

# 采用超图的微博群落感知方法

莫同, 褚伟杰, 李伟平, 吴中海  
(北京大学软件与微电子学院, 100871, 北京)

**摘要:** 针对微博群落的识别与形成演化机理的获取问题, 提出一种基于超图的微博群落感知方法. 归纳微博用户的交互关系, 提出以用户为节点、交互关系为边的微博用户交互关系超图模型; 分析微博用户交互环境的情境特征, 通过 FP-TREE 方法挖掘用户交互与情境特征的关联规则; 根据关联规则对超图模型进行划分, 得到具有相同情境的微博群落. 以新浪微博为例进行了模拟验证, 结果表明该方法能够感知导致微博群落形成的情境特征, 且较传统数据挖掘方法能够更加准确地识别微博群落.

**关键词:** 微博; 社交网络; 微博群落; 超图; 情境感知

**中图分类号:** TP393 **文献标志码:** A **文章编号:** 0253-987X(2012)11-0120-07

## A Micro Blog User Community Aware Method Based on Hypergraph

MO Tong, CHU Weijie, LI Weiping, WU Zhonghai  
(School of Software and Microelectronics, Peking University, Beijing 100871, China)

**Abstract:** A method of micro-blog community awareness based on hypergraph is proposed to identify micro-blog community and explore its evolution mechanism. A hypergraph based relational model of micro-blog users is established by summarizing interactions among micro-blog users. Users are vertices, and interactions are edges in the hypergraph. Association rules among micro-blog interactions and context are mined using FP-TREE method by analyzing user profile and micro-blog message. Then, hypergraph is divided into communities with same context characters based on the association rules. A simulation test based on Sina micro-blog is performed to validate the effectiveness of the proposed method. The result shows that the method can be aware of the context of micro-blog community, and is more accurate than traditional data mining method in identifying the micro-blog communities.

**Keywords:** micro-blog; social networking; micro-blog community; hypergraph; context aware

随着网络的发展, 微博作为一种新兴的社交网络服务逐渐成为重要的沟通工具. 微博的应用范围拓展迅速, 涵盖信息分享、社交拓展、商业营销、政务发布等诸多方面, 并将全面渗透至社会的各个领域. 目前, 国内微博用户总数已超过 3 亿, 并仍在持续高速增长, 个人、企业、社会组织甚至政府机构均已将微博视作一种重要的信息交互渠道<sup>[1]</sup>.

在便于人们之间的沟通与交流的同时, 微博用

户量和信息量的增加也带来了一系列的挑战, 例如垃圾信息泛滥、谣言扩散、群体性事件产生与传播等, 因此研究微博用户群体的结构特征、建立微博用户交互模型、分析群体产生原因及演化规律、识别微博信息内容、发掘信息的传播模式与影响力等, 具有重大的学术价值和社会意义<sup>[2]</sup>.

由于微博的社会性, 用户根据小世界特性聚集形成若干群体, 称为微博群落 (micro-blog commu-

nity)<sup>[3]</sup>. 例如,一个用户和其粉丝即构成一个最基本的微博群落. 同一微博群落的用户之间交互频繁,来自群落内部的信息较易受到用户关注. 通常,一个用户可以同时处于多个群落,并与其他用户的交互关系加入新的群落,进而扩展信息获取和社交的范围. 目前,微博群落根据其形成方式大致可分为以下 3 种<sup>[3-5]</sup>. ①以信息为中心的群落. 群落形成的起因是一条或若干条信息,关注信息内容的用户互相关注形成群落. 这类群落结构松散,稳定性弱,生存周期受信息的时效性影响,存在时间通常较短,易消亡或转变成其他类型群落. ②以用户为中心的群落. 群落产生的核心是核心用户,其他用户关注核心用户形成群落. 群落主要由核心用户发布信息,粉丝对核心用户发布的信息进行评论和转发. 这类群落结构清晰,稳定性强,内部信息的发布和传播规律单一,生存周期由核心用户决定. ③以关系为中心的群落. 群落产生的核心是用户的社会关系,用户在现实社会中的关系映射到社交网络形成群落. 这类群落是社交网络中最为典型的群落,其结构受现实社会关系影响较为复杂,成员的加入和淡出与关系的亲疏相关.

前两类群落的用户收到关注对象发布的信息量随关注对象数量增长而急剧增加,导致普通信息的关注度随信息量的增加而下降. 相比之下,用户更加关注来自第 3 类群落的信息. 这是由于信息来自与之关系更为紧密的人,信息对用户的影响也更大. 因此,这类群落的结构与形成原因是研究微博用户群体的基础与重点.

目前,微博群落生成的主要方法有启发式搜索、聚类挖掘、连接矩阵分析、模拟退火等<sup>[5-7]</sup>,其主要基于群落用户交互信息的内容和频次. 但是,信息交互内容和频次只是群落关系的外在表现,其形成的内在因素是群落用户之间的社会关系,例如兴趣爱好、社会经历等. 这种关系不仅是两个用户之间的二元关系,也可能是多个用户间的多元关系,例如亲属、同学、同事等. 分析、挖掘群落的形成内因关系有助于更加深入理解微博用户群体结构,揭示群体产生及演化规律.

然而,与交互信息可直接获取不同,用户间的社会关系通常无法直接获得,需要根据用户的个人属性和相互间的交互信息进行推衍. 通过情境感知方法获取微博群落用户间的关系,进而指导微博群落的划分是一个可行的思路.

本文通过分析微博用户的交互,建立了基于超

图的微博用户交互关系模型;通过对微博用户情境的感知,挖掘用户情境与交互关系的关联规则;根据关联规则感知微博群落情境,提出微博群落的感知方法,并对相关方法进行了模拟验证.

## 1 基于超图的微博用户交互关系模型

### 1.1 超图基础

为定义超图,首先给出普通图的定义.

**定义 1** 图  $G$  是一种非线性结构,记作  $G=(V,E)$ . 式中: $V=\{v_1,v_2,\dots,v_n\}$  是顶点集, $V$  有穷非空; $E=\{e_1,e_2,\dots,e_n\}$  是边集,其中边  $e_m=(v_i,v_j)$ , $E$  可为空. 当边包含的节点偶对无先后顺序时,称作无向边或边,当边包含的节点偶对有先后顺序时,记作  $e_m=\langle v_i,v_j \rangle$ ,称作有向边或弧.

图用于描述对象(顶点)及对象间二元关系(边). 当对象间存在多元关系时,例如  $A$ 、 $B$ 、 $C$  互相是同学,则可以通过超图表示.

**定义 2** 超图  $G$  是一种非线性结构,记作  $G=(V,E)$ . 式中: $V=\{v_1,v_2,\dots,v_n\}$  是顶点集, $V$  有穷非空; $E=\{e_1,e_2,\dots,e_n\}$  是边集,其中边  $e_m$  是顶点的集合,且边  $e_m \subseteq V$ , $E$  可为空. 当边包含的节点无先后顺序时,称作无向边或边,当边包含的节点有先后顺序时,记作  $e_m=\langle V_i,V_j \rangle (V_i,V_j \subseteq V)$ ,称作有向边或弧.

由定义 2 可知,超图的边可以连接图中任意数目的顶点,可用于描述对象及对象间的多元关系. 微博用户及其交互关系可以通过超图表示.

### 1.2 微博用户交互分析

为理解微博用户之间的交互关系,建立了交互关系模型,并对其交互行为进行分析、归纳和总结如下. 目前,微博用户的交互行为主要包括以下 5 类. ①关注. 微博用户通过关注成为关注对象的粉丝,订阅关注对象发布的公开信息,用户将实时收到来自所有关注对象发布的所有公开信息. ②发布. 微博用户通过发布向外界传递信息,系统会将发布的信息实时发送给用户所有的粉丝. ③推送. 微博用户发布信息时可以使用推送将信息额外推送给其他用户,推送信息与普通的发布信息区分显示. ④评论. 微博用户可对收到的信息进行评论,用户评论以推送方式发给被评论信息的发布者及其关联用户,评论内容可由被评论信息发布者和其所有粉丝查看. ⑤转发. 微博用户可以将自己收到的信息进行转发,转发内容除原信息内容外,也包括原信息的发布者和原信息的转发与评论内容. 此外,转发的同时还可对信

息进行评论,转发和评论将分别产生一条推送信息发送给原信息发布者。

综合来看,关注是微博用户间建立信息交互渠道的主要行为。关注对象主要来源于指定条件的用户搜索,或通过转发、评论等渠道获得信息的发布者。微博用户通过微博信息进行交互,发布是产生微博信息的主要行为。转发和评论是用户对收到信息的有效反馈,推送用于向对方提示和推荐信息。从信息被关注的角度,转发、评论和推送的信息较易受到对方的关注,而发布的信息被关注的程度较上述3者相对较低。此外,由于用户对发布信息的浏览行为无法监测,因此本文研究的交互关系主要面向转发、评论和推送3类交互行为。

### 1.3 微博用户交互关系超图

根据超图的定义和微博用户的交互行为,微博用户的交互关系可以被定义成一个超图。

**定义3** 微博用户交互关系超图  $G=(V, E_f, E_c, E_@)$ , 式中:  $V=\{v_1, v_2, \dots, v_n\}$  是用户集,  $V$  中每个顶点  $v_m$  代表一个用户;  $E_f=\{e_{f1}, e_{f2}, \dots, e_{fn}\}$  是转发关系集, 其中边  $e_{fm}=\langle v_i, v_j, V_k \rangle (v_i, v_j \in V, V_k \subseteq V)$  表示用户  $v_i$  转发用户  $v_j$  的信息, 同时被推送给用户集  $V_k$ ;  $E_c=\{e_{c1}, e_{c2}, \dots, e_{cn}\}$  是评论关系集, 其中边  $e_{cm}=\langle v_i, v_j, V_k \rangle (v_i, v_j \in V, V_k \subseteq V)$  表示用户  $v_i$  评论用户  $v_j$  的信息, 同时被推送给用户集  $V_k$ ;  $E_@=\{e_{@1}, e_{@2}, \dots, e_{@n}\}$  是推送关系集, 边  $e_{@m}=\langle v_i, V_j \rangle (v_i \in V, V_j \subseteq V)$  表示用户  $v_i$  将信息推送给用户集  $V_j$ 。

根据定义3, 微博系统中所有用户及其间的交互关系可以构成一个微博用户交互关系超图。该超图可以根据3类交互操作划分成3个子图分别依下文方法进行分析处理, 为简化问题, 本文对此3种交互操作不做区分。微博用户交互关系超图  $G$  可简化为  $G=(V, E)$ , 其中  $V$  是用户集,  $E$  是交互关系集, 对  $\forall e \in E$  有  $e=\langle v_i, V_j \rangle$ 。

为进一步分析交互关系超图, 给出微博用户交互关系超图中边的度与顶点的度的定义。

**定义4** 微博用户交互关系超图  $G=(V, E)$ , 边  $e \in E$  的度  $w(e)=n(e)$ , 其中  $n(e)$  是边  $e$  出现的次数。

微博用户交互关系超图中边的度代表相同用户集合之间转发、评论或推送信息的次数。

**定义5** 微博用户交互关系超图  $G=(V, E)$  中, 顶点  $v_i \in V$  的度为  $w(v_i)=w_i(v_i)+w_o(v_i)$ 。式中: 顶点  $v_i$  的入度为  $w_i(v_i)=\sum n(e)(e=\langle v_x, V_y \rangle,$

$v_i \in V_y)$ ; 顶点  $v_i$  的出度为  $w_o(v_i)=\sum n(e)(e=\langle v_i, V_x \rangle)$ 。

微博用户交互关系超图中顶点的度代表用户参与转发、评论或推送信息的次数。该次数由两部分组成, 一部分是用户被转发、被评论和被推送的次数, 即入度; 另一部分是用户转发、评论别人信息和向别人推送信息的次数, 即出度。

根据上述定义, 可以根据微博系统日志中的交互记录, 建立或更新微博用户交互关系超图(初始为空), 超图更新算法的伪代码如下。

输入: 已有微博用户交互关系超图  $G=(V, E)$ , 交互记录集  $L$ , 其中每条记录  $l=\langle v_i, V_j \rangle$ 。

输出: 更新的微博用户交互关系超图  $G$ 。

```
For all  $l$  in  $L$  {
    For all  $v$  ( $v=v_i$  or  $v \in V_j$ ) {
        If ! ( $v \in V$ ) Add( $v$ ) into  $G.V$ ; }
        If  $\exists e=\langle v_i, V_j \rangle \in E$   $n(e)=n(e)+1$ ;
        Else Add( $e$ ) into  $G.E$ ; }
Return  $G$ 
```

对微博用户交互关系超图模型, 可以预设顶点与边的度的阈值进行剪枝, 过滤参与交互频次较低的顶点和相关的边, 然后通过超图分割算法挖掘微博群落<sup>[8]</sup>。然而, 这种方式只是从外在表象上获取了微博群落, 并没有解释微博群落形成的内在动因, 为揭示微博用户的内在交互动因, 需要对用户的交互关系进行情境感知。

## 2 基于情境关联挖掘的交互关系感知

### 2.1 情境驱动的交互关系

情境也称上下文, 指任何可用于表征实体状态的信息, 这里的实体可以是个人、位置、物理的或信息空间中的对象<sup>[9]</sup>。微博环境中, 与用户交互相关的情境可以分为 User ID、User profile 和微博信息3类, 如表1所示。其中 User ID 用于识别用户, User profile 表征交互双方用户的个人属性, 包括用户的基本信息、自定义的个性化特征标签、教育经历以及工作经历。微博信息表征交互双方传递的信息, 包括信息的内容以及交互的操作。

微博用户交互关系建立的内在因素是交互双方微博情境的内在关联。用户选择交互对象和交互信息不是随机、无目的的, 而是基于某些特定情境特征的有针对性的选择。这种基于情境的针对性来选择交互对象和交互信息的特性, 称之为用户情境驱动的交互关系, 情境对交互的驱动关系定义如下。

**定义 6** 微博用户交互关系超图  $G=(V,E)$ , 用户间交互关系  $e=<v_i,V_j>(e\in E),q$  是交互的信息,  $c$  是情境,  $C(v_i)$  是用户  $v_i$  的情境集合,  $C(V_j)$  是用户集  $V_j$  的情境集合,  $C(q)$  是信息  $q$  的情境集合, 则情境对交互的驱动关系记作  $C(e)=C(v_i)\cup C(q)\cup C(V_j)$ .

| 表 1 微博情境     |      |                  |
|--------------|------|------------------|
| 情境类型         | 情境子类 | 情境               |
| User ID      | ID   | 身份证号、姓名          |
|              | 基本信息 | 性别、所在地、生日、血型     |
|              | 标签   | 标签               |
| User profile | 教育信息 | 时间、地点、学校、班级、学历   |
|              | 工作信息 | 工作地点、单位、职位、行业、时间 |
| 微博信息         | 信息内容 | 内容类型、时间、地点       |
|              | 交互操作 | 操作类型             |

情境驱动的交互关系说明, 具有情境特征  $C(v_i)$  的用户会与具有情境特征  $C(V_j)$  的用户群进行交互  $e$ , 传递具有情境特征  $C(q)$  的信息.

用户集  $V_j$  的情境集合  $C(V_j)$  内部仍可能存在复杂的关联关系, 但为简化问题, 本文中  $C(V_j)$  特指用户集  $V_j$  中各用户取值相同的情境.

2.2 基于 FP-TREE 的交互关系感知方法

由于情境对交互具有驱动关系, 可以根据微博系统的交互历史记录和微博情境, 挖掘交互记录中情境的关联模式, 发现交互行为中情境特征的关联关系, 找出关联规则. 借鉴数据挖掘中关联规则挖掘的 FP-TREE 方法<sup>[10]</sup>, 提出基于 FP-TREE 的交互关系感知方法.

**定义 7** 情境关联 FP-TREE 是满足下列条件的一个树结构, 它由一个根节点、情境项(树干)和一个频繁项头表组成. 根节点为 null. 情境项是一个三元组  $C_{node}=<c_{id},c_{count},c_{link}>$ , 式中:  $c_{id}$  记录节点表示的情境标识;  $c_{count}$  记录到达节点的路径次数;  $c_{link}$  连接相同情境标识的下一个节点, 如果不存在, 则为 null. 频繁项头表保存各情境、用户的标识, 以及指向第一个情境、用户标识的树节点指针.

情境关联 FP-TREE 描述了引发用户交互的情境特征间的内在关联关系. 情境关联 FP-TREE 中每个节点代表一个情境特征值, 情境特征值按照交互出现的频次在频繁项头表中进行排序. 情境关联 FP-TREE 从根节点到每个叶子节点的一条分支代表一条情境关联规则, 该规则揭示了引发一类交互

的内在因素. 规则内的各情境特征值到根节点的距离由近及远根据频繁项头表依次排序.

情境关联 FP-TREE 的生成类似于传统 FP-TREE, 由对大量已知的用户交互历史记录进行分析处理<sup>[2]</sup>生成, 算法过程描述如下.

输入: 交互历史记录集  $L=\{<v_i,V_j,q>\}$ , 支持度阈值  $\lambda$ .

输出: 情境关联 FP-TREE.

扫描历史记录集  $L$ , 获取每条历史记录  $l$  的情境集合  $C(l)=C(v_i)\cup C(q)\cup C(V_j)$ ;

根据历史记录集的情境集合, 得到频繁情境集  $F$  及各情境的支持度, 对  $F$  中的情境按支持度  $\lambda$  进行降序排列, 通过过滤掉支持度低于  $\lambda$  的情境, 得到排序  $S$ ;

创建情境关联 FP-TREE 的根结点 root, 以“null”标记;

再次扫描历史记录集  $L$ , 对每条记录  $l$  的情境集合  $C(l)$  按照  $S$  进行排序和过滤, 设排序后的序列为  $[p|P]$ , 其中  $p$  是第 1 项,  $P$  是剩余项. 调用 insert\_tree( $[p|P]$ , root);

insert\_tree( $[p|P]$ , root) 的执行过程如下

如果 root 有子女 child 使 child.id= $p$ .id, 则 child.count 加 1; 否则创建 root 的一个子女 child, child.id= $p$ , child.count=1, 链接 child 到它的父结点 root; 在频繁项头表中查找名称为  $p$  的频繁项, 依次找到  $p$  指针的末尾, 将其指向 child;

如果  $P$  非空, 递归地调用 insert\_tree( $P$ , child).

表 2 给出交互历史记录集对应的情境集合及其过滤排序后的结果示例, 该示例只列举 6 类情境信息, 设支持度为 3.

| 表 2 交互历史记录集示例 |                                           |                                                  |
|---------------|-------------------------------------------|--------------------------------------------------|
| 交互            | 情境集合                                      | 整理后的情境集合                                         |
| $e_1$         | $C_{11}, C_{14}, C_{22}, C_{26}, C_{31},$ | $C_{52}, C_{11}, C_{14}, C_{43},$                |
|               | $C_{43}, C_{52}, C_{61}, C_{65}$          | $C_{22}, C_{26}, C_{61}, C_{65}$                 |
| $e_2$         | $C_{11}, C_{14}, C_{21}, C_{25},$         | $C_{52}, C_{11}, C_{14}$                         |
|               | $C_{32}, C_{42}, C_{52}, C_{64}, C_{68}$  |                                                  |
| $e_3$         | $C_{11}, C_{14}, C_{22}, C_{26},$         | $C_{52}, C_{11}, C_{14}, C_{22}, C_{26}$         |
|               | $C_{34}, C_{41}, C_{52}, C_{62}, C_{66}$  |                                                  |
| $e_4$         | $C_{11}, C_{14}, C_{23}, C_{27},$         | $C_{52}, C_{11}, C_{14}, C_{43}$                 |
|               | $C_{36}, C_{43}, C_{52}, C_{63}, C_{67}$  |                                                  |
| $e_5$         | $C_{12}, C_{15}, C_{24}, C_{28}, C_{33},$ | $C_{43}, C_{61}, C_{65}$                         |
|               | $C_{43}, C_{51}, C_{61}, C_{65}$          |                                                  |
| $e_6$         | $C_{13}, C_{16}, C_{22}, C_{26}, C_{35},$ | $C_{52}, C_{43}, C_{22}, C_{26}, C_{61}, C_{65}$ |
|               | $C_{43}, C_{52}, C_{61}, C_{65}$          |                                                  |

如表 2 所示, 各情境的频繁度排序为  $\{c_{52}:5,$

$c_{11}:4, c_{14}:4, c_{43}:4, c_{22}:3, c_{26}:3, c_{61}:3, c_{65}:3\}$ . 建立的情境关联 FP-TREE 如图 1 所示.

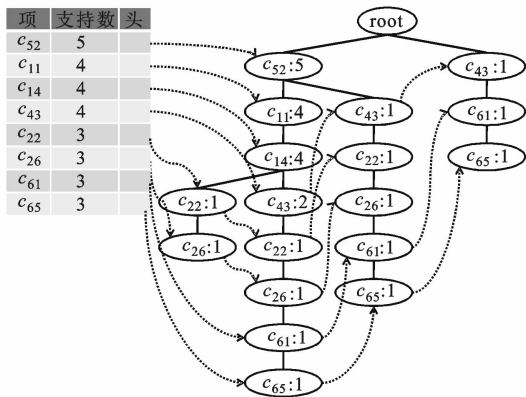


图1 情境关联 FP-TREE

### 3 微博群落感知方法

#### 3.1 微博群落

微博用户根据其社会属性的小世界特性聚集形成若干微博群落.

**定义 8** 微博群落  $m = \{u_1, u_2, \dots, u_n\}$  是具有相同情境特征的一类微博用户的集合.

超图的子图用于描述超图对象集合中的一个子对象团体及团体内部对象间的多元关系.

**定义 9** 超图  $G = (V, E)$ , 超图  $G' = (V', E')$  称作  $G$  的子图, 令  $V' \subseteq V$ , 且对  $\forall e_i' \in E'$ ,  $\exists e_j \in E$ , 使得  $e_i' \subseteq e_j$ .

由定义 3 可知, 对于超图  $G$  的子图  $G'$ ,  $G'$  中所有顶点均为  $G$  的顶点, 对  $G'$  中的每一条边  $e_i'$ ,  $G$  中存在一条边  $e_j$  与之对应, 并且  $e_j$  的顶点包含  $e_i'$  的顶点.

微博群落内用户及其交互组成微博用户交互关系超图的一个子图, 成为微博群落交互子图. 与普通子图概念的区别是, 微博群落交互子图中的边是微博用户交互关系超图中独立、完整的边. 借鉴超图子图的概念, 微博群落交互子图的定义如下.

**定义 10** 微博用户交互关系超图  $G = (V, E)$ , 超图  $G' = (V', E')$  称作  $G$  中微博群落  $m$  的交互子图, 微博群落  $m$  中的用户集对应图  $G$  中的节点集  $V'$ , 令  $V' \subseteq V$ , 且对  $\forall e' \in E'$ , 有  $e' \in E$ , 且  $e' \subseteq V'$ , 且  $\nexists e \in E, e \neq e'$ , 使得  $e' \subseteq e$ .

微博群落交互子图形成的内因是群落用户具有相同的情境特征, 如相同的 User profile、感兴趣的微博信息类型相同等. 这些情境称为微博群落情境.

**定义 11** 微博群落  $m = \{u_1, u_2, \dots, u_n\}$ , 微博群落情

境  $C(m) = C(u_1) \cap C(u_2) \cap \dots \cap C(u_n)$  是群落用户的相同情境集合.

情境关联 FP-TREE 每条分支上情境的集合表示一条情境关联规则, 该规则揭示了引发一类交互的内在因素. 根据定义 10, 一条情境关联规则可以看作是微博群落交互子图的内在形成因素. 因此, 情境关联 FP-TREE 的情境关联规则表明用户间交互特征的情境集合即为微博群落情境, 根据情境关联规则和微博用户交互关系超图可以感知微博群落.

#### 3.2 微博群落感知方法

微博用户交互关系超图  $G = (V, E)$ , 顶点  $v_i$  对应用户的情境集合  $C(v_i)$ , 情境关联 FP-TREE 中的每条情境关联规则为  $r = C_{r-ID} + C_{r-INTR}$ , 其中  $C_{r-ID}$  为用户标识情境特征集合,  $C_{r-INTR}$  为交互情境特征集合. 微博群落感知问题描述如下.

找到用户集  $V' = \{v_1', v_2', \dots, v_n'\}$ ,  $V' \subseteq V$ , 使得  $C_{r-ID}$  标识的用户集合  $U \subseteq V'$ , 且  $C(v_1') \cap C(v_2') \cap \dots \cap C(v_n') = C_{r-INTR}$ , 且对  $\forall v_i' \in V'$ , 使得  $v_i'$  在  $V'$  中的交互比重大于阈值  $\omega$ , 且  $\nexists$  用户集  $V'' = \{v_1'', v_2'', \dots, v_m''\}$ ,  $V'' \subset V'$ , 使得  $C(v_1'') \cap C(v_2'') \cap \dots \cap C(v_m'') = C_{r-INTR}$ , 且对  $\forall v_i'' \in V''$ , 使得  $v_i''$  在  $V''$  中的交互比重大于阈值  $\omega$ .

微博用户交互关系超图  $G = (V, E)$ , 用户集  $V' \subseteq V$  是一个微博群落,  $V'$  在  $G$  的微博群落交互子图  $G' = (V', E')$ , 用户  $v_i \in V'$ ,  $v_i$  在  $V'$  中的交互比重  $\varphi$  的计算公式为

$$\varphi = \frac{\alpha \omega_i(v_i) + \beta \omega_o(v_i)}{c(E')} \quad (1)$$

式中:  $\omega_i(v_i)$  和  $\omega_o(v_i)$  是顶点  $v_i$  的入度和出度;  $\alpha$  和  $\beta$  是顶点  $v_i$  的入度和出度的权值;  $c(E')$  是  $E'$  中边的数目.

微博群落感知算法过程描述如下.

输入: 情境关联 FP-TREE  $tree_{fp}$ , 微博用户交互关系超图  $G$ , 交互比重阈值  $\omega$ .

输出: 微博群落集合  $M$ .

New  $M = \Phi$ ;

深度优先遍历  $tree_{fp}$ , 得到  $tree_{fp}$  关联规则集  $R$ ;

For every  $r = C_{r-ID} + C_{r-INTR}$  in  $R$

New  $m = \Phi$ ;

New  $m' =$  All user in  $C_{r-ID}$ ;

While  $m \neq m'$

$m = m'$ ;

New  $m'' =$  All user in  $G$  and not in  $m'$  and

$C_{r-INTR} \subseteq C(\text{user})$  and  $G$  中存在边使得 user 与  $m'$  中

的节点相连;

```
While  $m'' \neq \Phi$ 
  Get user from  $m''$ ;
  Delete(user) in  $m''$ ;
  Add(user) into  $m'$ ;
  由  $G$  计算 user 在  $m'$  中的交互比重  $\varphi$ ;
  If  $\varphi < \omega$  Then delete(user) in  $m'$ ;
  Add( $m$ ) into  $M$ ;
Return  $M$ .
```

3.3 时间复杂度分析

微博群落感知方法的时间复杂度包含 3 个部分:根据微博系统的交互历史记录构建微博用户交互关系超图;根据微博系统的交互历史记录和微博情境生成情境关联 FP-TREE;在上述二者基础上建立微博群落感知方法。

假设微博系统的交互历史记录数量为  $l$ ,每条交互记录的平均参与用户数为  $v$ ,微博情境种类为  $c$ ,关联规则的平均数目为  $r$ ,每个用户的平均交互用户数为  $n$ ,则微博用户交互关系超图构建算法的时间复杂度为  $O(l * v)$ ;情境关联 FP-TREE 的生成算法的时间复杂度为  $O(2l * c)$ ;微博群落感知算法的时间复杂度受微博群落内用户的实际交互情况影响,其最小值为  $O(r * n)$ ,最大值为  $O(r * n^2)$ 。

4 应用验证

4.1 实验设计

本文以新浪微博作为案例背景,实验环境为 2.5 G 双核 cpu,3 GB 内存,windows XP 操作系统,使用 mySQL 作为微博交互记录的存储数据库。

受隐私保护因素和数据采集条件限制,本文以作者及其相关用户的微博作为实验数据,手工进行数据预处理,仅存储交互关系和用户的微博情境。

根据现实情况,作者及其新浪微博上的相关用户可以划分成 7 个群落。为验证微博群落感知方法的有效性,分析交互记录数据量对算法效果的影响,并与基于交互记录挖掘的传统方法进行比较,共设计了 4 组实验,分别搜集作者及其相关用户的 3、7、10 和 14 d 的微博作为实验数据进行群落划分。

微博群落感知中交互顶点的入度和出度的权值  $\alpha$  和  $\beta$  均设为 1,即出度和入度视作权重等同的交互;交互比重阈值  $\omega$  设为 10%。

交互记录的挖掘方法采用 k-means 聚类,以用户与簇内用户的交互记录次数作为节点到簇的距离,当距离大于簇内用户交互总次数的 10%时,将

节点凝聚入簇。

4.2 结果与分析

实验结果如图 2 所示。

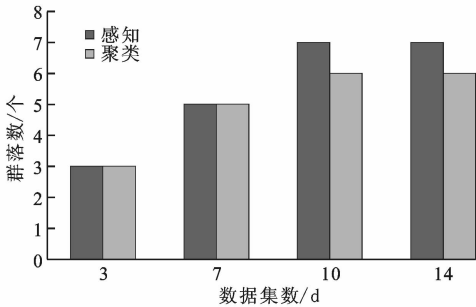


图 2 情境关联 FP-TREE

由图 2 可看出,随着数据搜集天数,即微博数据量的增加,两种方法识别出的数目均有所提高,并最终趋于稳定。微博群落感知方法的最终结果与实际情况相一致,而聚类挖掘方法的最终结果与实际情况相接近,但仍有差别。以 14 d 的数据为例,两种方法获得的群落结构如表 3 所示。

表 3 群落结构对比

| 群落用户                                                                    | 群落形成<br>社会因素 | 群落情<br>境特征          | 可识别<br>方法 |
|-------------------------------------------------------------------------|--------------|---------------------|-----------|
| $u_1, u_4, u_{13}, u_{15},$<br>$u_{18}, u_{19}, u_{20}, u_{21}, u_{22}$ | 同校好友         | 大学、所在地、<br>个性标签、年龄段 | 感知、挖掘     |
| $u_2, u_7, u_8, u_9, u_{11}$                                            | 科研伙伴         | 职业、个性标签             | 感知、挖掘     |
| $u_8, u_{13}, u_{14}, u_{21}, u_{22}$                                   | 工作同事         | 职业、单位、<br>工作地点      | 感知、挖掘     |
| $u_3, u_{23}, u_{24}$                                                   | 中学同学         | 中学、年龄段              | 感知、挖掘     |
| $u_{14}, u_{21}, u_{22}$                                                | 同事挚友         | 职业、单位、工作<br>地点、个性标签 | 感知        |
| $u_{10}, u_{12}, u_{16}, u_{17}$                                        | 生活好友         | 所在地、个性<br>标签、年龄     | 感知、挖掘     |
| $u_5, u_6$                                                              | 合作伙伴         | 所在地、<br>个性标签        | 感知、挖掘     |

根据实验结果,可以得出如下结论:对群落的识别需要交互数据的支持,当交互数据量过小时,无法体现各群落的特征;同一用户可能属于不同群落,不同群落间用户存在交叉;群落的形成内因是由于用户的相同情境而形成的内在关联关系,而交互则是这种关联关系的外部体现;对交互关系挖掘之后的每条情境关联规则对应一个微博群落;针对每条情

境关联规则,规则中标识情境的用户必属于该规则对应的微博群落;不在关联规则标识情境中的用户也可能属于该群落,取决于其交互情境特征和与群落中其他用户的交互情况;群落可能存在嵌套关系,即某个群落中的若干用户可能形成一个新的小群落,而聚类挖掘方法难以识别这类群落。

## 5 结 论

为识别微博群落,了解群落形成原因,进而对微博用户的管理以及增值业务的开展提供指导依据,提出了一种基于超图的微博群落感知方法.以用户为顶点,交互关系为边,通过超图描述微博用户交互关系.抽取个人属性和微博信息特征得到微博用户交互情境,利用 FP-TREE 挖掘用户情境与交互关系的关联规则,揭示用户交互的内在动因.根据关联规则标注微博群落情境,并提出根据微博群落情境和微博用户交互关系超图感知微博群落的方法.通过模拟分析,验证了所提方法的有效性。

### 参考文献:

- [1] 张晨逸,孙建伶,丁轶群. 基于 MB-LDA 模型的微博主题挖掘[J]. 计算机研究与发展, 2011, 48(10): 1795-1802.  
ZHANG Chenyi, SUN Jianling, DING Yiqun. Topic mining for microblog based on MB-LDA model [J]. Journal of Computer Research and Development, 2011, 48(10): 1795-1802.
- [2] 中华人民共和国科学技术部. 国家重点基础研究发展计划和重大科学研究计划 2013 年重要支持方向[EB/OL]. (2012-02-11)[2012-06-20]. <http://www.most.gov.cn/fz-gw/zfwj/zfwj2012/201202/W020120210626443434599.doc>.
- [3] YANG Bo, CHEUNG W K, LIU Jiming. Community mining from signed social networks [J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(10): 1333-1348.
- [4] LANCIERI L, DURAND N. Internet user behavior: compared study of the access traces and application to the discovery of communities [J]. IEEE Transactions on Systems, Man and Cybernetics: Part A Systems and Humans, 2006, 36(1): 208-219.
- [5] YANG Bo, HUANG Jing, LIU Dayou, et al. A multi-agent based decentralized algorithm for social network community mining [C]// Proceedings of the International Conference on Advances in Social Network Analysis and Mining. New York, USA: ACM,

2009: 78-82.

- [6] SUNDARAM H, LIN Yuru, DE CHOUDHURY M, et al. Understanding community dynamics in online social networks: a multidisciplinary review [J]. Signal Processing Magazine, 2012, 29(2): 33-40.
- [7] YANG Bo, LIU Jiming, FENG Jianfeng, et al. On modularity of social network communities: the spectral characterization [C]// Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology. Piscataway, NJ, USA: IEEE, 2008: 127-133.
- [8] REID F, MCDAID A, HURLEY N. Partitioning breaks communities [C]// Proceedings of the International Conference on Advances in Social Networks analysis and Mining. New York, USA: ACM, 2011: 102-109.
- [9] GU Tao, PUNG Hungkeng, ZHANG Daqing. A service-oriented middleware for building context-aware services [J]. Journal of Network and Computer Applications, 2005, 28(1): 1-18.
- [10] HAN Jiawei, PEI Jian, YIN Yiwen. Mining frequent patterns without candidate generation [C]// Proceedings of the ACM-SIGMOD International Conference on Management of Data. New York, USA: ACM, 2000: 1-12.

### [本刊相关文献链接]

- 翟治年, 奚建清, 卢亚辉, 等. 主动授权管理中的关联继承机制. 2012, 46(4): 24-31.
- 豆增发, 高琳. 利用膜粒子群优化和信息熵的医学文本特征选择. 2012, 46(4): 45-51.
- 陈刚, 蔡远利, 穆静, 等. 海量信息异常检测问题的异常概率排序算法. 2011, 45(4): 36-40.
- 李志雄, 严新平. 独立分量分析和流形学习在 VSC-HVDC 系统故障诊断中的应用. 2011, 45(2): 44-48.
- 刘京鑫, 孙剑, 孟德宇. 基于视觉原理的分类算法. 2010, 44(10): 116-119.
- 冯少荣, 张东站. 高效的访问预测新算法. 2010, 44(4): 28-33.
- 吴宁, 柏春霞, 祝毅博. 一种应用关联规则森林的改进贝叶斯分类算法. 2009, 43(2): 48-52.
- 王志文, 夏秦, 李平均. 一种面向网络行为因果关联的攻击检测方法. 2008, 42(8): 931-935.
- 朱虎明, 焦李成. 基于免疫记忆克隆的特征选择. 2008, 42(6): 679-682.

(编辑 武红江 赵大良)