flow modeling based on diffusion rate for prediction and ranking[C]//Proceedings of the 16th International Conference on World Wide Web, Banff, Canada, May 8-12, 2007. New York: ACM, 2007: 191-200.

[66] Richardson M, Domingos P. Mining knowledge-sharing sites for viral marketing[C]//Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, Jul 23-26, 2002. New York: ACM, 2002: 61-70.

[67] Leskovec J, Krause A, Guestrin C, et al. Cost-effective outbreak detection in networks[C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, USA, Aug 12-15, 2007. New York: ACM, 2007: 420-429.

[68] Chen Wei, Wang Yajun, Yang Siyu. Efficient influence maximization in social networks[C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, Jun 28-Jul 1, 2009. New York: ACM, 2009: 199-208.

[69] Chen Wei, Wang Chi, Wang Yajun. Scalable influence maximization for prevalent viral marketing in large-scale social networks[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, Jul 25-28, 2010. New York: ACM, 2010: 1029-1038.

[70] Goyal A, Bonchi F, Lakshmanan L V S. A data-based approach to social influence maximization[J]. Proceedings of the VLDB Endowment, 2011, 5(1): 73-84.

[71] Goyal A, Lu W, Lakshmanan L V S. Simpath: an efficient algorithm for influence maximization under the linear threshold model[C]//Proceedings of the 2011 IEEE 11th International Conference on Data Mining, Vancouver, Canada, Dec 11-14, 2011. Piscataway, USA: IEEE, 2011: 211-220.

[72] Kim J, Kim S K, Yu H. Scalable and parallelizable processing of influence maximization for large-scale social networks? [C]//Proceedings of the 2013 IEEE 29th International Conference on Data Engineering, Brisbane, Australia, Apr 8-12, 2013. Piscataway, USA: IEEE, 2013: 266-277.

[73] Li Hui, Bhowmick S S, Sun A. Cinema: conformity-aware greedy algorithm for influence maximization in online social networks[C]//Proceedings of the 16th International Conference on Extending Database Technologys, Genoa, Italy, Mar 18-22, 2013. New York: ACM, 2013: 323-334.

[74] Carnes T, Nagarajan C, Wild S M, et al. Maximizing influence in a competitive social network: a follower's perspective[C]//Proceedings of the 9th International Conference on Electronic Commerce, Minneapolis, USA, Aug 19-22, 2007. New York: ACM, 2007: 351-360.

[75] Bharathi S, Kempe D, Salek M. Competitive influence maximization in social networks[C]//LNCS 4858: Proceedings of the 3rd International Workshop on Web and Internet Economics, San Diego, USA, Dec 12-14, 2007. Berlin, Heidelberg: Springer, 2007: 306-311.

[76] Borodin A, Filmus Y, Oren J. Threshold models for competitive influence in social networks[C]//LNCS 6484: Proceedings of the 6th International Workshop on Internet and Network Economics, Stanford, USA, Dec 13-17, 2010.

Berlin, Heidelberg: Springer, 2010: 539-550.

[77] Li Hui, Bhowmick S S, Cui Jiangtao, et al. Getreal: towards realistic selection of influence maximization strategies in competitive networks[C]//Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Australia, May 31-Jun 4, 2015. New York: ACM, 2015: 1525-1537.

[78] Zhang H, Nguyen D T, Zhang H, et al. Least cost influence maximization across multiple social networks[J]. IEEE/ACM Transactions on Networking, 2016, 24(2): 929-939.

[79] Vaswani S, Lakshmanan L V S. Adaptive influence maximization in social networks: why commit when you can adapt?[J]. arXiv: 1604. 08171, 2016.

[80] Hu Xia, Liu Huan. Text analytics in social media[M]//Mining Text Data. New York: Springer US, 2012: 385-414.

[81] Forman G, Kirshenbaum E. Extremely fast text feature extraction for classification and indexing[C]//Proceedings of the 17th ACM Conference on Information and Knowledge Management, Napa Valley, USA, Oct 26-30, 2008. New York: ACM, 2008: 1221-1230.

[82] Dai Yue, Kakkonen T, Sutinen E. MinEDec: a decision-support model that combines text-mining technologies with two competitive intelligence analysis methods[J]. International Journal of Computer Information Systems and Industrial Management Applications, 2011, 3: 165-173.

[83] Negi P S, Rauthan M M S, Dhami H S. Language model for information retrieval[J]. International Journal of Computer Applications, 2010, 12(7): 13-17.

[84] ChandraShekar B H, Shoba G. Classification of documents using Kohonen's self-organizing map[J]. International Journal of Computer Theory and Engineering, 2009, 1(5): 610-613.

[85] Yuan Lichi. Improvement for the automatic part-of-speech tagging based on hidden Markov model[C]//Proceedings of the 2010 2nd International Conference on Signal Processing Systems, Dalian, China, Jul 5-7, 2010. Piscataway, USA: IEEE, 2010, 1: 744-747.

[86] Ling H S, Bali R, Salam R A. Emotion detection using keywords spotting and semantic network IEEE ICOCI 2006 [C]//Proceedings of the 2006 International Conference on Computing & Informatics, Kuala Lumpur, Malaysia, Jun 6-8, 2006. Piscataway, USA: IEEE, 2006: 1-5.

[87] Li J, Khan S U. MobiSN: semantics-based mobile ad hoc social network framework[C]//Proceedings of the Global Communications Conference, Honolulu, USA, Nov 30-Dec 4, 2009. Piscataway, USA: IEEE, 2009: 1-6.

[88] Hua J, Tembe W D, Dougherty E R. Performance of feature-selection methods in the classification of high-dimension data[J]. Pattern Recognition, 2009, 42(3): 409-424.

[89] Efron M, Organisciak P, Fenlon K. Improving retrieval of short texts through document expansion[C]//Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, Portland, USA, Aug 12-16, 2012. New York: ACM, 2012: 911-920.

[90] Lee C H, Wu C H, Chien T F. BursT: a dynamic term weighting scheme for mining microblogging messages [C]//LNCS 6677: Proceedings of the 8th International Symposium on Neural Networks, Guilin, China, May 29-Jun 1, 2011. Berlin, Heidelberg: Springer, 2011: 548-557.

[91] Yoshida K, Tsuruoka Y, Miyao Y, et al. Ambiguous part-of-speech tagging for improving accuracy and domain portability of syntactic parsers[C]//Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, Jan 6-12, 2007. San Francisco, USA: Morgan Kaufmann, 2007: 1783-1788.

[92] Fiscus J G, Doddington G R. Topic detection and tracking evaluation overview[M]//Topic Detection and Tracking. New York: Springer US, 2002: 17-31.

[93] Allan J, Papka R, Lavrenko V. On-line new event detection and tracking[C]//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, Aug 24-28, 1998, New York: ACM, 1998: 37-45.

[94] Allan J, Harding S, Fisher D, et al. Taking topic detection from evaluation to practice[C]//Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Big Island, USA, Jan 3-6, 2005. Piscataway, USA: IEEE, 2005: 101a.

[95] Sekiguchi Y, Kawashima H, Okuda H, et al. Topic detection from blog documents using users' interests[C]//Proceedings of the 7th International Conference on Mobile Data Management, Nara, Japan, May 9-13, 2006. Piscataway, USA: IEEE, 2006: 108.

[96] Blei D, Ng A, Jordan M. Latent Dirichlet allocation[J]. The Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[97] Ramage D, Dumais S T, Liebling D J. Characterizing microblogs with topic models[C]//Proceedings of the 4th International Conference on Weblogs and Social Media, Washington, May 23-26, 2010. Menlo Park, USA: AAAI, 2010: 130-137.

[98] Zhao W Xin, Jiang Jing, Weng Jianshu, et al. Comparing Twitter and traditional media using topic models[C]// LNCS 6611: Proceedings of the 33rd European Conference on Information Retrieval, Dublin, Ireland, Apr 18-21, 2011. Berlin, Heidelberg: Springer, 2011: 338-349.

[99] Hong L, Davison B D. Empirical study of topic modeling in Twitter[C]//Proceedings of the 1st Workshop on Social Media Analytics, Washington, Jul 25, 2010. New York: ACM, 2010: 80-88.

[100] Steyvers M, Smyth P, Rosen-Zvi M, et al. Probabilistic author-topic models for information discovery[C]//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, USA, Aug 22-25, 2004. New York: ACM, 2004: 306-315.

[101] Zhao Xujian, Yang Chunming, Li Bo, et al. A topic evolution mining algorithm of news text based on feature evolving [J]. Chinese Journal of Computers, 2014, 37(4): 819-832.

[102] Wang X, McCallum A. Topics over time: a non-Markov continuous-time model of topical trends[C]//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA, Aug 20-23, 2006. New York: ACM, 2006: 424-433.

[103] Xu Ge, Wang Houfeng. The development of the topic models in natural language processing[J]. Chinese Journal of Computers, 2011, 34(8): 1423-1436.

[104] Popescu A M, Pennacchiotti M, Paranjpe D. Extracting events and event descriptions from twitter[C]//Proceedings of the 20th International Conference Companion on World Wide Web, Hyderabad, India, Mar 28-Apr 1, 2011. New York: ACM, 2011: 105-106.

[105] Chakrabarti D, Punera K. Event Summarization using Tweets[C]//Proceedings of the 5th International Conference on Weblogs and Social Media, Barcelona, Spain, Jul 17-21, 2011. Menlo Park, USA: AAAI, 2011: 66-73.

[106] Ritter A, Etzioni O, Clark S. Open domain event extraction from Twitter[C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, Aug 12-16, 2012. New York: ACM, 2012: 1104-1112.

[107] Sharifi B, Hutton M A, Kalita J. Summarizing microblogs automatically[C]//Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Los Angeles, USA, Jun 2-4, 2010. Stroudsburg, USA: ACL, 2010: 685-688.

[108] Yang Y, Pierce T, Carbonell J. A study of retrospective and on-line event detection[C]//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, Aug 24-28, 1998, New York: ACM, 1998: 28-36.

[109] Petrović S, Osborne M, Lavrenko V. Streaming first story detection with application to Twitter[C]//Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Los Angeles, USA, Jun 2-4, 2010, Stroudsburg, USA: ACL, 2010: 181-189.

[110] Becker H, Naaman M, Gravano L. Beyond trending topics: real-world event identification on Twitter[J]//Proceedings of the 5th International Conference on Weblogs and Social Media, Barcelona, Spain, Jul 17-21, 2011. Menlo Park, USA: AAAI, 2011: 438-441.

[111] Long Rui, Wang Haofen, Chen Yuqiang, et al. Towards effective event detection, tracking and summarization on microblog data[C]//LNCS 6897: Proceedings of the 12th International Conference on Web-Age Information Management, Wuhan, China, Sep 14-16, 2011. Berlin, Heidelberg: Springer, 2011: 652-663.

[112] Weng J, Lee B S. Event detection in Twitter[C]//Proceedings of the 5th International Conference on Weblogs and Social Media, Barcelona, Spain, Jul 17-21, 2011. Menlo Park, USA: AAAI, 2011: 401-408.

[113] Cordeiro M. Twitter event detection: combining wavelet analysis and topic inference summarization[C]//Proceedings of the 7th Doctoral Symposium in Informatics Engineering, Porto, Portugal, Jan 26-27, 2012. Porto: Faculdade de Engenharia da Universidade do Porto, 2012: 123-138.

[114] Li C, Sun A, Datta A. Twevent: segment-based event detection from Tweets[C]//Proceedings of the 21st ACM International Conference on Information and Knowledge Management, Maui, USA, Oct 29-Nov 2, 2012. New York: ACM, 2012: 155-164.

[115] Robinson B, Power R, Cameron M. A sensitive Twitter

earthquake detector[C]//Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, May 13-17, 2013. New York: ACM, 2013: 999-1002.

[116] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: real-time event detection by social sensors[C]//Proceedings of the 19th International Conference on World Wide Web, Raleigh, USA, Apr 26-30, 2010. New York: ACM, 2010: 851-860.

[117] Li R, Lei K H, Khadiwala R, et al. Tedas: a Twitter-based event detection and analysis system[C]//Proceedings of the 2012 IEEE 28th International Conference on Data Engineering, Washington, Apr 1-5, 2012. Piscataway, USA: IEEE, 2012: 1273-1276.

[118] Popescu A M, Pennacchiotti M. Detecting controversial events from twitter[C]//Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Toronto, Canada, Oct 26-30, 2010. New York: ACM, 2010: 1873-1876.

[119] Benson E, Haghighi A, Barzilay R. Event discovery in social media feeds[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Portland, USA, Jun 19-24, 2011. Stroudsburg，USA: ACL, 2011: 389-398.

[120] Becker H, Chen F, Iter D, et al. Automatic identification and presentation of Twitter content for planned events[C]// Proceedings of the 5th International Conference on Weblogs and Social Media, Barcelona, Spain, Jul 17-21, 2011. Menlo Park, USA: AAAI, 2011: 1-2.

[121] Becker H, Iter D, Naaman M, et al. Identifying content for planned events across social media sites[C]//Proceedings of the 5th ACM International Conference on Web Search and Data Mining, Seattle, USA, Feb 8-12, 2012. New York: ACM, 2012: 533-542.

[122] Massoudi K, Tsagkias M, De Rijke M, et al. Incorporating query expansion and quality indicators in searching microblog posts[C]//LNCS 6611: Advances in Information Retrieval, Proceedings of the 33rd European Conference on IR Research, Dublin, Ireland, Apr 18-21, 2011. Berlin, Heidelberg: Springer, 2011: 362-367.

[123] Metzler D, Cai C, Hovy E. Structured event retrieval over microblog archives[C]//Proceedings of the 2012 Conference of the 9th American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montréal, Canada, Jun 3-8, 2012. Stroudsburg, USA: ACL, 2012: 646-655.

[124] Gu Hansu, Xie Xing, Lv Qin, et al. Etree: effective and efficient event modeling for real-time online social media networks[C]//Proceedings of the 2011 IEEE/WIC/ACM International Conference on Web Intelligence, Lyon, France, Aug 22-27, 2011. Piscataway, USA: IEEE, 2011: 300-307.

[125] Chen L, Roy A. Event detection from Flickr data through wavelet-based spatial analysis[C]//Proceedings of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, China, Nov 2-6, 2009. New York: ACM, 2009: 523-532.

[126] Lee R, Wakamiya S, Sumiya K. Discovery of unusual regional social activities using geo-tagged microblogs[J]. World Wide Web, 2011, 14(4): 321-349.

[127] Hayes A S, Singer J B, Ceppos J. Shifting roles, enduring values: the credible journalist in a digital age[J]. Journal of Mass Media Ethics, 2007, 22(4): 262-279.

[128] Lu Rong, Xu Zhiheng, Zhang Yang, et al. Life activity modeling of news event on Twitter using energy function [C]//Proceedings of the 16th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Kuala Lumpur, Malaysia, May 29-Jun 1, 2012. Berlin, Heidelberg: Springer, 2012: 73-84.

[129] Kleinberg J. Bursty and hierarchical structure in streams[J]. Data Mining and Knowledge Discovery, 2003, 7(4): 373-397.

[130] Kontostathis A, Galitsky L M, Pottenger W M, et al. A survey of emerging trend detection in textual data mining [M]//Survey of Text Mining. New York: Springer, 2004: 185-224.

[131] Lau C H. Detecting news topics from microblogs using sequential pattern mining[D]. Brisbane: Queensland University of Technology, 2014.

[132] Bernstein M S, Suh B, Hong L, et al. Eddi: interactive topic-based browsing of social status streams[C]//Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, New York, Oct 3-6, 2010. New York: ACM, 2010: 303-312.

[133] Marcus A, Bernstein M S, Badar O, et al. Twitinfo: aggregating and visualizing microblogs for event exploration [C]//Proceedings of the 2011 SIGCHI Conference on Hu-

man Factors in Computing Systems, Vancouver, Canada, May 7-12, 2011. New York: ACM, 2011: 227-236.

[134] Zhang Chao, Zhang Keyang, Yuan Quan, et al. GMove: group-level mobility modeling using geo-tagged social media[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, USA, Aug 13-17, 2016. New York: ACM, 2016: 1305-1314.

[135] Wang Weiqing, Yin Hongzhi, Chen Ling, et al. Geo-SAGE: a geographical sparse additive generative model for spatial item recommendation[C]//Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, Aug 10-13, 2015. New York: ACM, 2015: 1255-1264.

[136] Zhang Chao, Zheng Yu, Ma Xiuli, et al. Assembler: efficient discovery of spatial co-evolving patterns in massive geo-sensory data[C]//Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, Aug 10-13, 2015. New York: ACM, 2015: 1415-1424.

[137] Min Li. EKNOT: event knowledge from news and opinions on Twitter[C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, USA, Feb 12-17, 2016. Menlo Park, USA: AAAI, 2016: 4367-4368.

[138] Zhang Chao, Zhou Guangyu, Yuan Quan, et al. GeoBurst: real-time local event detection in geo-tagged Tweet streams [C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy, Jul 17-21, 2016. New York: ACM, 2016: 513-522.

[139] Wang Jingjing, Tong Wenzhu, Yu Hongkun, et al. Mining multi-aspect reflection of news events in Twitter: discovery, linking and presentation[C]//Proceedings of the 2015 IEEE International Conference on Data Mining, Atlantic City, USA, Nov 14-17, 2015. Piscataway, USA: IEEE, 2015: 429-438.

[140] Peng Min, Zhu Jiahui, Li Xuhui, et al. Central topic model for event-oriented topics mining in microblog stream[C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management, Melbourne, Australia, Oct 19-23, 2015. New York: ACM, 2015: 1611-1620.

[141] Lu Haokai, Caverlee J, Niu Wei. BiasWatch: a lightweight system for discovering and tracking topic-sensitive opinion bias in social media[C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management, Melbourne, Australia, Oct 19-23, 2015. New York: ACM, 2015: 213-222.

[142] Prasojo R E, Kacimi M, Nutt W. Entity and aspect extraction for organizing news comments[C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management, Melbourne, Australia, Oct 19-23, 2015. New York: ACM, 2015: 233-242.

[143] Ding Wanying, Shang Yue, Guo Lifan, et al. Video popularity prediction by sentiment propagation via implicit network[C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management, Melbourne, Australia, Oct 19-23, 2015. New York: ACM, 2015: 1621-1630.

[144] Tandon N, de Melo G, De A, et al. Knowlywood: mining activity knowledge from Hollywood narratives[C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management, Melbourne, Australia, Oct 19-23, 2015. New York: ACM, 2015: 223-232.

[145] Wu Fangzhao, Shu Jinyun, Huang Yongfeng, et al. Social spammer and spam message co-detection in microblogging with social context regularization[C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management, Melbourne, Australia, Oct 19-23, 2015. New York: ACM, 2015: 1601-1610.

[146] Qian S, Cao J, Mouël F L, et al. SCRAM: a sharing considered route assignment mechanism for fair taxi route recommendations[C]//Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, Aug 10-13, 2015. New York: ACM, 2015: 955-964.

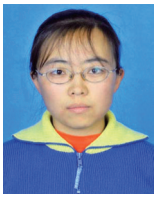
[147] Vallet D, Berkovsky S, Ardon S, et al. Characterizing and predicting viral-and-popular video content[C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management, Melbourne, Australia, Oct 19-23, 2015. New York: ACM, 2015: 1591-1600.

[148] Brody S, Diakopoulos N. Cooooooooooooooollllllllllllll-!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs[C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, Jul 27-31, 2011. Stroudsburg, USA: ACL, 2011: 562-570.

[149] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and Trends in Information Retrieval, 2007, 2 (1/2): 1-135.
[150] Liu Bing. Sentiment analysis and opinion mining[J]. Synthesis Lectures on Human Language Technologies, 2012, 5 (1): 1-167.

附中文参考文献：

[9] 李栋, 徐志明, 李生, 等. 在线社会网络中信息扩散[J]. 计算机学报, 2014, 37(1): 189-206.
[11] 吴信东, 李毅, 李磊. 在线社交网络影响力分析[J]. 计算机学报, 2014, 37(4): 735-752.
[32] 姜雅文. 复杂网络社区发现若干问题研究[D]. 北京: 北京交通大学, 2014.
[101] 赵旭剑, 杨春明, 李波, 等. 一种基于特征演变的新闻话题演化挖掘方法[J]. 计算机学报, 2014, 37(4): 819-832.
[103] 徐戈, 王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报, 2011, 34(8): 1423-1436.



DU Zhijuan was born in 1986. She is a Ph.D. candidate at Renmin University of China, and the member of CCF. Her research interests include Web data management and cloud data management, etc.
杜治娟(1986—),女,中国人民大学博士研究生,CCF学生会员,主要研究领域为Web数据管理,云数据管理等。

WANG Shuo was born in 1981. He is a Ph.D. candidate at Renmin University of China, and a lecturer at Hebei University. His research interests include data fusion and knowledge fusion, machine learning and soft computing, etc.
王硕(1981—),男,中国人民大学博士研究生,河北大学数学与信息科学学院讲师,主要研究领域为数据融合与知识融合,机器学习,软计算等。

WANG Qiuyue was born in 1974. She is an assistant professor at Renmin University of China. Her research interests include database and information systems, information retrieval, large-scale knowledge processing, natural language questions and answers, etc.
王秋月(1974—),女,博士,中国人民大学讲师,主要研究领域为数据库和信息系统,信息检索,大规模知识处理,自然语言问答等。

MENG Xiaofeng was born in 1964. He is a professor and Ph.D. supervisor at Renmin University of China, and the fellow of CCF. His research interests include cloud data management, Web data management, flash-based databases and privacy protection, etc.
孟小峰(1964—),男,中国人民大学教授、博士生导师,CCF会士,主要研究领域为云数据管理,Web数据管理,闪存数据库,隐私管理等。