基于交互相似度的细粒度社群发掘方法

张 星 於志文 梁韵基 郭 斌
(西北工业大学计算机学院 西安 710129)

摘 要 发现在线社交网络中的社群结构有助于深入研究和分析信息传播规律,同时在社会推荐、群体特征发现等应用领域具有重要的意义。但是现有的社群结构发掘方法多忽略了用户之间的社会属性,导致获取的社群结构难以反映细粒度的结构特征。文中将用户的社会属性引入到社群结构发掘算法中。为了衡量用户的社会交互属性,提出了用户交互相似度模型。基于用户交互相似度模型,提出了一种面向在线社交网络的细粒度社群发掘方法。该算法可以有效衡量用户之间的社会属性,通过层次聚类的手段获得不同粒度的社群,并过滤无关数据。为了验证算法的有效性,以社交网站人人网的用户交互记录为数据集,比较了与其他社区挖掘算法的性能差异。实验结果表明,该方法发掘出的细粒度社群具有较高的准确性,在发现社群之间的不同话题上有着较好的应用。

关键词 交互相似度,细粒度,在线社交网络,社群发掘
中图法分类号 TP399 **文献标识码** J

Community Development Method Based on Interactive Similarity

ZHANG Xing YU Zhi-wen LIANG Yun-ji GUO Bin
(Department of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710129, China)

Abstract It was found that online social network community structure contributes to in-depth research of information propagation, social recommendation and the application of group identity discovery. Existing community structure excavation method ignores the many social attributes among users, which makes it difficult that the obtained community structure reflects the fine-grained structure. We combined user's social attributes into community structures excavated algorithm, then proposed the user interaction model in order to measure the user's social interaction properties. We proposed a community development method based on interactive similarity. The algorithm can effectively measure social attributes between users, get different size groups through hierarchical clustering, and filter noise data. In order to verify the effectiveness of the algorithm, we collected user interaction recorded from social networking site as data sets, compared the performance differences with other community mining algorithm. The experimental results show that this method discovers fine-grained communities with high accuracy, and may be used to discover different topic between communities.

Keywords Interactive similarity, Fine grain, Online social network, Community detection

1 引言

随着网络技术的发展,互联网正逐渐成为人们交往的重要渠道,人与人之间的社会网络关系蕴藏进了互联网中,诞生了以社交网站为代表的社交网络服务[1]。社交网站(如facebook、微博)上的用户通过显式交互(评论、点赞和转发)或者隐式交互(例如访问同一页面、关注共同话题),形成了一个错综复杂的社交网络,其中内部节点联系较为密切,而与其他节点联系较为稀疏的节点集合被称为社群[2]。研究社交网络的社区结构有助于发现社群中的热点话题和演化特征[3],进而有利于探索社交网络中的信息传播规律[4]。

社交网络的社群挖掘技术分为两类:1)面向显式网络的社群挖掘,基于网络中用户之间的相互关系,将社交网络映射为连通的拓扑图形,用图论相关技术发掘其中的子模块,或者利用概率模型计算出每个点属于不同社群的概率;2)面向间接网络的社群挖掘,基于网页中包含的文本信息,通过文档话题模型,得到不同网页之间的共同话题,从而得到对应的社群。

传统的社交网络社群发掘方法具有其局限性,解立群[5]分析了微博用户信息传播的网络拓扑关系,得出社交网络的"围观模型":名人间交流频繁,形成小圈子;大量普通人置于圈外"围观"形成大圈子。在围观模型中,使用用户之间的好友关系或者文本信息进行分析,不能有效区分大量共同关注了很多名人用户的普通用户,对社群中的噪音数据的过滤效果有限。张佰明[6]等研究了网络微博发展的根本逻辑,发现社交网络的用户形成了一个又一个的子社群,通过子社群之

到稿日期:2013-05-30 返修日期:2013-10-21 本文受国家重点基础研究发展计划(2012CB316400),国家自然科学基金(61222209,61103063),教育部高等学校博士学科点专项科研基金(博导类)(20126102110043),陕西省自然科学基础研究计划项目(2012JQ8028)资助。

张 星(1989-),男,硕士,主要研究方向为普适计算、社会感知计算,E-mail:357550902@qq.com;於志文(1977-),男,博士生导师,主要研究方向为普适计算、社会感知计算。

间的相互嵌套传播消息。因此，对社交网络进行更细粒度的社群划分，可以将原本用户特征不明显的普通用户群体划分为更多的子社群，进一步探究话题在普通人群之间的传播关系。

用户之间的交互操作对社交网络的拓扑结构具有重要的影响。沟通中形成的某种群体规范使不同的人具有了某种共同身份，这就形成了有实际意义的网络群体[7]。因此我们将用户的交互特征作为一个重要的社会属性，引入到社群结构发现工作中，从而实现细粒度社群挖掘的目的。具体地，文本首先介绍了如何度量用户的交互特征及交互特征相似度。然后基于用户交互相似度模型，提出了一种层次聚类算法发现细粒度的社群结构。该算法可以过滤与其他用户交互较少的噪音数据。同时，为了对该算法进行验证分析，我们从人人网获取真实的网络数据，对算法正确性进行了分析试验。

本文第 2 节介绍了相关工作；第 3 节对用户相关度模型和社区发掘方法进行说明；第 4 节介绍对本文提出的方法的实验，通过话题发现的结果说明细粒度社群划分的有效性；最后对接下来的工作进行了总结和展望。

2 相关工作

随着万维网技术的发展，更多的信息可以被在线处理，传统的社交网络挖掘技术也有了更大的需求，开始向在线社群发展，许多的相关技术和算法被研究了出来。Borgatti 等[8]探讨了社交网络分析在研究组织行为(organizational Behavior)中的应用，Garton 等[9]探讨了社会网络分析在虚拟社群(Virutal Community)领域中的应用，这些领域大多是基于互联网环境的。

在社交网络中，社群是一组节点，其中成员内部之间的联系远远多于其与网络的其余部分的联系。例如，Flake 等[10]将社群定义为一组指向内部的边高于指向外部网络的节点集合。Radicchi 等[11]也使用边介数计算的方法来定义强社群和弱社群。在以 Julien 等[12]为代表的一系列话题模型研究中，社群指的是具有共同话题的节点集合。

当前在在线社交网络中的社群发掘上已经存在着大量的相关工作。根据对象不同可以分为对显式网络和间接网络的社群发掘。

对显式网络的话题发掘是将社交网络中的用户关系映射为拓扑图，通过一定的计算方法，如迭代添加节点或者删除边，获得子分区，比如最小边切割[13,14]和模块化[15]。较为著名的是 Newman 和 Girvan[16]提出的 GN 算法，其根据网络中社群内部高内聚、社群之间低内聚的特点，逐步去除不相关社群之间的边，取得相对内聚的社群结构。概率模型也被广泛用于显式网络的社群发现。比如随机块模型[17]，假设链接生成的概率只有依靠社群的节点，计算每个节点属于对应社群的概率，将节点对应到概率最大的社群中。随机块的变体模型包括混合成员随机块模型[18]和贝叶斯随机块体模型[19]。

对间接网络的社群发掘指的是基于网络中不同网页的文本信息，发现共同话题的页面集合，从而对应出其背后隐藏的话题社群。普遍的方法是主题模型，其中每个主题自然地对应为一个社群。两个著名的主题模型是 PLSA[20]和 LDA[21]。

在上述工作中所研究的社群发掘均没有考虑利用社交网络用户之间的交互关系。本文工作的贡献在于：(1)将用户的交互特征作为一个重要的社会属性，引入到社群结构发现工作中，构造了描述社交网络中用户相关度的模型；(2)提出了相应的社群发掘算法，它能够实现更加细粒度的社群发掘。

3 细粒度社群发掘方法

为实现更细粒度的社群发掘，本文提出了一种基于交互相似度的细粒度社群发掘方法。该方法的具体流程包括用户相关度计算、社区关联矩阵构建、社群聚类树生成和面向不同粒度的社群分割。通过用户相关度模型计算用户之间的相关度，将社交网络中的用户之间交互操作转化为权值，进而生成社群聚类树，根据需求分割为不同粒度的社群，如图 1 所示。

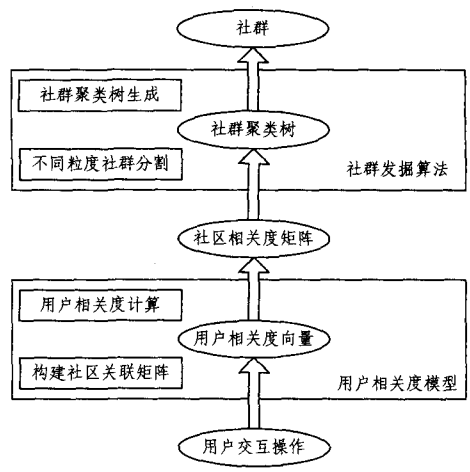

图 1 细粒度的社群发掘方法流程

3.1 用户相关度模型

网络社群中的用户通过社交网站为媒介进行联系。用户之间的相关程度，可以通过他们基于网站的强关联操作和弱关联操作体现。强关联操作指的是用户之间较为直接的交互关系，如转发、评论、分享等操作。弱关联操作指的是用户之间不太明显的交互关系，如关注同一个公共页面，经常到达相同的地理位置或共同使用同一个网站应用等等。

为了表示用户之间的相关程度，本文采用用户之间的几个常见的强关联操作作为依据来衡量用户之间的相关程度，具体如式(1)所示。

$$\eta_{ij}=\frac{\alpha_{ij}}{\sum_{k=1}^{n}\alpha_{ik}+1}\cdot h_1+\frac{\beta_{ij}}{\sum_{k=1}^{n}\beta_{ik}+1}\cdot h_2+\frac{\lambda_{ij}}{\sum_{k=1}^{n}\lambda_{ik}+1}\cdot h_3 \quad (1)$$

式中，η_{ij} 表示用户 i 和用户 j 的相关度，α_{ij} 表示用户 i 对用户 j 的评论次数，β_{ij} 表示用户 i 对用户 j 的转发次数，γ_{ij} 表示用户 i 对用户 j 的分享次数；h_1，h_2，h_3 分别表示评论、转发、分享这 3 种操作的权值。为了确定 h_1，h_2，h_3，我们分别随机抽取 N 条评论、转发和分享记录，分析发生在好友之间的评论、转发和分享的发生次数，分别记为 n_1，n_2，n_3，则 $h_1=n_1/(n_1+n_2+n_3)$，$h_2=n_2/(n_1+n_2+n_3)$，$h_3=n_3/(n_1+n_2+n_3)$，具体地，根据分析，本文中 $h_1=0.45$，$h_2=0.3$，$h_3=0.25$。

基于用户相关度，我们可以对社交网络中的任何两个用户之间的交互相似度进行度量，确定用户之间的强交互的相似程度。下面对于该问题进行形式化描述。对于一个含有 n 个用户的社群 Q，设其中的用户分别为 U_1，U_2，…，U_i，…，U_n，对于其中任意用户 U_i，通过式(1)的用户相关度公式，可以计算出他和其它 $N-1$ 个用户的相关度 η_{ij}；定义向量 $\vec{A}_i=(\eta_{i1}$，

$\eta_{i2},\cdots,\eta_{ij},\cdots,\eta_{in}$)为用户 i 的社群相关度向量，则该向量表示用户 i 对于社群中所有用户的相关度。

计算出社群中所有用户的相关度向量 A_i 后，定义矩阵 $RM_Q=(\vec{A}_1,\vec{A}_2,\cdots,\vec{A}_i,\cdots,\vec{A}_n)^T$ 为社群 Q 的相关度矩阵。矩阵 RM_Q 包含了社群 Q 中所有用户之间的相关度数据，反映了所有用户之间的相关性情况。

3.2 基于层次聚类的社群发掘算法

通过用户相关度模型，我们将网络社群内的用户之间的关系表示为向量空间的形式。如果矩阵中两个用户的相关度向量相似性越高，说明他们的交际圈重合度越高，则他们属于一个社群的概率也越大。因此，社群挖掘就转化为了向量聚类问题。将社群内的用户分为 N 个不同的社群，即将相关度模型中的用户矩阵聚类为 N 个向量集合。考虑到聚类的数量无法事先得知，本文提出了一种考虑噪音过滤的层次聚类方法，其可以根据需求，给定不同的聚类数量，得到不同粒度的聚类结果。

层次聚类方法是对给定的数据集中的数据，通过计算两两之间的距离，进行层次的分解或合并，直到某种条件满足为止。传统的层次聚类方法的结果是将一棵二叉树根据需求分割为若干数量的子树，这种方法的缺陷是对样本信息的完备性依赖太强，无法区分噪音数据。

为了解决这个问题，本文提出了一种改进的层次聚类算法，如图 2 所示。首先，基于采集的用户之间的交互操作信息，对于一个包含 n 个用户在线社交网络 Q，我们可以计算出其用户之间彼此的相关度矩阵 RM_Q。

$RM_Q=(\vec{A}_1,\vec{A}_2,\cdots,\vec{A}_i,\cdots,\vec{A}_n)^T$，是 n 个 n 维向量的集合。每个相关度向量 $\vec{A}_i$ 代表了社交网络中的用户 U_i 和其他所有用户的相关度信息，反映了用户 U_i 社交网络 Q 中的社会属性。因此，我们就把用户聚类转换成了向量聚类问题。

对 n 个相关度向量层次聚类，首先计算彼此之间的马氏距离，从而得到一个 $n*n$ 维的距离矩阵 DM_Q，显而易见，该矩阵相对于对角线对称且对角线元素为 0，其他的元素 d_{ij} 就代表了向量 $\vec{A}_i$ 和向量 $\vec{A}_j$ 之间的马氏距离。基于层次聚类的思想，我们选取 DM_Q 的下三角矩阵 DM_Q'，选取其中最小的元素 d_{pq}，这说明了向量 $\vec{A}_p$ 和向量 $\vec{A}_q$ 距离最接近，将这两个向量合并为一个新向量 $\vec{A}_r$，在 DM_Q' 中去掉对应于 $\vec{A}_p$ 和 $\vec{A}_q$ 的两行两列，再加上 $\vec{A}_r$ 与其余向量之间的距离，其中 $\vec{A}_r$ 与其他向量的距离取 $\vec{A}_p$ 和 $\vec{A}_q$ 与该向量距离的较小值。这样，就把用户 p 和用户 q 合并成了一个名为用户 r 的子树，反复重复此过程，最终会形成整个社交网络所有用户的聚类树。

在得到聚类树后，我们可以将这棵聚类树裁剪为任意数量若干子树，每个子树就对应了一个社群，通过聚类数量的改变就实现了不同粒度的社群发掘。具体裁剪方法是：不断地选取根节点距离最大的两个左右子树，将其拆分为两个子树，或者将根节点距离最小的两个子树合并，直到数量和给定的聚类数量相等。在裁剪的过程中，我们不断删除子节点数量小于一定阈值的子树，视其为游离于社群之外的噪音数据。

Process："基于层次聚类的社群发掘算法"
Input：用户之间相关度矩阵 RM_Q，社交网络置信度 p，用户数量 N，聚类数量 K
Output：聚类树集合 $C=\{Cluster_1,Cluster_2,\cdots,Cluster_K\}$
Begin：
1. for RM_Q 中的每个相关度向量 A
2. 计算 A_i 和其他向量之间的距离
3. 把结果写入向量距离矩阵 DM_Q
4. endfor
5. while DM_Q 不为空
6. 从 DM_Q 的下三角矩阵 DM_Q' 中寻找值最小的非对角线元素 d_{pq} 以 A_p 和 A_q 为左右孩子节点，合并出一个新向量 A_r
7. 在 DM_Q' 中去掉对应于 A_p 和 A_q 的两行两列，再加上 A_r 与其余向量之间的距离，A_r 与其他向量的距离取 A_p 和 A_q 与该向量距离的较小值
8. endwhile
9. 根据向量合并顺序，绘制出聚类二叉树 ClusterTree
10. 将 ClusterTree 加入聚类集合 C 中
11. while C 中的聚类集合数量不为 K
12. if(C 中的聚类集合数量小于 K)
13. 选取集合中根节点距离最大的聚类树拆为两棵子树
14. if(子树数量小于一定阈值)
15. 删除该子树
16. Else
17. 将子树添加到 C 中
18. endif
19. else if(C 中的聚类集合数量大于 K)
20. 选取集合中根节点距离最小的两棵聚类树合并为一棵子树
21. endif
22. endwhile
End

图 2 基于层次聚类的社群发掘算法流程

4 实验与分析

为了评估算法在细粒度社群发掘上的效果，基于社交网站采集的数据集，我们比较了在相同数据集下不同传统社群发掘方法和基于交互相似度的细粒度社群发掘方法在社群发掘的粒度差异。同时，在使用相同的层次聚类方法进行相同粒度社群发掘的前提下，我们比较了只考虑好友关系和考虑用户交互这两种方法的准确率。

在获得一个社交网络中的细粒度社群后，我们尝试发现网络社区中不同社群的话题分布，探究细粒度社群与话题传播的关系，从而证明细粒度社区发掘的有效性，进而通过社群之间明显的话题独立性间接说明划分结果的有效性。

4.1 数据采集

通过网络爬虫技术，我们从实名制社交网站人人网上采集了 953 个用户的信息作为实验数据。具体方法是：通过种子用户，依次进入该用户的好友页面，通过对网页 HTML 代码的解析，获取这些用户从 2009 年 11 月到 2013 年 1 月发布的全部新鲜事：包括状态、日志和照片等，共 182443 条，详见表 1。通过新鲜事信息中包含的评论列表、转发列表等操作，我们又获得了围绕这些新鲜事的交互信息 368883 条，详见表 2。

表 1 实验数据中新鲜事分布

NewsType	Status	Other	Photo	Video	Total
NewsNum	57756	6897	20542	96707	182443

表 2 实验数据中操作信息分布

OperType	Repost	Share	Reply	Other	Total
OperNum	28983	78273	261627	1002	369885

在本文中，我们选取了3个不同大学的学生作为用户，获取了他们所有好友的新鲜事和交互数据。社交网络中存在大量僵尸用户，他们虽然拥有社交网络账号但是几乎很少发布新鲜事，也很少和其他用户互动。为消除僵尸用户的影响，我们对数据进行预处理，从用户集合中抽取了其中操作信息在300条以上的190个活跃用户作为实验数据。由于人人网是个实名制网站，通过对用户的好友关系分析和个人资料分析，我们得到了这些用户的社交关系，可以粗粒度地分为4个社群。详细信息见表3。

表3　实验数据中的用户社交关系信息

集合编号	集合A	集合B	集合C	集合D	合计
用户数量	92	38	10	50	190
用户属性	通讯工程专业本科大班同学。年龄21岁左右	理工科院校的研究生，性别男，年龄25岁左右	文科院校的本科同学，性别女，年龄23岁左右	较为独立用户，和数据集中的其他用户几乎没有任何交互信息	无
交互特征	3种操作都很明显	评论、分享操作较为明显	评论操作较多，转发和分享操作较少	无	无

4.2　社群挖掘结果及分析

基于本文第4.1节采集的实验数据，我们计算了数据集中用户之间的交互信息，得到了这190人的社交网络中的用户相关度矩阵，并以此作为基于层次聚类的社群发掘算法的输入，比较了在几种常用聚类算法下，数据集的社群挖掘结果。聚类结果用不同颜色表示，黑色为算法过滤出的噪音数据，得到了图3所示的可视化表示。可以看出，在处理小规模社交网络时，本文定义的层次聚类算法可以将原本已经内聚的模块划分为更细粒度的社群。

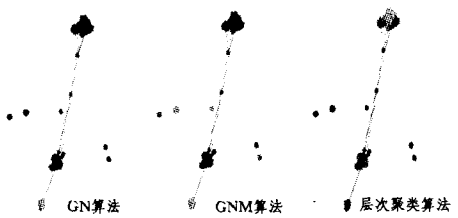

图3　不同算法下社群挖掘的结果

进一步地，我们比较了使用交互关系和不使用交互关系的层次聚类结果，通过对50位实验用户的调研，得到了两种聚类结果的准确度评价，见图4。可以发现，考虑了用户间交互关系的层次聚类方法在细粒度社群划分时具有较高的准确率，而只考虑好友关系的聚类，由于无法彻底刻画用户之间关系强弱，因此准确率较差。

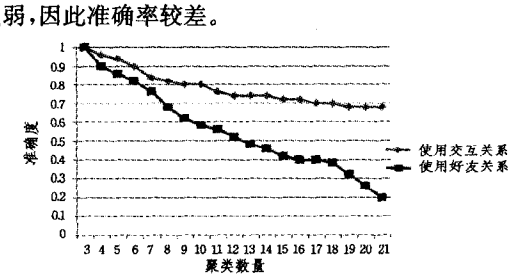

图4　使用交互关系和好友关系的社群发掘准确率

为了探究在线社群与网络话题之间的关系，在获得一个社交网络中的若干社群后，我们尝试发现网络社区中不同社群的话题分布，探究社群与话题传播的关系。我们把每个用户看作一个文档，利用分词软件把用户发布的新鲜事分割为词语集合，构建了话题-文档-词语的对应关系。利用Topic Model中经典LDA模型进行训练，得到了不同文档的话题分布，通过文档和用户的一一对应关系，最终得到了不同社群的话题分布。

图5比较了将社交网络划分为细粒度社群时，不同社群之间的话题分布情况。横坐标表现了不同的社群，纵坐标表现了该社群中每个话题的百分比。可以看出，当将该社交网络分割为一个个较小的细粒度社群后，不同社群之间的话题具有明显的独立性，从而可以更好地发现不同社群之间的独特话题。因此，发现在线社交网络中的细粒度社群，可以对在线社交网络的话题传播研究提供帮助。

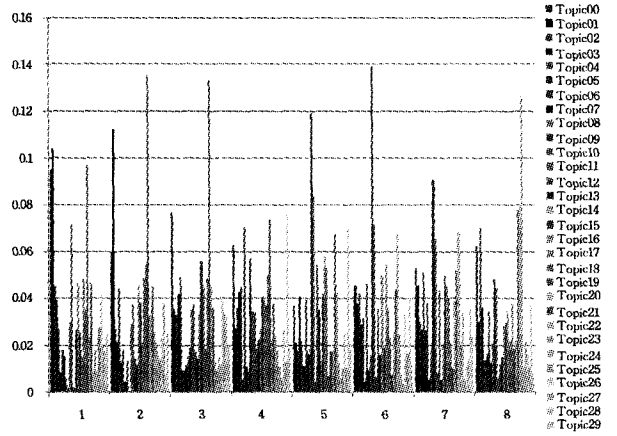

图5　细粒度社群之间的话题分布

结束语　本文提出了一种新的社交网络社群发掘方法——基于交互相似度的细粒度社群发掘方法，该方法基于用户间交互数据，建立用户相关度模型，结合层次聚类算法，有效地过滤噪音数据，得到细粒度的社群聚类结果。实验证明，该方法相对传统方法在细粒度社群划分上具有更好的性能。最后，经过实验分析，在线社交网络中的细粒度社群之间的话题具有较为明显的独立性，对于发现社群之间的话题传播有着较大的借鉴意义。本文中的实验数据选取的是社交网络上的活跃用户，未考虑交互关系较少的用户对模型的影响。同时，由于采用了层次聚类，算法的效率还有待提高。因此，下一步的研究将着力于在聚类的过程中如何更好地获取和利用非活跃用户之间的交互信息，改善社群发掘方法的性能，并基于细粒度社群的发掘，对社交网络中的话题传播及话题演变做相关工作。

参考文献

[1] http://zh.wikipedia.org/wiki/%E7%A4%BE%E4%BA%A4%E7%B6%B2%E8%B7%AF%E6%9C%8D%E5%8B%99

[2] Berger-Wolf T Y, Saia J. A framework for analysis of dynamic social networks [C]// Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2006: 523-528

[3] 阎艳. 关联网络的社区发现研究[D]. 重庆：西南大学，2009

[4] 徐渝. 高校BBS社群结构与信息传播的影响因素[J]. 西安交通大学学报，2007，27(1)：93-96

（下转第229页）

Springer,1996:204-218

[9] Salotti S, Ventos V. Study and formalization of a case based reasoning system using a description logic[C]// Lecture Notes in Computers Science 1488. Springer,1998:286-297

[10] d'Amato C, Fanizzi N, Esposito F. A semantic similarity measure for expressive description logics[C]// Proceedings of Convegno Italiano di Logica Computazionale(CILC05). Rome, Italy, 2005

[11] d'Amato C, Fanizzi N, Esposito F. A dissimilarity measure for ALC description logic[C]// Proceedings of the 21st Annual ACM Symposium of Applied Computing (SAC2006). Dijon, France,2006,2:1695-1699

[12] Fanizzi N, d'Amato C. A similarity measure for the ALN description logic[C]// Proceedings of Convegno Italiano di Logica Computazionale(CILC06). Bari, Italy,2006

[13] Janowicz K. Sim-DL: Towards a semantic similarity theory for the description logic ALCNR in geographic information retrieval [C]// Lecture Notes in Computers Science 4278. Springer, 2006:1681-1692

[14] Janowicz K, Wilkes M. SIM-DLA: A novel semantic similarity measure for description logics reducing inter-concept to inter-instance similarity[C]// Lecture Notes in Computers Science 5554. Springer,2009:353-367

[15] Gomez-Albarran M, Gonzalez-Calero P, Diaz-Agudo B, et al. Modelling the CBR life cycle using description logics[C]// Lecture Notes in Computers Science 1650. 1999:147-161

[16] Baader F, Calvanese D, McGuinness D. The description logic handbook: theory, implementation and applications[M]. Cambridge University Press,2003

[17] Brandt S, Küsters R, Turhan A Y. Approximating ALCN-Concept Descriptions[C]// Proc. of the 2002 Int. Workshop on Description Logics. 2002

[18] 常亮,王娟,古天龙,等. 时态描述逻辑 ALC-LTL 的 Tableau 判定算法[J]. 计算机科学,2011,38(8):150-154

[19] 蒋运承,汤庸,王驹,等. 面向语义 Web 的描述逻辑[J]. 模式识别与人工智能,2007,20(1):48-54

[20] 常亮,史忠植,陈立民,等. 一类扩展的动态描述逻辑[J]. 软件学报,2010,21(1):1-13

[21] Stanchev L. On efficient access to knowledge bases[C]// Proceedings of The 20th Midwest Artificial Intelligence and Cognitive Science Conference. Indiana University-Purdue University Fort Wayne, Fort Wayne,2009

[22] Sanchez-Ruiz A A, Ontanon S, Gonzalez-Calero P A, et al. Measuring similarity in description logics using refinement operators [C]// Lecture Notes in Artificial Intelligence 6880. Springer, 2011:289-303

[23] Larson J, Michalski R S. Inductive inference of VL decision rules [J]. ACM SIGART Bulletin,1977(63):38-44

[24] 鞠可一,周德群,吴君民. 混合概念格在案例相似性度量中的应用[J]. 控制与决策,2010,25(7):987-992

(上接第 218 页)

[5] 解立群,颜清华,陈颖. 从"围观模型"看交流困境——微博社会网络图谱分析[J]. 中国传媒科技,2011,8:92-95

[6] 张佰明. 嵌套性:网络微博发展的根本逻辑[J]. 国际新闻界,2010(6):81-85

[7] Backstrom L, Kumar R, Marlow C, et al. Preferential behavior in online groups[C]// Proceedings of the International Conference on Web Search and Web Data Mining. ACM,2008:117-128

[8] Borgatti S P, Foster P C. The network paradigm in organizational research: A review and typology[J]. Journal of management, 2003,29(6):991-1013

[9] Garton L, Haythornthwaite C, Wellman B. Studying online social networks[J]. Journal of Computer-Mediated Communication,2006,3(1)

[10] Flake G W, Lawrence S, Giles C L. Efficient identification of Web communities[C]// Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,2000:150-160

[11] 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

[12] Lacoste-Julien S, Sha F, Jordan M I. DiscLDA: Discriminative learning for dimensionality reduction and classification[C]// Advances in Neural Information Processing Systems (NIPS). Vancouver, British Columbia, Canada,2008,21

[13] Hofmann T. Probabilistic latent semantic indexing [C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM,1999:50-57

[14] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research,2003,3:993-1022

[15] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm[J]. Advances in neural information processing systems,2002,2:849-856

[16] Shi J, Malik J. Normalized cuts and image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2000,22(8):888-905

[17] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical review E, 2004, 69(2): 026113

[18] Newman M E J. Modularity and community structure in networks[J]. Proceedings of the National Academy of Sciences, 2006,103(23):8577-8582

[19] Nowicki K, Snijders T A B. Estimation and prediction for stochastic blockstructures[J]. Journal of the American Statistical Association,2001,96(455):1077-1087

[20] Airoldi E M, Blei D M, Fienberg S E, et al. Mixed membership stochastic block models for relational data with application to protein-protein interactions[C]// Proceedings of the international biometrics society annual meeting. 2006

[21] Hofman J M, Wiggins C H. Bayesian approach to network modularity[J]. Physical review letters,2008,100(25):258701