微博用户模型复杂网络中多维有向社区发现

刘大海 张博锋 邹国兵 顾程伟
(上海大学计算机工程与科学学院 上海 200444)

摘 要 大多数社区发现是基于一种信息的,即从一个维度来划分社区。但在现实场景中,用户之间社区构成是受兴趣、社交关系、地域、教育背景等诸多因素共同影响形成的。这些多维信息有些是无向的,如兴趣相似度等;有些是有向的,如关注关系等。根据有向社区发现的原理,将多个维度的信息融合,提出一种面向多维复杂网络的有向社区发现(MDCD)算法。通过实验证明,MDCD 算法相对于传统的多维社区发现方法 AMM 算法,社区发现结果准确率提高了 17.7%、F-measure 值提高了 0.068;与一维的兴趣相似度网络进行对比,MDCD 算法的三维复杂网络社区发现结果的准确率提高了 36.1%、召回率提高了 25.3%。由于多维有向社区发现综合考虑了多维的信息,得到的社区结构具有更重要的社会意义。

关键词 用户模型 复杂网络 多维有向社区发现

中图分类号 TP3 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2016.07.031

MULTI-DIMENSIONAL DIRECTED COMMUNITY DETECTION IN COMPLEX NETWORK OF MICROBLOGGING USER MODEL

Liu Dahai Zhang Bofeng Zou Guobing Gu Chengwei
(*School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China*)

Abstract Most of community detections are based on one kind of information, i.e. to partition the community using one dimension. However in reality scene, the composition of communities between users is formed by the combined effect of many factors, such as interests, social relationships, geography, educational background, etc. Moreover, some of these multi-dimensional information are undirected, for ex., the similarity of interests, but some others, like the relationship of follow, are directed. Based on the principle of directed community detection, in this paper we fuse the multi-dimensional information and propose a multi-dimensional complex network-oriented directed community detection algorithm (MDCD). It is proved through experiment that the MDCD algorithm, relative to conventional multi-dimensional community discovery algorithm AMM, improves the accuracy of community detection result by 17.7% and the F-measure value by 0.068; Furthermore, by comparing the MDCD algorithm with the one-dimensional interests similarity network, it improves the precision rate of three-dimensional complex network detection result by 36.1% and the recall rate by 25.3%. Since the multi-dimensional directed community detection considers the multi-dimensional information comprehensively, the community structure obtained has more important social significance.

Keywords User model Complex networks Multi-dimensional directed community detection

0 引 言

以 Twitter、新浪微博、腾讯微博为代表的社交网络及微博客服务网站已经成为一种社会媒体的代表,已经在人们的生活中成为重要的信息交流平台。现在微博用户模型是以用户作为社交网络的节点,好友相互关注关系、用户喜欢的内容、地理位置作为社交网络的边,以此形式构成了多个维度的复杂网络,进而进行社交网络的社区划分。

复杂网络中存在的社区结构是以子网络内部存在多条相互关联的边,每个子网络与其他网络之间存在少量相互关联的边构成的。子网络本身构成一个社区,网络内部的节点也具有相同或者相似的属性[1]。现在的社交发现算法在划分复杂网络过程中,大多是基于单维度因素进行社区划分的。但现实场景中,用户间社区是受兴趣、社交关系、地域、教育背景等诸多因素共同影响的,所以只考虑了一个维度的信息的社区划分不符合现实应用的。同时,鉴于用户关系和用户信息具有数据关系冗杂、数据存储量巨大、数据分布离散等特点,复杂网络划分算法根据其属性特点,对其进行综合利用、融合分析,是现在研究的主要方面。

现有社区划分算法主要有 GN 算法[2]、谱平分法[3]、近似 GN 算法[4]、结构相似度算法[5]、最大化模块度算法[6]、优化模块度算法[7]、基于模块度快速算法[10]、多维信息融合的社区划分方法[11]等。

GN 算法[2]是优先删除边介数最大的边,直至整个网络退化成一个社区为止。此类算法不需要预先知晓社区个数,针对层级结构社区划分较为友好,但其计算复杂度较高。

收稿日期:2015-02-05。国家自然科学基金项目(61303096);上海市自然科学基金项目(13ZR1454600)。**刘大海**,硕士生,主研领域:复杂网络,数据挖掘。**张博锋**,研究员。**邹国兵**,讲师。**顾程伟**,硕士生。

近似 GN 算法[4]是 Tyler 等人将统计方法引入 GN 算法中，Radicchi 等人提出近似 GN 的层次算法，继续对 GN 算法进行的改良，其算法是以边聚系数 Edge Clustering Coefficient 作为判断环路个数的参数。

结构相似度算法[5]是由刘大有等人提出的，其算法思想由结构相似度代替 GN 算法中的边介数概念，但是该方法主要是针对社会网络结构设计的，不适用于其他类型的复杂网络。

最大模块度算法[6]是根据社区模块度指标来衡量社区分类是否符合实际的社区划分。模块度作为社区划分的目标函数，其主要思想是将实际网络中的社区内部已经存在的边减去伪随机网络中连接社区内部节点所需的边的期望值。

优化模块度算法[7]不局限于无向图网络，实际生活中关联往往是有方向性的。所以，Newman 又提出了有向网络的模块度，如式(1)所示：

$$Q = \frac{1}{m}\sum_{ij}(A_{ij} - \frac{k_i^{in}k_j^{out}}{m})\delta(c_i,c_j) \tag{1}$$

Chauhan[8]等人通过网络的邻近矩阵的最大特征值来划分社区。Newman[9]将已有的社区划分方法例如谱分析法映射到图的最小割方法中，并使用最大似然估计将全局搜索变为局部搜索，并实现对无向图的社区划分。

基于模块度的快速算法[10]是 Blondel 等人提出的一种在大规模复杂网络中的快速社区发现算法，算法是基于模块度最优的探索式算法，首先将社区中每个节点视为一个社区，之后将此节点置入邻近节点社区中，计算其相应的模块度，如果模块度增加，则将此节点社区并入邻近社区中；否则保持不变。以此类推，循环执行这个过程，直至所有社区达到稳定程度。之后，构建上层网络，重复进行上述社区合并过程，直到模块度不再增加为止。该算法过程简单，但其仅考虑无向网络的社区划分情况。

Lei Tang 等人提出了多维信息融合的社区划分方法 AMM[11]。该方法是将多维信息网络先融合成一个维度网络，然后用经典的方法进行社区划分。它的融合方法是非常粗糙的，只是进行了多个维度网络的简单加和求平均，它的结果优劣主要侧重在之后的单维划分方法的效果。

上述各类算法都是经典的社区划分方法，但它们都是针对单一维度的信息进行划分或粗糙地进行多维融合划分的。对此，本文针对微博用户模型复杂网络，提出了一种多维有向社区发现方法(MDCD)，并通过实验对 MDCD 算法进行了测试和对比，相比其他多维算法和单维网络，得到了比较良好和有意义的社区划分结果。

1 多维用户模型网络的构建

通过腾讯微博收集了 12 761 个用户的信息、微博和相互关注关系，然后对其中用户进行筛选，选出活跃用户 3 856 个。对此 3 856 个用户进行用户模型网络的建立。本文采用用户社交关系、地理位置信息和兴趣相似度三个维度构建多维用户模型复杂网络。

用户社交关系网络，即是用户之间的关注与被关注关系，这些现实生活中用户的主动发出的行为，是一种强关系，构成了用户社会关系网络的有向边。这种强关系，非有即无，所以，定义边的权重为 1。即社交关系网络是一个有向带权图。

用户地理位置网络，既是用户的基本信息里的地理信息，如果是同一个地方的则有边的联系。在用户地理位置信息里，分为三个粒度：国家、省份和城市。所以定义国家相同为第一等级，国家与省份相同为第二等级，国家、省份和城市相同为第三等级。量化用户之间的地理关系 $\varphi(x_1,x_2,y_1,y_2,z_1,z_2)$ 如式(2)所示：

$$\varphi(x_1,x_2,y_1,y_2,z_1,z_2)=\begin{cases}0.2 & x_1=x_2\\ 0.5 & x_1=x_2 \quad \text{且} \ y_1=y_2\\ 1 & x_1=x_2 \quad \text{且} \ y_1=y_2 \quad \text{且} \ z_1=z_2\end{cases} \tag{2}$$

其中，x_1、x_2表示两个用户所属的国家，y_1、y_2表示两个用户所属的省份，z_1、z_2表示两个用户所属的城市。

出于隐私等原因，有些用户的地理位置信息是不完全的，再者如果存在大量的同一个国家的用户，会造成用户之间的连接过多从而导致的计算量过大，所以根据实际拥有地理位置的用户数与总用户数之比，来确定地理关系 $\varphi(x_1,x_2,y_1,y_2,z_1,z_2)$ 的取舍。当实际拥有地理位置的用户数与总用户数之比大于 50% 时，$\varphi(x_1,x_2,y_1,y_2,z_1,z_2)$ 只取等于 1(即相同城市)的地理关系；否则，$\varphi(x_1,x_2,y_1,y_2,z_1,z_2)$ 就取等于 1(即相同城市)和 0.5(即相同省份)的地理关系。这样即构建成了一个地理位置无向带权图。

前两个网络都很好构建，构建最复杂的就是第三个：用户兴趣相似度网络。它是由用户模型[12]中的用户兴趣度计算出来的。用户兴趣度则是将用户的微博内容用 TF-IDF(Term frequency - inverse document frequency)方法[13,14]和本体库进行结合计算得到的。

有些用户之间的兴趣相似度很低，而且如果两两用户的兴趣度都放进网络，那么这个网络就是个完全图，这样不仅增加了计算量，而且这样的图进行社区划分是没有意义的。所以，根据复杂网络的边数定义：边数 $= n\ln n$，n 是图的节点数，选定一个阈值 α，小于 α 的兴趣相似度都裁剪掉。然后，再计算出一些被裁剪掉而没有边的孤立节点的最大相似度边，将其加入到裁剪后的网络中，从而形成了用户兴趣相似度网络。这个网络就代表用户与用户之间对于某个领域或某个话题具有共同的爱好或兴趣。所以，用户兴趣相似度网络是一个无向带权图。

2 多维有向社区发现方法

本文的多维有向社区发现方法是对社交关系(有向)、兴趣相似度(无向)和地理位置信息(无向)三个维度的融合网络，充分考虑其边的方向性，进行社区发现的。

2.1 无向边分析

在微博网络中，用户与用户之间的兴趣相似度是很重要的关联。兴趣相似度是根据用户所发的或者转发的微博内容，经过本体库的基础计算出来的语义度，再计算出来的用户之间的语义兴趣相似度组成用户之间的网络联系的，这样的网络图代表了用户所感兴趣的东西或者话题，是具有很高的价值的。

而另一种社交信息组成的网络则是不同类型的，用户与用户的社交关系网络是有向无权图，而用户之间的兴趣相似度网络是无向有权的图，这其中最关键的是有向图与无向图，即边的方向性。有向图的方向性代表了非常重要的信息，不能直接将其方向性抹掉。那么思考从无向图入手，无向图本身其实代表的是两个用户之间的关系，用户 1 对用户 2 与用户 2 对用户 1，都是具有等价的意义的。所以，就将无向图的一条边，拆分成两条有向边：一条是用户 1 指向用户 2，一条是用户 2 指向用户 1，

且这两条有向边都具有权重并且权重相等,如图 1 所示。

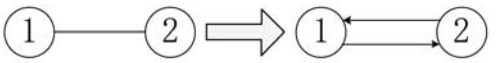

图 1　无向边转换成为有向边

所以,无向图也就可以转换成有向图。

在阐述算法之前,先进行分析和定义几个概念[15,16],本节的算法也是根据文献[15]中一部分算法改进的。

2.2　有向边的影响力分析

在有向图中,每个节点的度分为入度和出度。入度和出度分别代表了不同的意义,对于出度大的节点,可以看到出度是这个节点本身对别的节点的跟随,而对于入度大的节点,入度则是这个节点对别的节点的吸引。很明显,入度大的节点在这个网络图中是更重要的,它们是整个网络中的枢纽,是具有高的影响力的。所以,对于每一个节点,入度的意义大于出度的意义,在算法中,入度的统计是非常重要的,它是节点之间的相似性的基础[17]。

在有向带权图中,如果单纯只把边的权重值算成边的影响力是不合适的。分析入度对节点的影响,可以看到一个节点如果入度很大,代表在这个网络中对它关注的人或节点很多,这个节点的影响力很大,它如果改变,关注它的人必然也会受到一定的影响。关注的边越多,影响范围越大。那么,它所发出的边,即关注他人的边,必然要比一个很少被关注的人的影响力大。举个例子,微博的中的大 V 用户,即明星或名人用户,他被关注的人很多;微博中的一般用户,如我,被关注的人很少。可以很明显的看到,如果名人用户的影响力是要大于我的,每天“看着”他的人很多,而我几乎没有。这样一来,如果名人用户主动关注了另外一个人,无论这个人是否是大 V 用户,关注名人用户的人,肯定也会注意到这个人,至少作用到这个人的影响力要比我关注他来的大。所以,定义一个边的影响力 $E_{j\to i}$,如式(3)所示:

$$E_{j\to i} = k_j^{in} \times w_{ji} \tag{3}$$

其中,$E_{j\to i}$表示的是节点 j 到节点 i 这条边的影响力;k_j^{in}表示节点 j 的入度;w_{ji}表示的是节点 j 到节点 i 这条边的权重。

这样,就可以把节点的影响因素加到边的上面,使边在网络中的影响力更加合理。

2.3　节点权重的定义

节点的入度是关键,那么,节点的权重应该怎么定义呢? 单纯用一个节点的入度来代表显然是不够的。可以看到有向边对一个节点是有很明显的作用的,同样的,对于上一节的例子,名人用户关注(指向)的另一个人,即使这个人的入度为 1,即只有名人用户关注他,他的影响力也比有好多个一般用户关注的人大。这说明名人用户关注他的那条边对他(这个节点)的影响力是巨大的。

所以,定义节点 i 的权重 V_i,如式(4)所示:

$$V_i = \sum_{j\in V} (k_j^{in} \times w_{ji}) \tag{4}$$

其中,V 是指向节点 i 的节点的集合;k_j^{in}表示节点 j 的入度;w_{ji}表示的是节点 j 到节点 i 这条边的权重。

这样,就能计算出整个网络中每个节点的权重值,如果一个节点没有入度,那么它的权重值就是 0。

2.4　节点对社区的归属度

在划分出一个社区之后,为了衡量其周围的节点是否会加入到这个社区,这时候定义一个节点 i 对社区 C 的归属度 $A_{(i,C)}$。当节点 i 与社区 C 内的节点的边的联系比和非社区 C 内的节点边的联系多,那么节点 i 就很大可能是属于社区 C 的。所以归属度 $A_{(i,C)}$计算如式(5)所示:

$$A_{(i,C)} = \frac{\sum_{j\in C} (w_{ij} + w_{ji})}{\sum_{j\notin C} (w_{ij} + w_{ji})} \tag{5}$$

其中,w_{ij}和 w_{ji}分别是节点 i 到节点 j 的边的权重和节点 j 到节点 i 的边的权重。当 $A_{(i,C)} > 1$ 时,节点 i 加入到社区 C 中,否则,则不加。

2.5　算法过程

通过以上分析,MDCD 算法描述如下:

1) 将单维的用户兴趣相似度无向带权网络转换成有向带权网络;将单维的地理位置无向带权网络转换成有向带权网络。

2) 把 1)中的用户兴趣相似度有向带权网络和地理位置有向带权网络加入到社区关系有向带权网络中,由此形成一个新的融合三维的有向带权网络。

3) 统计新的网络中的每个节点的入度。

4) 根据式(3)计算整个网络中每条边的影响力。

5) 根据式(4)和上一步中计算的边的影响力,计算整个网络中每个节点的权重。

6) 对整个网络的所有节点按照其的权重值进行从大到小的排序。找出权重值最大的核心节点,如果有多个权重相等的节点,随机选一个。从这个核心节点开始,将与这个核心节点相连接的所有其他节点和这个核心节点一起化成初始社区 C。

7) 在剩余节点中,对和社区 C 内的节点有边相连的节点逐一计算其与社区 C 的归属度,式(5),如果归属度大于 1,则加入社区 C,否则,不加入。

8) 对再剩余的节点,再回到步骤 6),直到所有节点都被划分到社区中为止。如果整个网络中有入度为零的节点,将其归入到与之边联系最多的社区。

9) 由此得到了初始的社区划分,然后根据有向模块度[7],对每个节点进行移动操作。即将其重新分别放入与之有边联系的社区,再计算网络的有向模块度,选取最大的有向模块度所对应的社区划分,成为最终最优的社区划分。

至此,算法结束。

算法流程如图 2 所示。

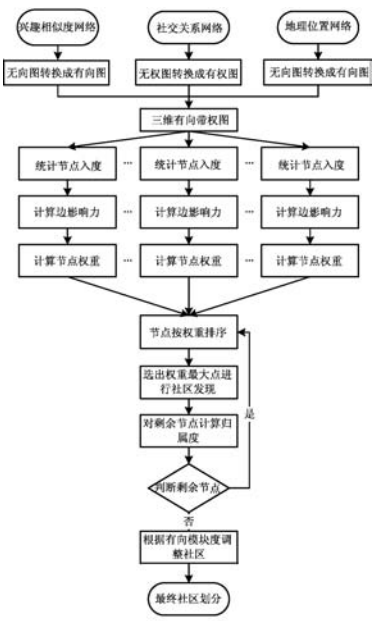

图 2　三维复杂网络融合社区发现算法

3 实验分析和对比

3.1 实验数据准备

首先通过腾讯微博的 API 来收集用户的基本信息、微博内容和用户之间的关注与被关注关系，每隔 15 天收集一次，然后针对这些数据进行数据预处理，筛选出活跃的用户。以用户所发的微博为标准，选出在本次收集中，发微博数超过 70 以上的用户作为活跃用户，而且筛选出来的与之有关系（不论是地理关系还是社会关系）的用户，都是活跃用户，筛选出 3856 个活跃用户。

3.2 二维社区发现

首先对社交关系和兴趣相似度通过 MDCD 算法进行二维融合社区发现。因为社交关系网络是一个强社区，所以它呈现除了极高的有向模块度，而且两个维度融合所代表的意义也是不同于一种维度的。本文通过研究它们的准确率和召回率[18]，并细化到用户与用户之间的关系，来进行对比。本文选择第二批数据当作训练集；选择第六批数据作测试集，它们之间相差三个月。单个维度网络和二维融合网络的社区发现对比如表 1 所示。

表 1 单维和二维网络的准确率、召回率对比

网络类型 / 划分结果	社交关系网络	兴趣相似度网络	二维融合网络
准确率	87.4%	26.3%	56.4%
召回率	83.9%	26.0%	54.6%

首先说明，单维的社交关系网络的准确率和召回率都非常高，这也是很符合实际情况的。在实际生活中，用户之间的关注与被关注这本身就是一个强关系，而且用户对另一个用户的关注或者取消关注，这本身也是一个强动作，要用户自身发出，所以它发生的频率会很低，表现在数据上就是高的准确率和召回率。而用户的兴趣度和兴趣相似度是经过它们的微博内容文本语义计算出来的，在这三个月中，用户无论是关心的话题还是热点都是会发生漂移的，所以在数据上也可以看到，单维兴趣相似度网络的准确率和召回率也充分表现了兴趣度的漂移。再来看看它们的二维融合网络的社区发现的准确率和召回率，可以发现，在社交关系和兴趣相似度的共同作用下，二维融合网络对比单维的兴趣相似度网络的准确率和召回率，提高了 2 倍以上的百分点，这个提高是十分可观的。融合之后的社区发现是把社交关系和兴趣相似度一起考虑的，比单纯的单维网络的社区发现有更好的意义。

通过实际的数据发现，用户的社交关系网络是很稳定的，不易发生漂移的，而用户兴趣度是容易改变的，容易发生漂移。那么，联想实际生活想想，有多少用户是因为兴趣度相似，产生了社交强关系。例如，我喜欢曼联足球俱乐部，你也喜欢；我喜欢读乔治·R·R·马丁的小说，你也喜欢读他的小说，虽然，我和你是实际生活不认识的，但因为兴趣度极度相似，会产生例如我关注你、你关注我或我们相互关注的强关系。相似的，也有部分用户本来相互关注或单方向关注，但因为兴趣度的转移，兴趣度逐渐不同，可能也会发生相互取消关注的行为。兴趣度和社交关系之间的影响正是融合所要研究意义所在。那么，可将用户之间的这类关系分成如下两种：

1）第二批社交关系存在，第二批兴趣相似度不存在，导致第六批社交关系不存在；

2）第二批社交关系不存在，第二批兴趣相似度存在，导致第六批社交关系存在。

首先来看关系 1），通过计算，得到从第二批数据到第六批数据这段时间，有 4637 条社交关系（含有 2894 个用户）是因为用户之间没有兴趣相似度改变了。而原来社交关系中，共有 14770条关系和 3856 个用户，由此可以得到总共有 31.4% 的关系被改变了，75.1% 的用户受到了影响。这只是一个粗略的数据，因为在这里面也会有兴趣度相似的用户，但可能因为在社交关系上两人闹矛盾等原因，直接取消关注的情况。再经过计算，得出这其中有 138 条关系（包含 242 个人），它们两两用户之间其实是属于同一个兴趣度社区的，也就是说，精确的数据是4499 条关系（30.5%）和2652个用户（68.8%），它们的变化是受到了兴趣度的影响，这个比例是非常巨大的。

关系 2）更能精确地表达出社交关系和兴趣相似度之间的影响，计算得出，本来没有社交关系的用户，因为有共同的极强的兴趣相似度，他们最后发生了关注的社交关系，这样的用户一共有 34 个，他们做出的关注的强关系有 17 条，数据如图 3 所示。

Id	name1	name2	ModularityCl...
1	a529973002	wjy977890801	1
2	dongchun210	z_y1030	3
3	feihe365	gpssss2888	22
4	fwaini1314	zhengcheng2010	28
5	gong2963559	tiramisu_lin	22
6	guqingyun660	zsp729	15
7	hxmailjp	z627036253	1
8	lichun19920709	poppy-_forever	155
9	liuguangdi1987	yswsyy	15
10	ma12856515	wsf549053034	121
11	maomaoyu_xingy	ncogg277829543	3
12	mmsgdzt	shiloulou2012	3
13	p315932265	wangyuliang0000	53
14	qinyunqing9675	yongge31327419	3
15	sololoveforever	yu-qiuli	46
16	t2282088146	weibosdsd4975	3
17	niuhaiming8	niuhuayun	1

图 3 受到兴趣相似度影响而发生社交关系的用户

可以看到，这 34 个用户共同属于的兴趣度社区也已经标识出来。他们两两都是属于同一个兴趣度社区的，说明这 17 条关系是 100% 受到兴趣度影响而形成的。同样的，从整个社区来看，有 0.9% 的用户收到了影响。结合实际，这是非常符合现实社交关系的，当两个用户，要发出关注这种很强的需要人为主动的社交关系时，如果两个人只有一点儿兴趣度相似是不行的，只有他们有着极强的兴趣相似度，他们才有可能发出关注的行为。

通过分析可以知道，用户的社交关系和兴趣相似度是密切相关的，将它们融合在一起进行社区发现是非常有意义的，而且结果也是理想的，本文所做工作也正是为了这些相互有影响的用户和关系。

3.3 三维社区发现

（1）与二维社区发现对比

对社交关系、地理位置关系和兴趣相似度通过 MDCD 算法进行三维融合社区发现。将三维融合和二维融合结果进行对比，如表 2 所示。

表 2　二维和三维网络的对比

网络类型 / 划分结果	二维融合网络	三维融合网络
有向模块度 Q	0.543	0.687
准确率	56.4%	62.4%
召回率	54.6%	51.3%

无论在有向模块度 Q 还是准确率上,三维的社区发现结果都要更好。三维的社区划分,不仅在社区结构上表现的更加好,而且在用户的准确率上也提升明显。召回率虽然有所下降,但其实差别不大,而且影响也不大。所以三维的融合社区划分有着更丰富的意义和理想的结果。

(2) 与传统多维社区发现对比

本文 MDCD 算法与 Lei Tang 等人提出的多维信息融合的社区发现方法 AMM 算法进行比较。其中,Lei Tang 等的方法中单维的社区发现算法选用效果非常好的 Blondel 等人提出的快速社区发现算法[9]。将两种算法同时运用在三维网络中,得到的准确率、召回率和 F-measure 值对比如表 3 所示。

表 3　两个方法的对比

算法 / 划分结果	AMM	MDCD
准确率	44.7%	62.4%
召回率	55.5%	51.3%
F-measure	0.495	0.563

可以得出,MDCD 算法在准确率上比 AMM 方法高 17.7%。本文的 MDCD 算法将不同维度网络的方向性考虑进去,并且进行了合理的方向分析,有效地规避了 AMM 方法的不足。从而得到了不错的社区结构和划分结果。

4　结　语

以往的社区发现方法大多都是基于单个维度的,它们只采用一个方面信息进行社区发现。在多维发现中,也没有将边的方向进行充分的考虑。本文在多维微博用户模型网络的基础上,提出了多维有向网络社区发现方法 MDCD,将多个维度综合考虑,并且考虑了边的方向,得到了良好的结果。在准确率、召回率和 F-measure 值上比单维和已有的多维方法有了一定的提升,并且社区结构也具有更重要的社会意义。

参　考　文　献

[1] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(9): 2658 - 2663.

[2] 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.

[3] Pengli Lu, Shenglong Zhang. Node Similarity Reveals Community Structure in Complex Networks[J]. International Journal of Advancements in Computing Technology, 2013, 5(3): 387 - 392.

[4] Tyler J R, Wilkinson D M, Huberman B A. E-mail as spectroscopy: Automated discovery of community structure within organizations[J]. The Information Society, 2005, 21(2): 143 - 153.

[5] 金弟,刘杰,贾正雪,等. 基于 k 最近邻网络的数据聚类算法[J]. 模式识别与人工智能,2010,23(4):546 - 551.

[6] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical review E, 2004, 69(6): 066133.

[7] Leicht E A, Newman M E J. Community structure in directed networks[J]. Physical review letters, 2008, 100(11): 118703.

[8] Chauhan S, Girvan M, Ott E. A network function-based definition of communities in complex networks[J]. Chaos: An Interdisciplinary Journal of Nonlinear Science, 2012, 22(3): 033129.

[9] Newman M E J. Community detection and graph partitioning[J]. EPL, 2013, 103: 330 - 337.

[10] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics Theory and Experiment, 2008, 30(2): 155 - 168

[11] Tang L, Liu H. Uncovering cross-dimension group structures in multi-dimensional networks[C]//SDM workshop on Analysis of Dynamic Networks. 2009.

[12] 潘建国. 基于语义的用户建模技术与应用研究[D]. 上海:上海大学,2009.

[13] 黄承慧,印鉴,侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报,2011,34(5):856 - 864.

[14] Chowdhury G. Introduction to modern information retrieval[M]. Facet publishing, 2010.

[15] 张博. 有向网络的社区发现算法研究[D]. 成都:电子科技大学,2013.

[16] Newman M E J. The mathematics of networks[J]. The new palgrave encyclopedia of economics, 2008, 2: 1 - 12.

[17] Rosvall M, Bergstrom C T. Maps of random walks on complex networks reveal community structure[J]. Proceedings of the National Academy of Sciences, 2008, 105(4): 1118 - 1123.

[18] 李金城. 大规模图像检索和识别中的神经网络学习及其应用[D]. 华南理工大学,2013.

(上接第 117 页)

[9] Wanying Ding, Xiaoli Song, Lifan Guo. A novel hybrid HDP-LDA model for sentiment analysis[C]//Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technology, Atlanta, USA, 2013, 1(1): 329 - 336.

[10] Kevin P Murphy. Machine Learning-A Probabilistic Perspective[M]. Cambridge, Massachusetts London, England: The MIT Press, 2012: 2 - 3.

[11] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[C]//International Conference on Learning Representations, 2013.

[12] Yoshua B, Rejean D, Pascal V, et al. A neural probabilistic language model[J]. The Journal of Machine Learning Research, 2003, 3(6): 1137 - 1155.

[13] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[C]//Neural Information Procesing Systems Foundation, 2013.

[14] Frederic M, Yoshua B. Hierarchical probabilistic neural network language model[C]//Proceedings of the international workshop on artificial intelligence and statistics, 2005.

[15] 董振东,董强. HowNet 情感词典[EB/OL]. [2013-07-28]. http://www.keenage.com.

[16] Ku Lunwei, Lo Yongsheng, Chen Hsinhsi. Using Polarity Scores of Words for Sentence-level Opinion Extraction[C]//Proc. of NTCIR-6 workshop meeting, 2007: 316 - 322.

[17] 张伟,刘缙,郭先珍. 学生褒贬义词典[M]. 北京:中国大百科全书出版社,2004.

[18] 中国计算机学会. 微博情感分析评测数据[EB/OL]. [2012-09-12]. http://tcci.ccf.org.cn/conference/2012/.