ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 35,No. 6,Nov. 2020,pp. 1134－1146
DOI:10. 16337/j. 1004-9037. 2020. 06. 013

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

# 多异构社交网络的全局建模及应用例证

王艺霖[1]，仲兆满[2,3]，樊继冬[2]，管　燕[2]

(1. 江苏海洋大学海洋科学与水产学院，连云港，222005;2. 江苏海洋大学计算机工程学院，连云港，222005;3. 江苏省海洋资源开发研究院(连云港)，连云港，222005)

**摘　要**：给出了面向多异构社交网络(Multi-heterogeneous social network，MHSN)的全局表示模型，建立了MHSN用户空间及内容空间的关联模型，为基于MHSN的后续研究提供借鉴。以MHSN的地域突发事件检测为例，论述了基于内容空间的MHSN的融合方法，并以微博和贴吧进行了数据采集和突发事件检测的实验分析；以MHSN的用户兴趣挖掘为例，论述了基于用户空间的MHSN的融合方法，并以微博和贴吧进行了数据采集和用户兴趣挖掘的实验分析。结果表明，本文所提的面向MHSN的突发事件融合检测及用户兴趣融合挖掘方法可以有效地改善突发事件检测和用户兴趣挖掘的效果。

**关键词**：多异构社交网络；内容空间关联；用户空间关联；突发事件融合检测；用户兴趣融合挖掘

**中图分类号**：TP391　　**文献标志码**：A

## Global Modeling and Application Examples of Multi-heterogeneous Social Networks

WANG Yilin[1]，ZHONG Zhaoman[2,3]，FAN Jidong[2]，GUAN Yan[2]

(1. School of Marine Science and Fisheries，Jiangsu Ocean University，Lianyungang，222005，China;2. School of Computer Engineering，Jiangsu Ocean University，Lianyungang，222005，China;3. Jiangsu Academy of Marine Resources Development，Lianyungang，222005，China)

**Abstract**：The global representation model for multi-heterogeneous social networks (MHSN) is presented，and the user space and content space association models of MHSN are established，which can be used as a reference for the follow-up research based on MHSN. Furthermore，taking the detection of localized emergencies in MHSN as an example，this paper discusses the integration method of MHSN based on content space，and conducts experimental analysis of data collection and emergency detection in Weibo and Tieba. Taking the user interest mining of MHSN as an example，this paper discusses the integration method of MHSN based on user space，and conducts experimental analysis on data collection and user interest mining of Weibo and Tieba. The results show that the proposed methods of emergency integration detection and user interest integration mining for MHSN can effectively improve the effect of emergency detection and user interest mining.

**Key words:** multi-heterogeneous social network; content space aligning; user space aligning; emergency integration detection; user interest integration mining

---

**基金项目**：国家自然科学基金(61403156)资助项目；江苏省高校自然科学研究基金(19KJB52004)资助项目；连云港高新区科技计划(ZD201912)资助项目。

**收稿日期**：2020-06-20；**修订日期**：2020-09-01

## 引 言

诸多媒体包含了大量的用户及用户创造的内容,包括Facebook、Twitter、MySpace、LinkedIn、Google+、微博、人人网、论坛、贴吧以及微信等,这类媒体被称为在线社交网络(Online social networks, OSNs)。单个社交网络包含了不同类型的实体以及实体之间建立了不同的关联,是典型的异构社交网络,即网络上的实体或者关系是多类型的。在单异构社交网络的基础上,多个社交网络通过某些实体产生关联,比如用户账户、发表的信息等,这样多个社交网络又建立了更加复杂的网络结构。Bartunov等[1]的研究表明,约有84%的互联网用户拥有多于一个的社交网站账户。2015年,Global Web Index面向50个社交媒体的调研发现,每个人平均拥有5.54个账号,经常活跃在2.82个社交网络上。由于社交网络信息传播性强,具有复杂网络的结构特征,内部蕴含了丰富的潜在有价值信息,近几年引起了学术界和产业界的高度重视。跨多个社交网络的研究可以有效连接不同社交网络的独立异构数据,实现网络的深层融合和数据的综合利用。在多异构社交网络的研究过程中,以用户为中心的分析方法相对充分,尤其是同一自然人在多个社交网络的对齐关联。因为人们更多地关注了用户在多个社交网络的社交圈子、社交行为、生活习惯和兴趣爱好,在兴趣推荐、社区发现以及特殊人员监控等领域有着广泛的应用价值。

## 1 相关工作

### 1.1 单异构社交网络表示模型

异构社交网络是指网络中包含了不同的实体以及实体之间形成了不同的关系。因此,单异构社交网络的表示模型多是围绕网络中的对象及其关系加以描述。根据单异构社交网络表示模型包含的要素个数,可分为二元组、三元组以及多元组等模型。二元组是对社交网络的节点及其关系的直观抽象描述形式。Yang等[2]在研究社交推荐系统的协同过滤时,提出的社交网络模型为有向图$G=(U,F)$,$U$是用户集合,$F$是朋友链接集合。Chen等[3]面向问答型社交网络,将网络描述为一个由用户、问题及类别3种节点,用户之间、用户与问题之间、问题与类别之间3种联系边的异构网络。Seo等[4]定义的异构信息网络为二元组$G=\{V,E\}$,$V$是信息对象,$E$是信息对象之间的关系。

有些研究者对社交网络的节点和边进行了细分,或者为边添加了权重,进而形成了异构社交网络的三元组表示模型。Li等[5]定义社交网络为三元组$SNL=<U,N_{U\times U},P>$,$U$为用户集,$N_{U\times U}\subseteq U\times U$表示用户之间的好友关系集,$P=P_{u_1}\cup,\cdots,\cup P_{u_m}$是用户发表、评论和交互的集合。Tang等[6]将大规模复杂信息网络定义为:$G=(V,E,W)$,$V$代表网站的节点结合,$E$是边的集合,$W$为边的权重,表示关系的强度。齐金山等[7]在文献[6]的基础上,添加了$C$表示所有数据对象的多媒体内容构成,进而定义大规模复杂信息网络为$G=(V,E,W,C)$。Zhu等[8]在度量影响力扩散时,认为社交网络是一个有向二部图$G(V,E,W)$,节点$V=U\cup B$,$U$是用户集合,$B$是用户发表的内容集合;边$E=E_{U\to B}\cup E_{B\to U}$,$E_{U\to B}$用户指向内容的边集合,$E_{B\to U}$为内容提及到用户的边集合;$W$是边的权重。周小平等[9]将社交网络表示为$SN=(U,F,C)$,其中$U$为用户集合,$F$为用户关系集合,$C$为用户创造的内容集合。汪潜等[10]定义一个社交网络为$G=(U,E,A)$,其中$U$为用户集合,$E$代表用户之间的关系集合,$A$为用户的属性集合。Qin等[11]定义异构社交网络为三元组$G=\{X,Y,E\}$,$X$是社交网络的节点集合,$Y$是节点产生的内容集合,$E$是边的集合。琚春华等[12]定义的电商化社交网络包含了用户$U=\{u_1,u_2,\cdots,u_n\}$、好友$F=\{F_1,F_2,\cdots,F_n\}$和用户信用$R=\{r_1,r_2,\cdots,r_n\}$。

针对特定研究目标,一些研究者进一步对社交网络的对象进行了更精细化的描述,由此形成了包含了4个要素以上的多元组表示模型。Vu等[13]在总结了Facebook、Twitter、LinkedIn及Google+等媒

体特点的基础上,定义了社交网络模型的5个主要维度,分别是包含了用户名、描述、城市、E-mail、性别和地点的用户背景,用户之间建立的朋友关系,包含了用户的群组、用户兴趣以及用户发表的帖子。Kundu等[14]提出了模糊粒社交网络的概念FGSN,融合了粒计算理论和模糊邻居系统,将有向的社交网络表示为四元组$S=(C,V,G_{\mathrm{IN}},G_{\mathrm{OUT}})$,其中$V$是网络中的节点,$C\subseteq V$是粒表示的有限集,$G_{\mathrm{IN}}$是入度关系的有限集,$G_{\mathrm{OUT}}$是出度关系的有限集。已有的社交网络表示模型将个体作为活动节点,但FGSN可以从不同的粒度出发重新定义节点,比如将一些个体形成的群体作为活动节点。吴奇等[15]将社交网络描述为五元组$G=<V,E;A,R;\varphi>$,其中$V$是节点集合,$E\subseteq V\times V$是边集合,$A$是节点的种类,是节点$V$经过$\varphi$函数的投影,$R$是边的类型,是边$E$经过$\varphi$函数的投影,$\varphi$是投影函数。仲兆满等[16]面向特定的社交网络——新浪微博,对其进行了细化的描述,给出了九元组表示模型:$\mathrm{MBN}=(U,MB,E_{UMB},E_{MBC},E_{MBF},E_{UU},E_{UForU},E_{UCU},E_{UPU})$,其中,$U$为用户集,$MB$为微博集,$E_{UMB}$、$E_{MBC}$、$E_{MBF}$、$E_{UU}$、$E_{UForU}$、$E_{UCU}$、$E_{UPU}$为用户与用户、微博与微博、用户与微博之间形成的关系集。

### 1.2 多异构社交网络表示模型

由于单个异构社交网络包含的信息量有限,面向多个异构社交网络的融合问题是近期研究的热点。在单一的社交网络的表示模型基础上,已有的融合多个异构社交媒体的研究多是以围绕用户的对齐关联展开的。

Kong等[17]首先提出了以用户为中心的多个社交网络对齐的概念,$g=((G^1,G^2,\cdots,G^n),(A^{1,2},A^{1,3},\cdots,A^{1,n},A^{2,3},\cdots,A^{(n-1),n}))$,其中,$G^i=(V^i,E^i)(i\in\{1,2,\cdots,n\})$是单一的包含了各种类型节点和链接的社交网络,$A^{i,j}$是$G^i$和$G^j$锚链接集合。如果$G^i$和$G^j$的所有用户都存在锚链接,$G^i$和$G^j$是全对齐,否则,$G^i$和$G^j$是部分对齐。现实中的社交网络用户之间多是部分对齐。Zhan等[18]选取了Foursquare和Twitter进行了跨社交媒体的链接预测的研究。在借鉴文献[17]定义的社交网络对齐概念的基础上,将社交网络的节点和边细化为$G=(\{U\cup L\cup W\cup T\},\{E_{u,u}\cup E_{u,l}\cup E_{u,w}\cup E_{u,t}\})$,其中$U$、$L$、$W$和$T$分别是用户集、地点集、文本集和时间戳集,$E_{u,u}$、$E_{u,l}$、$E_{u,w}$和$E_{u,t}$分别为用户链接集、地点链接集、文本链接集和时间戳链接集。通过采集用户在Foursquare主页上的Twitter账号,使得用户在两个平台上的信息对齐。Buccafurri等[19]定义了社交互联网络图为$G=<N,E>$,其中,$N$是节点集合,$E$是边集合。以Facebook、LinkedIn和Twitter为例,给出了多社交网络场景的表示方法,将Facebook、LinkedIn抽象为无向图,而Twitter是有向图。Liu等[20]提出了社交实体连接的概念。假设$P$为现实世界的所有自然人集合,对一个社交网络平台$S$而言,$C_S$是社交网络平台$S$的所有用户集合,社交实体连接定义为$f:C_S\times C_{S'}\mapsto\{0,1\}$。进一步地,提出了通过异构行为模型建立跨平台的用户关联。李国良等[21]开展了多社交网络影响力最大化分析,选取了DBLP、Citeseer、Linkedin和Aminer进行了实验比较。给定$n$个网络$G_1(V_1,E_1)$、$G_2(V_2,E_2)$,$\cdots$,$G_n(V_n,E_n)$,对于每个网络$G_i(V_i,E_i)(0\leqslant i\leqslant n)$,$V_i$表示网络$G_i$的节点集合,$E_i$表示网络$G_i$的边集合,$\mathrm{ET}_i$表示网络$G_i$的实体集合。$V=\bigcup_{i=1}^{n}V_i$,$V$表示所有节点集合,同样,$E$表示所有边集合,$ET$表示所有实体集合。Huang等[22]面向Flickr社交网络研究了社交朋友的推荐,论文提到了多网络协作推荐的方法,但其使用的还是一个社交网络因不同的实体和关系而构建的不同子网,比如联系网、标签网和评论网等。Zhou等[23]将一个社交媒体定义为一个三元组,$\mathrm{SMN}=\{U,C,I\}$,$U$指用户,$C$指连接,$I$指用户之间的交互。$C$和$I$是用户$U$生成的,因此$U$是社交媒体的核心。进一步地,给出了不同社交媒体的用户匹配对的概念$\mathrm{UMP}_{A\sim B}(i,j)$。选取新浪微博和人人网,重点研究了基于用户的朋友网络结构实现不同社交媒体对同一真实用户的对齐关联。Wang等[24]面向两个社交网络$G^x=\{U^x,E^x\}$、$G^y=\{U^y,E^y\}$定义了用户身份链接为$u_i^x\in U^x$和$u_j^y\in U^y$是同一个自

然人,每个社交网络仍然为包含了节点和边的二元组。

Shi等[25]系统地论述了当前异构网络分析的现状和存在的不足,指出需要进一步研究的方向包括不同异构网络信息的融合、实体间关系的清晰梳理、面向不同应用的异构网络挖掘方法等。

### 1.3 存在的问题

已有社交网络表示模型的研究存在的问题概述如下:

(1)对单社交网络而言,表示模型仍然以包含了节点和边的二元组、三元组为主,部分研究者根据不同社交网络的特点,对节点和边进行了一定的细化分析,进而形成了包含4个要素以上的多元组表示模型。已有研究多是面向特定的目标而构建社交网络表示模型,在研究目标的约束下,构建的表示模型多是为特定研究内容服务,没能根据社交媒体具有的宏观和微观特点进一步揭示其包含的各种复杂实体和联系。

(2)对多社交网络的融合而言,同一自然人在不同社交网络的账号对齐关联是研究重点,因此面向多个社交网络构建的表示模型也受限于此。跳出研究目标的约束,系统地梳理不同社交网络的内在本质联系,面向各种类型社交媒体的全局建模方法还没有文献提及。

## 2 多异构社交网络全局建模

基于OSNs的用户空间、内容空间的关联以及不同OSNs之间的分类关系,在理清每个OSN包含的节点及其关系的基础上,给出的多异构社交网络(Multi-heterogeneous social networks, MHSN)的全局表示模型如图1所示。MHSN从纵向和横向两个角度刻画了多个社交网络OSNs的关联关系。显然,用户及内容在不同OSNs的关联与传播,构建了更加复杂的多异构社交网络。多异构社交网络MHSN全局表示模型描述如下:

(1)多异构社交网络表示为MHSN$=(G,R)$,其中$G$表示不同社交网络类OSN和实例osn集合,$R$表示不同OSNs建立关联关系的集合;

(2)最高层OSNs类表示为OSNs$=(US,CS,R_{UU},R_{CC},R_{UC},T)$,以用户空间$US$和内容空间$CS$为实体类型,进而在用户之间、内容之间及用户和内容之间形成了3种关系$R_{UU}$、$R_{CC}$和$R_{UC}$,以时间戳集合$T$刻画OSNs类的动态特性;

(3)不同的OSNs类之间通过继承形成了分类关系,ER$=\{(\mathrm{OSN}_i \ \ \mathrm{Extend} \ \ \mathrm{OSN}_j)|\mathrm{OSN}_i,\mathrm{OSN}_j\in G, i\neq j\}$;

(4)社交网络$\mathrm{OSN}_i,\mathrm{OSN}_j$通过用户的对齐形成了关联关系,$UR=\{(u_1 \ \ \mathrm{Align} \ \ u_2)|u_1\in\mathrm{OSN}_i,u_2\in\mathrm{OSN}_j,i\neq j\}$,$u_1,u_2$是同一自然人在不同社交媒体的账号描述;

(5)社交网络$\mathrm{OSN}_i,\mathrm{OSN}_j$通过内容的对齐形成了关联关系,$CR=\{(c_1 \ \ \mathrm{Align} \ \ c_2)|c_1\in\mathrm{OSN}_i,c_2\in\mathrm{OSN}_j,i\neq j\}$,$c_1,c_2$是同一信息内容在不同社交媒体的呈现描述;

(6)社交网络类$\mathrm{OSN}_i$通过实例化生成具体的社交网络实例$\mathrm{osn}_i$,$OR=\{(\mathrm{osn}_{ij} \ \ \mathrm{Object} \ \ \mathrm{OSN}_i)|\mathrm{OSN}_i \ \mathrm{generates\ object}\ \mathrm{osn}_{ij},\ \mathrm{OSN}_i,\mathrm{osn}_{ij}\in G\}$。

每个社交网络都包含了复杂的实体及其关系。比如,Twitter包含用户和tweets两种实体,用户与tweet之间存在发表、回复、转发和点赞关系,tweets之间可以建立回复和转发关系,用户之间可以直接建立关注关系,并通过tweet建立用户间的回复和转发关系。又如,百度贴吧包括贴吧、帖子和用户实体,用户与帖子之间存在发表、回复和收藏关系,帖子之间可以建立回复关系,用户之间可以直接建立关注关系,并通过帖子建立用户间的回复关系。同一用户在不同社交媒体上有不同的表现形式,但对应的都是同一自然人。基于MHSN用户空间的关联,可以分析多个OSNs上用户的社交行为和影响

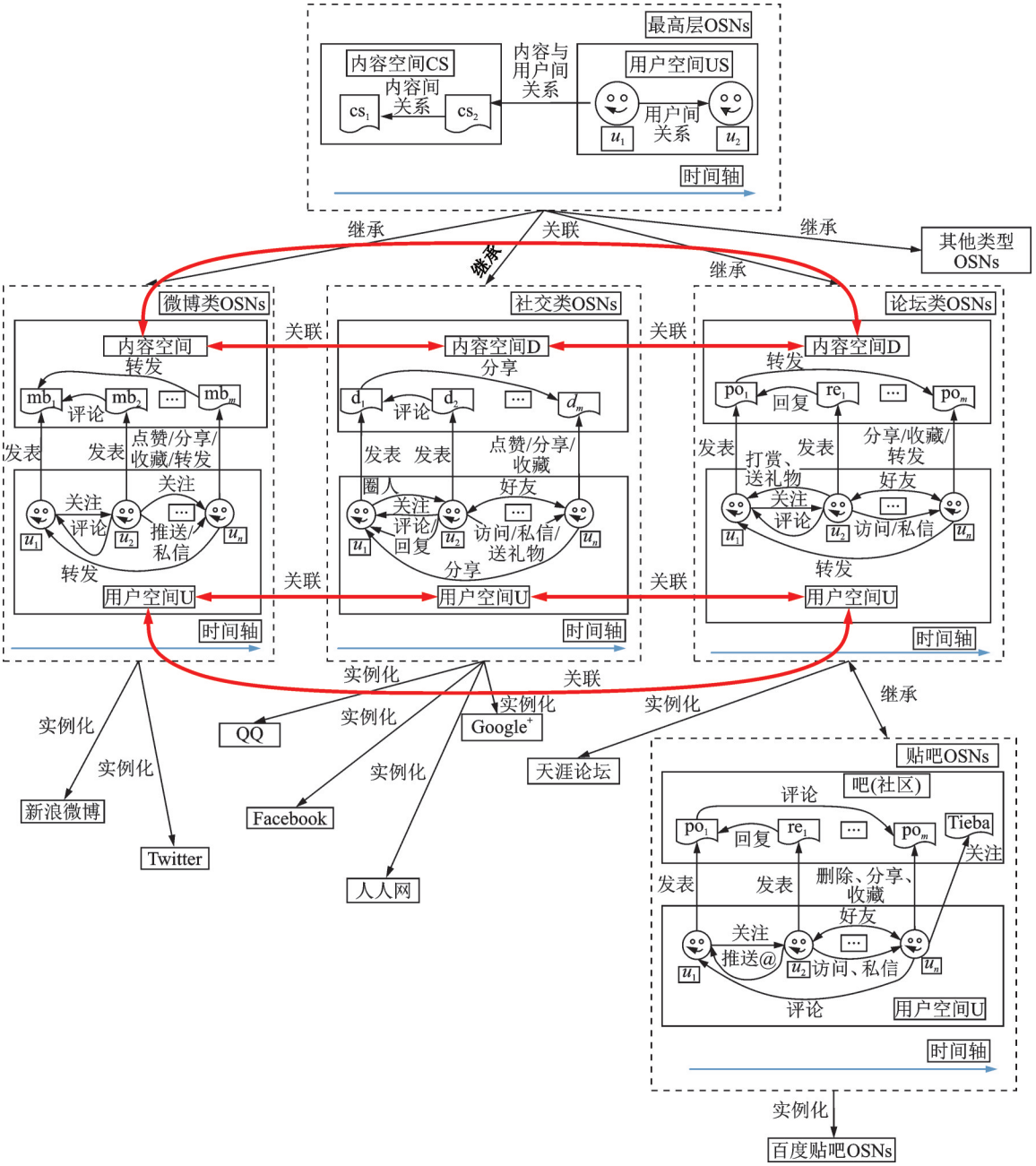



图1 多异构社交网络MHSN全局表示模型

Fig.1 Global representation model of multiple heterogeneous social networks

力,可以进行全面的用户画像描述。图2是同一真实用户在多个不同社交媒体的对齐关联示例。

网络上的内容在不同社交媒体的呈现有两种模式:一种是显式的,指同一篇信息在不同网络上的传播,比如新浪媒体发表的一篇新闻在贴吧、微博中以转发的形式进行传播;另一种是隐式的,指对同一内容的描述采用了不同的表达方式,比如不同用户对同一突发事件从不同侧面进行了描述和分析,各个内容是独立的,但又内在关联到了同一突发事件。不同的社交媒体产生的内容有所差异,总体上

包括文本、图片和音视频等类型。基于MHSN内容空间的关联,可以分析信息在不同OSNs上关联的用户数,阅读、评论及转发数,进而可以全面地计算信息的影响力、热度值等。图3是社交媒体显式内容对齐关联示例。


图2 MHSN用户对齐关联示例

Fig.2 User alignment association example of MHSN


图3 MHSN显式内容的对齐关联示例

Fig.3 Explicit content alignment association example of MHSN

## 3 多异构社交网络表示模型应用例证

本文选取基于异构社交网络的内容空间关联(突发事件检测)及用户空间关联(用户兴趣挖掘)的两个应用场景,阐述多异构社交网络全局建模的应用策略。

### 3.1 基于MHSN的地域突发事件检测

#### 3.1.1 多异构社交网络突发事件检测融合策略

本文使用的社交网络地域突发事件检测如定义1所述。

**定义1**[26] 地域Top-$k$突发事件,形式化描述为一个三元组:LEE=($l$,$t$,$E$),$l$表示地域,$t$表示时间段,$E$表示Top-$k$个突发事件集合,$E=\{e_1,e_2,\cdots,e_k\}$,$e_i=\{kw_1,kw_2,\cdots,kw_n\}$。从语义上讲,地域Top-$k$突发事件指地域$l$在时间段$t$发生的,产生较大影响的$k$个事件。多个社交网络的内容空间融合问题可以简化为两两社交网络的内容融合。基于内容空间的社交网络$SN_1$、$SN_2$突发事件检测融合策略如图4所示。从自上而下的角度看,单异构社交媒体的突发事件检测包含3个核心步骤,可以完成各自的突发事件检测任务。从水平的方向看,两个异构社交媒体突发事件检测可以有3种融合策略,分别是信息融合、突发词融合和突发词簇融合,不同的融合策略对突发事件检测效果的影响见3.1.4小节结果对比部分。

基于内容空间的社交网络$SN_1$、$SN_2$突发事件检测融合策略描述如下:

(1) 融合策略1(信息融合)。假设$SN_1$、$SN_2$采集的信息集合分别为$DS\text{-}SN_1$、$DS\text{-}SN_2$,将$DS\text{-}SN_1$、$DS\text{-}SN_2$合并为一个信息集合$DS\text{-}SN$。从信息集合$DS\text{-}SN$计算得到突发词集为$EW$,后续可看作是基于同一社交网络的突发词聚类、词簇热度计算和Top-k突发事件排序输出。

(2) 融合策略2(突发词融合)。假设$SN_1$、$SN_2$计算得到的突发词集合分别为$EW_1$、$EW_2$,将$EW_1$、$EW_2$合并为一个突发词集$EW$。由于不同的社交媒体用户的活跃度不同,导致信息量、阅读数和关联用户等有较大差异,不能简单地根据计算的指标值直接排序选取,需要分别对$EW_1$和$EW_2$中的词突发值进行归一化处理,选取$m$个词构成突发词集合为$EW$,后续可基于$EW$进行聚类、词簇热度计算,进而排序得到Top-$k$突发事件。

(3) 融合策略3(突发词簇融合)。假设$SN_1$、$SN_2$计算得到的突发词簇集合分别为$EWC_1$、$EWC_2$,

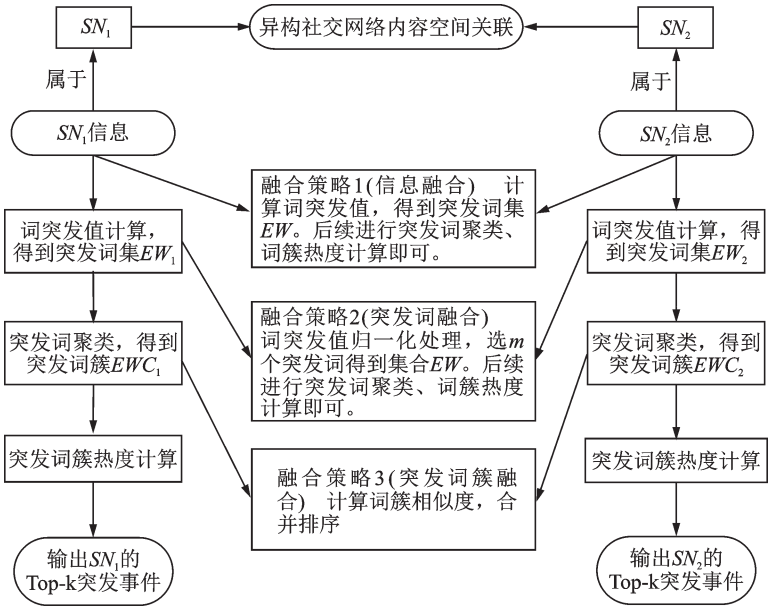



图4 基于内容空间的突发事件检测融合策略

Fig.4 Emergency detection and fusion strategy based on content space

将$EWC_1$、$EWC_2$合并为一个突发词簇集$EWC$。在融合的过程中，需要计算两个词簇的相似度，达到一定阈值两个词簇应合并在一起，形成一个词簇。两个词簇$ewc_i$、$ewc_j$相似度计算方法采用Jaccard相似系数，有

$$\mathrm{Sim}(ewc_i, ewc_j)=\frac{\left|ewc_i \cap ewc_j\right|}{\left|ewc_i \cup ewc_j\right|} \tag{1}$$

实验验证，当$\mathrm{Sim}(ewc_i, ewc_j)\geqslant 0.6$时，两个词簇进行合并效果较好。

3.1.2 单异构微博网络的地域突发事件检测方法

2018年，面向单异构微博社交网络，本文研究提出了地域Top-$k$突发事件检测方法，简记为LocBED-WB，详见文献[26]。该研究内容包含3个核心步骤，简介如下：

(1) 词突发值计算

词$w_i$在$k$时间段的突发值为

$$\mathrm{BurstyScore}(w_i)=\alpha\times F(w_i)+\beta\times U(u|w_i)+\chi\times GT(gt|w_i)+\delta\times SB(sb|w_i) \tag{2}$$

式中：$F(w_i)$、$U(w_i)$、$GT(w_i)$、$SB(w_i)$分别为词$w_i$的频率突发性、用户突发性、地域突发性和社交行为突发性；$\alpha$、$\beta$、$\chi$、$\delta$为权重系数，$\alpha+\beta+\chi+\delta=1$，$\alpha\geqslant 0$，$\beta\geqslant 0$，$\chi\geqslant 0$，$\delta\geqslant 0$。在实际应用中，可以根据社交网络的特点，对上述指标进行删减。计算得到每个词的突发值后，使用四分差选出$m$个突发特征词，按照词突发值进行降序排序，得到突发特征词集$EW$。

(2) 突发词聚类

基于突发特征集$EW$，构建突发词关联网络$EWN=(V, E)$，$V$是突发词集$EW$，$E$表示突发词之间的关联强度。突发词$ew_i$、$ew_j$关联强度是统计两个词在同一篇信息中共现的次数。突发词网络$EWN$构建完成后，使用开源的CLUTO工具包对$EWN$进行聚类，获取突发词簇$EWC=\{ewc_1, ewc_2, \cdots, ewc_q\}$，假设有$q$个词簇。

(3) 突发词簇热度计算

词簇$ewc_i$的热度值为

$$\mathrm{Score}(ewc_i)=(1+LN(ewc_i))\times(F(ewc_i)+MN(ewc_i)+MBI(ewc_i)+UN(ewc_i)) \tag{3}$$

式中$LN(ewc_i)$、$F(ewc_i)$、$MN(ewc_i)$、$MBI(ewc_i)$、$UN(ewc_i)$分别为词簇$ewc_i$的地域、频率、关联博文、关联博文影响力和关联用户指标。

3.1.3 实验数据及评测指标

新浪微博数据集BEWeiboDS为采集北京、南京两个大城市的2016年12月1日—12月30日的带有地理标签的博文,采集连云港和日照两个中小规模城市2016年5月1日—10月31日的带有地理标签的博文,形成微博数据集BEWeiboDS。百度贴吧数据集BETiebaDS为采集北京、南京两个大城市的2016年12月1日—12月30日的贴吧内容,采集连云港和日照两个中小规模城市2016年5月1日—10月31日的贴吧内容,每个市包括了区县级以上的贴吧,形成百度贴吧数据集BETiebaDS。两个社交网络数据集的情况如表1所示。

**表1 突发事件检测的两个数据集**

**Table 1 Two data sets for emergency detection**

| 城市 | 采集周期 | 新浪微博数据集/条 | 百度贴吧数据集/条 |
|---|---|---|---|
| 北京 | 2016年12月1日—12月30日 | 346 863 | 367 394 |
| 南京 | 2016年12月1日—12月30日 | 174 539 | 196 303 |
| 连云港 | 2016年5月1日—10月31日 | 63 744 | 116 452 |
| 日照 | 2016年5月1日—10月31日 | 58 227 | 104 075 |

采用精准率$P@n$作为评测指标。$P@n$是一个拟人化的指标,目前在搜索评测中用的较多。突发事件检测类似于从给定的批量信息中搜索挖掘出密切相关的地域突发事件。$P@n$指标关心的是返回的$n$个结果中,是否存在相关的信息,不考虑返回信息相关性的顺序。$P@n=m/n$,其中$n$指返回的突发事件个数,$m$指人工判断后符合突发事件检测结果的个数。由于Top-$k$突发事件检测返回的事件数量很少,人工参与评测工作量并不大。

3.1.4 结果对比

本文使用5种方法基于新浪微博数据集BEWeiboDS和百度贴吧数据集BETiebaDS进行突发事件检测对比。5种方法简介如下。(1) 方法1(LocBED-WB):使用单异构社交网络新浪微博数据集BEWeiboDS,使用3.1.2小节介绍的方法进行突发事件检测,具体方法详见文献[26]。(2)方法2(LocBED-TB):使用单异构社交网络百度贴吧数据集BETiebaDS,使用3.1.2小节介绍的方法进行突发事件检测。(3)方法3(LocBED-WB&TB-BW):使用两个异构社交网络新浪微博数据集BEWeiboDS和百度贴吧数据集BETiebaDS,在突发词计算层面进行融合,然后进行突发事件检测。(4)方法4(LocBED-WB&TB-BWC):使用两个异构社交网络新浪微博数据集BEWeiboDS和百度贴吧数据集BETiebaDS,在突发词聚类层面进行融合,然后进行突发事件检测。(5)方法5(LocBED-WB&TB-BEH):使用两个异构社交网络新浪微博数据集BEWeiboDS和百度贴吧数据集BETiebaDS,在突发词簇热度计算层面进行融合,然后进行突发事件检测。

5种方法使用两个社交网络数据集,在$P@1$、$P@2$、$P@3$、$P@4$、$P@5$和Average的评测指标结果如表2所示。

如表2所示,单独使用新浪微博数据集,方法LocBED-WB的平均准确率为0.79,精准率已经比较高了,说明单独使用新浪微博进行突发事件检测的优势。单独使用百度贴吧数据集,方法LocBED-TB的平均准确率为0.56,精准率比较低,一方面百度贴吧活跃用户数相对少,发表的信息量偏少,另外贴吧发表的帖子没有地理标签的标记,检测的很多突发事件多是广域突发事件,地域特征型不强。使用两

**表2 5个评测指标检测结果**

**Table 2 Detection results of five evaluation indicators**

| 方法 | 数据集 | *P*@1 | *P*@2 | *P*@3 | *P*@4 | *P*@5 | Average |
|---|---|---|---|---|---|---|---|
| LocBED-WB | BEWeiboDS | 0.80 | 0.80 | 0.82 | 0.80 | 0.75 | 0.79 |
| LocBED-TB | BETiebaDS | 0.60 | 0.65 | 0.50 | 0.45 | 0.60 | 0.56 |
| LocBED-WB@TB-BW | BEWeiboDS & BETieba DS | 0.80 | 0.85 | 0.75 | 0.90 | 0.80 | 0.82 |
| LocBED-WB@TB-BWC | BEWeiboDS & BETieba DS | 0.80 | 0.80 | 0.80 | 0.80 | 0.85 | 0.81 |
| LocBED-WB@TB-BEH | BEWeiboDS & BETieba DS | 0.85 | 0.82 | 0.85 | 0.88 | 0.80 | 0.84 |

个社交网络,从3个层面进行融合检测突发事件,第3种融合策略,即突发词簇热度计算融合的方法,效果最理想,准确率达到0.84,比单独使用新浪微博数据集的方法LocBED-WB提高了0.05,比单独使用百度贴吧数据集的方法LocBED-TB提高了0.28。

### 3.2 基于MHSN的用户兴趣挖掘

#### 3.2.1 多异构社交网络用户兴趣挖掘融合策略

本文使用的社交网络用户兴趣表示模型如定义2和3所述。

**定义2**[16] 用户静态兴趣是指从用户背景中挖掘出的兴趣点,$UI=\{Int_1, Int_2, \cdots, Int_m\}$,每个兴趣点是一个二元组$Int_i=(kw_i, w_i)$,$kw_i$为关键词;$w_i$为用户对$kw_i$的喜好权重。

**定义3** 用户动态兴趣是指从用户生成中挖掘出的随时间变化而变化的兴趣点,$UI=\{Int_1, Int_2, \cdots, Int_m\}$,每个兴趣点为一个三元组$Int_i=(\text{topic}_i, w_i, T)$,其中,$\text{topic}_i$是由多个关键词组成的话题;$w_i$为用户对$\text{topic}_i$的喜好权重;$T=\{t_1, t_2, \cdots, t_s\}$,$t_i$为用户讨论话题$\text{topic}_i$的各个时间点,即话题在不同时间点的分布情况。

同样,多个社交网络的用户空间融合问题可以简化为两两社交网络的用户融合。两个社交网络$SN_1$、$SN_2$在挖掘用户兴趣时,用户的静态兴趣可以从简介、标签和职位等背景信息方面融合,用户的动态兴趣可以从用户生成的内容方面进行融合。基于用户空间的社交网络$SN_1$、$SN_2$用户兴趣挖掘融合策略如图5所示。单异构社交网络的用户兴趣挖掘分为静态兴趣和动态兴趣两类,使用社交网络上用户的背景和内容信息,可以完成各自的兴趣挖掘任务。对两个社交网络$SN_1$、$SN_2$而言,静态兴趣和动态兴趣挖掘都有两种融合策略,分别是背景和生成内容的融合,以及静态兴趣和动态兴趣的融合。不同的融合策略对用户兴趣挖掘效果的影响见3.2.4小节结果对比部分。

基于用户空间的社交网络$SN_1$、$SN_2$用户兴趣挖掘融合策略描述如下:

(1) 融合策略1(背景和生成内容的融合)。假设$SN_1$、$SN_2$用户的背景信息分别为$\text{profile}_1$、$\text{profile}_2$,$SN_1$、$SN_2$用户的生成内容分别为$\text{content}_1$、$\text{content}_2$,将$\text{profile}_1$、$\text{profile}_2$合并为一个背景信息profile,将$\text{content}_1$、$\text{content}_2$合并为一个生成内容content。后续分别从profile和content中挖掘用户的静态兴趣和动态兴趣。

(2) 融合策略2(静态兴趣和动态兴趣的融合)。假设$SN_1$、$SN_2$用户的静态兴趣分别为$SN_1$-*SI*、$SN_2$-*SI*,$SN_1$、$SN_2$用户的动态兴趣分别为$SN_1$-*DI*、$SN_2$-*DI*,将$SN_1$-*SI*、$SN_2$-*SI*合并为*SN*-*SI*,将$SN_1$-*DI*、$SN_2$-*DI*合并为*SN*-*DI*。在融合用户动态兴趣时,需要计算兴趣点的相似度,然后调整权重$W$和时间点$T$的分布,$SN_1$、$SN_2$用户的一个兴趣点分别记为$SN_1\text{-}DI\text{-}Int_i=\{\text{topic}_i, W_i, T_i\}$、$SN_2\text{-}DI\text{-}Int_j=\{\text{topic}_j, W_j, T_j\}$,用户兴趣点相似度计算使用Jaccard相似系数,有

$$\text{Sim}(SN_1\text{-}DI\text{-}Int_i, SN_1\text{-}DI\text{-}Int_j)=\frac{\left|\text{topic}_i \cap \text{topic}_j\right|}{\left|\text{topic}_i \cup \text{topic}_j\right|} \tag{4}$$

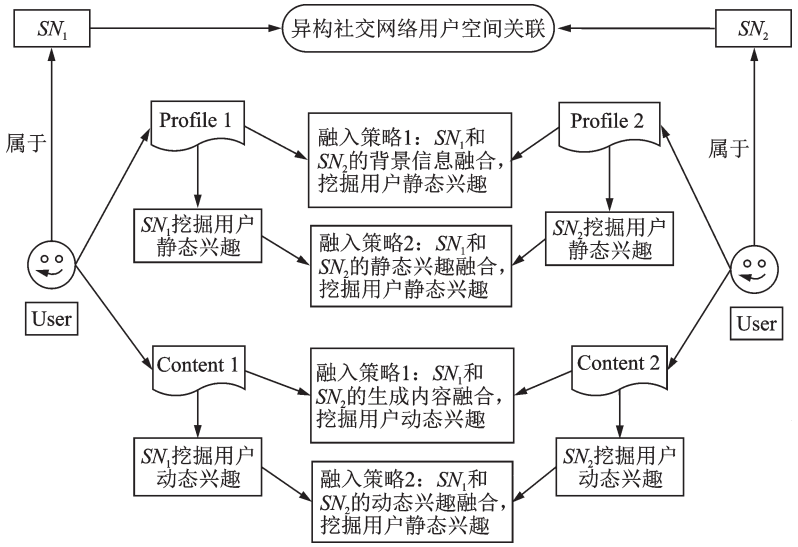



图5 基于用户空间的用户兴趣挖掘融合策略

Fig.5 User interest mining and fusion strategy based on user space

实验验证,当$\mathrm{Sim}(SN_1\text{-}DI\text{-}Int_i,\ SN_1\text{-}DI\text{-}Int_j)\geqslant 0.6$时,两个兴趣点合并效果较好。

3.2.2 单异构微博网络的用户兴趣挖掘方法

2017年,作者提出了面向微博的用户兴趣静态和动态兴趣挖掘方法,简记为USDInt-WB,详见文献[16]。该研究内容包含3个核心步骤,简介如下:

(1) 用户静态兴趣挖掘。挖掘新浪微博用户的简介、标签和职位等背景信息,得到用户的静态兴趣为$\mathrm{USInt}=\{(kw_1,w_1),(kw_2,w_2),\cdots,(kw_m,w_m)\}$。

(2) 用户动态兴趣挖掘。挖掘用户原创、转发和评论等方式的微博,得到用户的动态兴趣为$\mathrm{UDInt}=\{(\mathrm{topic}_1,w_1,T_1),(\mathrm{topic}_2,w_2,T_2),\cdots,(\mathrm{topic}_m,w_m,T_m)\}$。

(3) 用户兴趣相似度计算。两个用户兴趣相似度整合,有

$$\mathrm{UISim}(u_1,u_2)=\alpha\cdot\mathrm{USISim}(u_1.USI,u_2.USI)+(1-\alpha)\cdot\mathrm{UDISim}(u_1.UDI,u_2.UDI) \tag{5}$$

式中$\alpha$是静态兴趣和动态兴趣权重系数,$0\leqslant\alpha\leqslant 1$。

用户$u_1$,$u_2$的静态兴趣相似度计算使用Jaccard方式。用户$u_1$,$u_2$的动态兴趣中的两个兴趣点$Int_i$,$Int_j$的相似度计算公式为

$$\mathrm{UDISim}(u_1.Int_i,u_2.Int_j)=\frac{Int_i.KW\cdot Int_j.KW}{\left\|Int_i.KW\right\|\cdot\left\|Int_j.KW\right\|}\cdot\frac{\min(Int_i.|T|,Int_j.|T|)}{\max(Int_i.|T|,Int_j.|T|)} \tag{6}$$

式中综合考虑了用户兴趣点内容的相似度和兴趣点的时间周期。

3.2.3 实验数据及评测指标

本文的研究内容没有涉及不同用户在跨社交媒体的对齐关联方法。因此人工选取了100个用户,已知他们在新浪微博和百度贴吧的账号,然后从两个社交媒体中融合挖掘用户兴趣进行实验分析。对于100个用户,采用滚雪球的方式分别采集其关注和粉丝用户共计2层,即采集到了用户$u_1$关注的关注集和粉丝的粉丝集。对于采集的用户,分别从新浪微博和百度贴吧采集用户背景和生成内容信息,每个用户的背景信息合并为1条,得到的新浪微博数据集UserWeiboDS和百度贴吧数据集UserTiebaDS情况如表3所示。

**表3 用户兴趣挖掘的两个数据集**

**Table 3 Two data sets of user interest mining**

| 数据集 | 采集周期 | 用户数/个 | 背景信息/条 | 生成内容/条 |
|---|---|---|---|---|
| 新浪微博UserWeiboDS | 2018年1月1日—6月30日 | 81 394 | 81 394 | 1 139 516 |
| 百度贴吧UserWeiboDS | 2018年1月1日—6月30日 | 42 376 | 42 376 | 466 136 |

新浪微博数据集中用户$u_1$的关注集记为$u_1$.follower,作为标准答案。通过方法method$_1$计算用户间的兴趣相似度选取出的关注集记为$u_1$.follower-method$_1$,令$\left|u_1.\text{follower}\right|=\left|u_1.\text{follower-method}_1\right|$,方法method$_1$选取关注的准确率计算公式为

$$\text{RUA}=\frac{\left|u_1.\text{follower}\cap u_1.\text{follower-method}_1\right|}{\left|u_1.\text{follower}\cup u_1.\text{follower-method}_1\right|} \tag{7}$$

3.2.4 结果对比

本文使用4种方法基于新浪微博数据集UserWeiboDS和百度贴吧数据集UserTiebaDS进行用户兴趣挖掘对比。4种方法简介如下。(1)方法1(USDInt-WB):使用单异构社交网络新浪微博数据集UserWeiboDS,使用3.2.2小节介绍的方法挖掘用户兴趣,具体方法详见文献[16]。(2)方法2(USDInt-TB):使用单异构社交网络百度贴吧数据集UserTiebaDS,使用3.2.2小节介绍的方法进行用户兴趣挖掘。(3)方法3(USDInt-WB&TB-PC):使用两个异构社交网络新浪微博数据集UserWeiboDS和百度贴吧数据集UserTiebaDS,在背景和生成内容层面融合,然后挖掘用户兴趣。(4)方法4(USDInt-WB&TB-SD):使用两个异构社交网络新浪微博数据集UserWeiboDS和百度贴吧数据集UserTiebaDS,在静态和动态兴趣层面融合,然后挖掘用户兴趣。4种方法使用两个社交网络数据集,在RUA指标的评测结果如表4所示。单独使用UserWeiboDS,方法USDInt-WB推荐用户准确率RUA为0.61,说明单独使用新浪微博挖掘用户兴趣进行关注用户推荐已经比较准确。单独使用UserTiebaDS,方法USDInt-TB推荐用户准确率为0.37,准确率比较低,主要原因是百度贴吧中,用户往往对特定的贴吧感兴趣,用户之间的关注关系相对较少,不像新浪微博用户之间构建了丰富的社交关系。使用两个社交网络,从两个层面进行融合挖掘用户兴趣,第2种融合策略,即在静态和动态兴趣层面融合,效果最理想,推荐用户准确率达到0.69。比单独使用新浪微博数据集的方法USDInt-WB提高了0.08,比单独使用百度贴吧数据集的方法USDInt-TB提高了0.32,比使用第1种融合策略提高了0.04。

**表4 RUA指标的评测结果**

**Table 4 Evaluation results of RUA indicators**

| 方 法 | 数据集 | RUA |
|---|---|---|
| USDInt-WB | UserWeiboDS | 0.61 |
| USDInt-TB | UserTiebaDS | 0.37 |
| USDInt-WB&TB-PC | UserWeiboDS& UserTiebaDS | 0.65 |
| USDInt-WB&TB-SD | UserWeiboDS& UserTiebaDS | 0.69 |

## 4 结束语

本文在社交网络的用户空间和内容空间关联、不同OSNs的分类关系的基础上,给出了多异构社交网络的全局表示模型,为面向多异构社交网络的后续研究提供参考。选取多异构社会网络的地域突发事件检测、用户兴趣挖掘两个应用场景,阐述了基于内容空间和用户空间的多异构社会网络的融合策

略。以新浪微博和百度贴吧两大社交网络,进行了实验对比和分析。还需进一步提升的研究内容:(1)基于多异构社交网络的不同应用场景的抽象分析,以期为多异构社交网络的实际应用提供借鉴;(2)扩大社交网络分析的范围,选取主流的社交网络,进行更大规模的数据采集和分析;(3)基于隐式内容空间的社交媒体关联分析,使用自然语言处理、社交网络分析等技术,挖掘隐式内容在多异构社交网络的关联,进而实现突发事件、热点信息等的精准挖掘。

**参考文献:**


[1] BARTUNOV S, KORSHUNOV A, PARK S, et a1. Joint link-attribute user identity resolution in online social networks [C]//Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis. Beijing, China:[s.n.], 2012: 104-109.

[2] YANG X W, GUO Y, LIU Y, et al. A survey of collaborative filtering based social recommender systems[J]. Computer Communications, 2014, 41: 1-10.

[3] CHEN Z Q, ZHANG C, ZHAO Z, et al. Question retrieval for community-based question answering via heterogeneous social influential network[J]. Neurocomputing, 2018, 285: 117-124.

[4] SEO J W, CHOI S J, KIM Y A, et al. Word embedding-based relation modeling in a heterogeneous information network[J]. Multimed Tools Appl, 2018, 77: 18529-18543.

[5] LI Y M, HSIAO H W, LEE Y L. Recommending social network applications via social filtering mechanisms[J]. Information Sciences, 2013, 239: 18-30.

[6] TANG J, QU M, WANG M. Line: Large-scale information network embedding[C]//Proceedings of the 24th International Conference on World Wide Web. New York, USA:[s.n.], 2015: 1067-1077.

[7] 齐金山,梁循,李志宇,等. 大规模复杂信息网络表示学习: 概念、方法与挑战[J]. 计算机学报, 2018, 41(10): 2394-2420.
QI Jinshan, LIANG Dun, LI Zhiyu, et al. Representation learning of large-scale complex information network: Concepts, methods and challenges[J]. Chinese Journal of Computers, 2018, 41(10): 2394-2420.

[8] ZHU Z G, SU J Q, KONG L P. Measuring influence in online social network based on the user-content bipartite graph[J]. Computers in Human Behavior, 2015, 52: 184-189.

[9] 周小平,梁循,赵吉超,等. 面向社会网络融合的关联用户挖掘方法综述[J]. 软件学报, 2017, 28(6): 1565-1583.
ZHOU Xiaoping, LIANG Dun, ZHAO Jichao, et al. Correlating user mining methods for social network integration: A survey [J]. Journal of Software, 2017,28(6): 1565-1583.

[10] 汪潜,申德荣,冯朔,等. 全视角特征结合众包的跨社交网络用户识别[J]. 软件学报, 2018, 29(3): 811-823.
WANG Qian, SHEN Derong, FENG Shuo, et al. Identifying users across social networks based on global view features with crowdsourcing[J]. Journal of Software, 2018,29(3): 811-823.

[11] QIN H C, YUAN Y, ZHU F D, et al. Group identity matching across heterogeneous social networks[C]//Proceedings of WISE 2018, LNCS 11233. Dubai, United Arab Emirates: [s.n.], 2018: 230-246.

[12] 琚春华,赵凯迪,鲍福光. 融入紧密度中心性与信用的社交网络用户影响力强度计算模型[J]. 情报学报, 2019, 38(2): 170-177.
JU Chunhua, ZHAO Kaidi, BAO Fuguang. A user influence strength model in e-commerce social networks based on closeness and users' credit[J]. Journal of The China Society for Scientific and Technical Information, 2019, 38(2): 170-177.

[13] VU X T, ABEL M H, MAHOUDEAUX P M. A user-centered approach for integrating social data into groups of interest[J]. Data & Knowledge Engineering, 2015, 96/97: 43-56.

[14] KUNDU S, PAL S K. FGSN: Fuzzy granular social networks-model and applications[J]. Information Sciences, 2015, 314: 100-117.

[15] 吴奇,陈福才,黄瑞阳,等. 基于语义路径的异质网络社区发现方法[J]. 电子学报, 2016, 44(6): 1465-1471.
WU Qi, CHEN Fucai, HUANG Ruiyang, et al. Community detection in heterogeneous network with semantic paths[J]. Acta Electronica Sinica, 2016, 44(6): 1465-1471.

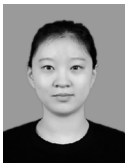
[16] 仲兆满,管燕,胡云,等. 基于背景和内容的微博用户兴趣挖掘[J]. 软件学报, 2017, 28(2): 278-291.
ZHONG Zhongman, GUAN Yan, HU Yun, et al. Mining user interests on microblog based on profile and content[J]. Journal of Software, 2017,28(2): 278-291.

[17] KONG X, ZHANG J, YU P. Inferring anchor links across multiple heterogeneous social networks[C]//Proceedings of the CIKM. San Francisco, CA, USA: [s.n.], 2013: 179-188.

[18] ZHAN Q Y, ZHANG J W, YU P S. Integrated anchor and social link predictions across social networks[J]. Knowledge &

Information Systems, 2019, 60: 303-326.
[19] BUCCAFURRI F, LAX G, NOCERA A, et al. Discovering missing me edges across social networks[J]. Information Sciences, 2015, 319: 18-37.
[20] LIU S Y, WANG S H, ZHU F D. Structured learning from heterogeneous behavior for social identity linkage[J]. IEEE Transactions on Knowledge and Engineering, 2015, 27(7): 2005-2019.
[21] 李国良,楚娅萍,冯建华,等. 多社交网络的影响力最大化分析[J]. 计算机学报, 2016, 39(4): 643-656.
LI Guoliang, CHU Yaping, FENG Jianhua, et al. Influence maximization on multiple social networks[J]. Chinese Journal of Computers, 2016, 39(4): 643-656.
[22] HUANG S R, ZHANG J, WANG L, et al. Social friend recommendation based on multiple network correlation[J]. IEEE Transactions on Multimedia, 2016, 18(2): 287-299.
[23] ZHOU X P, LIANG X, ZHANG H Y, et al. Cross-platform identification of anonymous identical users in multiple social media networks[J]. IEEE Transactions on knowledge and engineering, 2016, 28(2): 411-424.
[24] WANG Y Q, FENG C Y, CHEN L, et al. User identity linkage across social networks via linked heterogeneous network embedding[J]. World Wide Web, 2019, 22: 2611-2632.
[25] SHI C, LI Y T, ZHANG J W, et al. A survey of heterogeneous information network analysis[J]. IEEE Transactions on Knowledge & Data Engineering, 2017, 29(12): 87-99.
[26] 仲兆满,管燕,李存华,等. 微博网络地域Top-k突发事件检测[J]. 计算机学报, 2018, 41(7): 1504-1516.
ZHONG Zhaoman, GUAN Yan, LI Cunhua, et al. Localized top-k bursty event detection in microblog[J]. Chinese Journal of Computers, 2018, 41(7): 1504-1516.

**作者简介:**



**王艺霖**(2001-),女,硕士研究生,研究方向:社交网络分析,E-mail: 1411683071@qq.com。

**仲兆满**(1977-),男,副教授,研究方向:互联网大数据采集与挖掘应用。

**樊继冬**(1999-),男,硕士研究生,研究方向:大数据采集与分析。

**管燕**(1976-),女,讲师,研究方向:人工智能应用。

(编辑:刘彦东)