基于社交圈的在线社交网络朋友推荐算法

王 玓[1),2)] 高 琳[1)]

[1)](西安电子科技大学计算机学院 西安 710071)

[2)](西安电子科技大学经济与管理学院 西安 710071)

摘 要 为用户推荐朋友是在线社交网络的重要个性化服务.社交网站通过用户之间是否有相同属性信息或公共邻居判断他们能否成为朋友,但由于用户注册信息不完善和对公共邻居之间关系的忽略,推荐精度不高.事实上用户的朋友可以组成多个社交圈,拥有相似社交圈的用户更易成为朋友.因此,首先提出了社交圈检测算法,进而定义用户间的社交圈相似性,基于社交圈相似程度为用户推荐新朋友.使用YouTube数据验证了该文假设;使用Facebook自我网络数据,验证了社交圈检测方法的有效性,并与3种典型检测算法比较;使用区域Facebook数据,通过与公共邻居、Jaccard相似性比较,进一步验证了朋友推荐方法的准确性.

关键词 社交网络;社交圈;朋友推荐;社团发现;相似性;社会计算

中图法分类号 TP311 **DOI号** 10.3724/SP.J.1016.2014.00801

Social Circle-Based Algorithm for Friend Recommendation in Online Social Networks

WANG Yu[1),2)] GAO Lin[1)]

[1)](*School of Computer Science and Technology, Xidian University, Xi'an* 710071)

[2)](*School of Economics and Management, Xidian University, Xi'an* 710071)

Abstract Recommending friends to registered users is a crucial personal service of Online Social Networks (OSN). OSN will recommend a friend to a user if they share some common attributes or neighbors. But the recommendation accuracy is usually not so good since users' profile information may be incomplete and the relationships between neighbors are ignored. In fact, users can group their friends into several social circles and two users are more likely to become friends if they share similar social circles. Therefore, a social circle detection algorithm is suggested at first, and then the social circle similarity is defined. Based on this similarity, we can recommend friends to a user. Our hypothesis is verified by statistically analyzing the YouTube dataset. To verify the efficiency of the social circle detection algorithm, the ego networks of Facebook are used. The experimental results show that compared with three typical detection methods, our approach can identify social circles efficiently and accurately. We utilize social circle similarity, common neighbor similarity and Jaccard similarity to predict friend relationships in Facebook New Orleans network. The experimental results provide strong evidence that our algorithm is more precise in friend recommendation.

Keywords social network; social circle; friend recommendation; community detection; similarity; social computing

收稿日期:2013-06-20;**最终修改稿收到日期**:2014-01-20.**本课题得到国家自然科学基金**(60933009,91130006,61303122)、**陕西省社科基金资助项目**(11M016)、**中央高校基本科研业务费**(K5051106004)**资助**.**王 玓**,女,1980年生,博士研究生,讲师,主要研究方向为数据挖掘、复杂网络模块分析.E-mail:cheerwangyu@163.com.**高 琳**(**通信作者**),女,1964年生,教授,博士生导师,主要研究领域为计算生物信息学、数据挖掘、图论与组合优化算法及其应用.E-mail:lgao@mail.xidian.edu.cn.

1 引 言

随着 Web2.0 技术的成熟，在线社交网络(Online Social Network，OSN)，如 Facebook、Twitter、人人网等吸引了大量用户. 用户们不仅把现实生活中的人际关系搬到了网络上，还建立了与线下无关的单纯线上朋友关系，在社交网络上搭建起全新的沟通和分享信息的平台.

在社交网络提供的众多服务中，为注册用户推荐朋友是其中一项关键个性化服务，在提高用户体验和促进网络增长方面起重要作用[1]. 新朋友不仅能为用户提供新的信息来源、满足用户和其他具有共同兴趣爱好用户交流的需求[2]，还能帮助社交网络利用朋友关系推荐用户潜在感兴趣的服务项目，如 Facebook 的新闻反馈、LinkedIn 的产品推荐和 Dopplr 的旅游同伴等[3].

当朋友数量爆炸性增长时，用户在社交网络上的沟通效率反而会变得低下，大量朋友每天更新的信息都会推送到用户个人主页上，而用户可能只关心某些朋友关于某个主题的更新，如同事发布的工作相关内容、家人发布的生活相关消息、一起玩网游的朋友发布的游戏相关信息等. 为了过滤噪声、使用户快速发现他们真正需要的信息，朋友分组管理成为在线社交网络必须解决的问题.

2 相关工作

作为信息推荐问题的一种[4]，朋友推荐通常被转换为复杂网络上的链路预测问题[5]. 两个用户越相似，他们越有可能成为朋友. 社交网站在注册时要求用户填写个人资料，如性别、年龄、就读过的学校、工作单位等特征信息，可以利用特征信息的相似程度预测朋友关系. 但用户的特征信息常常是不完善的，考虑到隐私问题，用户可能不愿填写完整个人信息或填写虚假信息，网站为了简化注册过程，对此也不强求，所以仅利用注册信息推荐朋友的精确性较低. 除了特征信息，主要有 3 种方法计算用户相似性：基于局部信息的相似性、基于全局信息的相似性和基于随机游走的相似性. 其中基于全局和基于随机游走的相似性计算方法都需要整个网络的拓扑信息，不适合用于在线社交网络这种顶点规模巨大的网络上，而且朋友预测也不需要考虑整个网络上所有顶点. 因此在线社交网络都采用基于局部信息的相似性计算方法，主要是公共邻居相似性(Common Neighbors)及其不同的规范化. 公共邻居相似性定义如下 $s_{xy}=|\Gamma(x)\cap\Gamma(y)|$，其中 $\Gamma(x)$ 是用户 x 的邻居顶点的集合，即用户 x 的朋友集合. 如果考虑公共邻居的相对数量，即可得到 Jaccard 相似性 $s_{xy}=|\Gamma(x)\cap\Gamma(y)|/|\Gamma(x)\cup\Gamma(y)|$. 此类相似性最大的缺陷是仅考虑了公共邻居个数，忽略了公共邻居之间的关系. 我们注意到在社交网络中，用户的朋友会自然形成多个社交圈亲人、同事、同学等，如果两个用户拥有重叠程度很高的社交圈，那他们很可能会建立朋友关系. 为了利用社交圈发现潜在朋友，首先需要识别社交圈，即实现对朋友的自动分组.

朋友自动分组问题通常转换为用户自我网络(ego network)社团发现问题[6]. 自我网络是用户及其朋友以及他们之间的连接关系构成的网络. 如图 1 所示为某用户 A 的自我网络，用户 A 将他的朋友分成邻居、同事、大学同学和驴友 4 组，其中同事分组和大学同学分组有重叠.

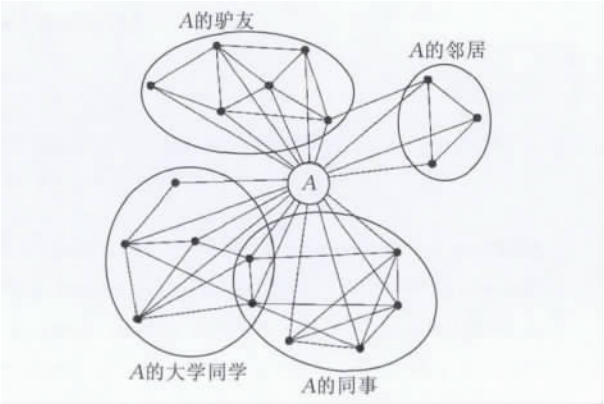

图 1 用户 A 的自我网络示意图

McAuley 等人[7]在对自我网络聚类时，除了网络拓扑结构，还考虑用户特征信息，认为拓扑连接稠密且特征信息相似的朋友们应属于同一社交圈，并在此基础上提出基于概率模型的社交圈发现算法. 该算法的主要缺陷是运行时间太长，当网络规模大于 1000 时，需要的运行时间以小时计算. Yoshida[8]提出的低秩嵌入算法能够聚类顶点有属性信息的网络，也可以用来发现自我网络中的社交圈，但该算法不能识别重叠圈子.

为了高效准确地预测朋友关系，本文首先提出一个基于关系的社交圈检测算法，快速聚类自我网络，完成社交圈的自动分类. 其次充分考虑公共邻居之间的关系，给出基于社交圈的用户相似性定义，并利用该相似性设计朋友推荐算法. 已有的研究将朋友分组与朋友预测看作两个独立问题，本文将这两

个问题结合在一起，在朋友推荐时充分利用朋友分组信息，不仅提高了推荐精度，还对新朋友所属社交圈给出建议，实现了朋友分组的自动更新.

3　朋友分组与预测算法

3.1　基于关系的社交圈检测算法

用户拥有不同社会角色，其朋友可以按照他们与用户不同的关系被分成多个社交圈：亲人、同事、同学等，同一圈子里的人相互联系紧密[7]. 通过对边进行聚类来识别社交圈更符合人们对社交圈的直观理解——由于拥有紧密相连的相同关系，人们才形成了社交圈. 边聚类能够有效发现重叠社交圈，因为一个顶点可以与多条边相连，当这些边属于不同圈子时，这个顶点相应也属于不同社交圈.

由于在线社交网络的朋友关系是逐步建立的，用户的某个真实社交圈可能会因为其朋友之间暂时还未形成稠密连接而被忽略. 为了发现这种“隐社交圈”，我们利用“有公共属性的用户更可能成为朋友并构成稠密社团”[9]这一理论，结合网络拓扑与用户特征信息，定义相邻边的相似性. 边 e_{ik} 与边 e_{jk} 的相似性定义为 $s(e_{ik},e_{jk})=\alpha\times ts(e_{ik},e_{jk})+(1-\alpha)\times fs(e_{ik},e_{jk})$，其中，$ts(e_{ik},e_{jk})$ 表示边 e_{ik} 与 e_{jk} 的拓扑相似性，$fs(e_{ik},e_{jk})$ 表示边 e_{ik} 与 e_{jk} 的特征相似性，参数 α 平衡两种相似性所占比重. 拓扑相似性定义为 $ts(e_{ik},e_{jk})=|n_+(i)\cap n_+(j)|/|n_+(i)\cup n_+(j)|$，其中 $n_+(i)$ 表示顶点 i 及其邻居的集合. 特征相似性定义为 $fs(e_{ik},e_{jk})=\cos(feature(i),feature(j))$，其中 $feature(i)$ 是表示用户 i 特征信息的特征向量. 可以把自我网络中用户及其朋友的所有特征信息集成到一个树型结构中，称为特征树. 图 2 是一个自我网络特征树的示意图. 在这个特征树的例子中，用户注册时需要填写职业、居住地、就读大学等信息，叶子节点表示该自我网络中至少有一个人具有这种特征. 利用特征树，自我网络中每个人的特征信息都可以用向量表示，向量长度与叶子数相等，当用户具有某个叶子节点描述的特征时，就将该用户特征向量的对应位置置 1，反之置 0. 例如用户 i 毕业于复旦大学，现在在上海做一名精算师，那么 i 的特征向量表示为 $feature(i)=(01001010)$. 需要说明的是，每个自我网络的特征树都不相同.

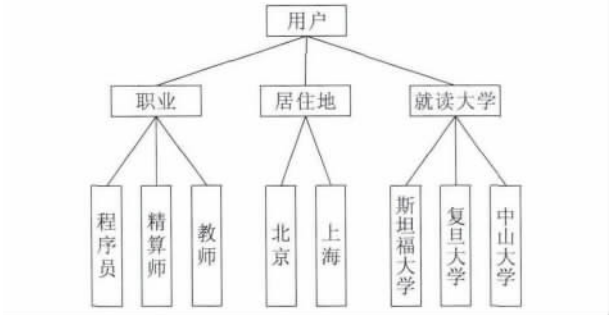

图 2　特征树示意图

利用连边相似度，可以用单链接分层聚类算法构建连边层次树图，并通过最大化划分密度函数来确定最优的社团划分[10]. 划分密度函数 $D=\frac{2}{M}\sum m_c\frac{m_c-(n_c-1)}{(n_c-2)(n_c-1)}$ 定义为一个划分下所有社团边密度的平均值. 社团 c 的边密度定义为 $D_c=\frac{m_c-(n_c-1)}{n_c(n_c-1)/2-(n_c-1)}$，这里 m_c 表示社团 c 中的边数，n_c 表示社团 c 中的顶点数. 该密度是在标准图密度的分子与分母上同时减去 $|n_c-1|$ 得到的，这种变化使得树型社团的密度为 0. 因为在随机网络中，树结构几乎处处可见[11]，所以在真实网络中，当检测到的社团是一棵树时，有理由认为它是没有意义的. 边密度函数使得树结构社团的密度为 0，让树结构社团在最大化划分密度函数时不做贡献，用这样的方法避免检测到没有意义的树型社团. 由于层次树图的其他分割线也能得到有意义的社团结构[12]，我们简化了边社团聚类，将相似性大于某个阈值的边的集合构成一个边社团，其导出的顶点集合（即与这些边相连的顶点）就构成了一个社交圈.

边聚类算法划分网络中的边，一些社团之间的边也会被看做是一个边社团. 在图 3 所示的网络中，直观上看应该包括 2 个社团：由顶点 1，2，3，4，5，6 构成的社团 1 和由顶点 7，8，9，10，11，12，13 构成的社团 2. 顶点 1，2 的相似度很高，所以边 $e_{1,7}$ 与边 $e_{2,7}$ 属于同一类；顶点 7，8 的相似度很高，所以边 $e_{2,7}$ 与 $e_{2,8}$ 属于同一类. 由于采用单链接分层聚类，最终边 $e_{1,7}$、边 $e_{2,7}$ 与边 $e_{2,8}$ 被聚为一个边社团. 虽然 3 条边是相似的，但它们的导出顶点集合（1，2，7，8）不符合社团“内紧外松”的特点，不能作为一个社交圈. 我们

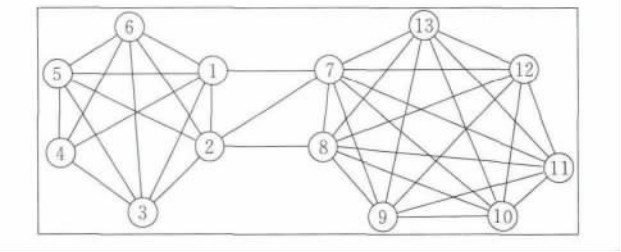

图 3　需要过滤的边社团示意图

用边聚集系数[13]过滤此类边社团，当一个边社团中大部分边的聚集系数都较小时，认为它的导出顶点集合不是社交圈. 由于这种边社团规模通常较小(大部分只包括 2 条边)，不需要对所有边社团进行判断，在本文实验中，仅对边数小于 4 的边社团进行筛选.

基于以上分析，提出基于关系的社交圈检测算法. 算法步骤如算法 1 所示.

算法 1. 基于关系的社交圈检测算法.

输入：用户 u 的自我网络 $G=(V,E)$；网络中所有顶点的特征向量；平衡参数 α；相似性阈值 θ；边聚集系数阈值 η

输出：用户 u 的社交圈，SC；

步骤：

1. 初始化集合 $SC=\varnothing$；
2. 计算图 G 中所有连边对的相似度；
3. 初始化集合 $Searched\_Edges=\varnothing$；
   //集合 $Searched\_Edges$ 中存储已知社团中的边
4. FOR $e_i \in E$
5. IF $e_i \notin Searched\_Edges$
6. 将 e_i 添加到 $Searched\_Edges$ 中；
7. 初始化集合 $Edge\_Circle=\varnothing$；//集合 $Edge\_Circle$ 中存储当前正在检测的社团中的边
8. 将 e_i 添加到 $Edge\_Circle$ 中；
9. FOR $e_j \in Edge\_Circle$
10. IF 存在 $e_k \in E$ 且 $e_k \notin Searched\_Edges$ 且 $s(e_k,e_j)>\theta$
11. 将 e_k 添加到 $Edge\_Circle$ 中；
12. 将 e_k 添加到 $Searched\_Edges$ 中；
13. END IF
14. END FOR
15. IF $size(Edge\_Circle)<4$ 且 $Edge\_Circle$ 中每条边的聚集系数都小于 η
16. $Edge\_Circle=\varnothing$；
17. END IF
18. 将 $Edge\_Circle$ 的导出顶点集合添加到 SC 中；
19. END IF
20. ENDFOR
21. 返回 SC

算法涉及 3 个参数：平衡参数 α、相似性阈值 θ 和边聚集系数阈值 η. 平衡参数 α 用于调节拓扑相似性与特征相似性所占比例，其取值与网络成熟度有关. 当自我网络比较成熟，即用户朋友之间的关系较为稳定时，α 取值较大；而当朋友关系正在完善，很多真实的朋友在网络中并未建立连接时，α 取值较小. 相似性阈值 θ 控制社交圈的规模，当 θ 增大时，检测到的社交圈数目增多. 边聚集系数阈值 η 用于过滤噪声社团，η 取值越小，过滤掉的社团越可能是噪声. 在后续实验中，取 $\alpha=0.5$，θ 等于网络中所有连边对的平均相似度，η 等于网络中平均边聚集系数的一半.

3.2 基于社交圈的朋友推荐算法

基于公共邻居的朋友预测只考虑了公共朋友的个数，周涛等人[14]进一步考虑了公共朋友间的关系，认为如果两个用户共同拥有的是一些经常联系的朋友，那么他们更可能成为朋友. 本文从社交圈的角度出发，更精细的分析公共朋友间的关系，认为如果两个用户的公共邻居在同一社交圈，那么他们更倾向于建立朋友关系. 以图 4 为例，在图 4(a)与图 4(b)中，顶点 1 与顶点 20 都有 3 个公共邻居，且这些邻居互为朋友. 在图 4(a)中，顶点 17、18、19 既是顶点 1 的一个社交圈也是顶点 20 的一个社交圈，由于同一个社交圈中的用户拥有相同关系，所以可以认为顶点 1 与顶点 20 也可能拥有这种关系. 在图 4(b)中，公共邻居 2、7、15 是顶点 20 的一个社交圈，但分别属于顶点 1 的 3 个不同社交圈. 这说明顶点 20 与顶点 2、7、15 彼此间有某种关系，而顶点 1 与顶点 2、7、15 没有这种关系，所以顶点 1 与顶点 20 可能也没有这种关系. 基于以上分析，有理由认为，

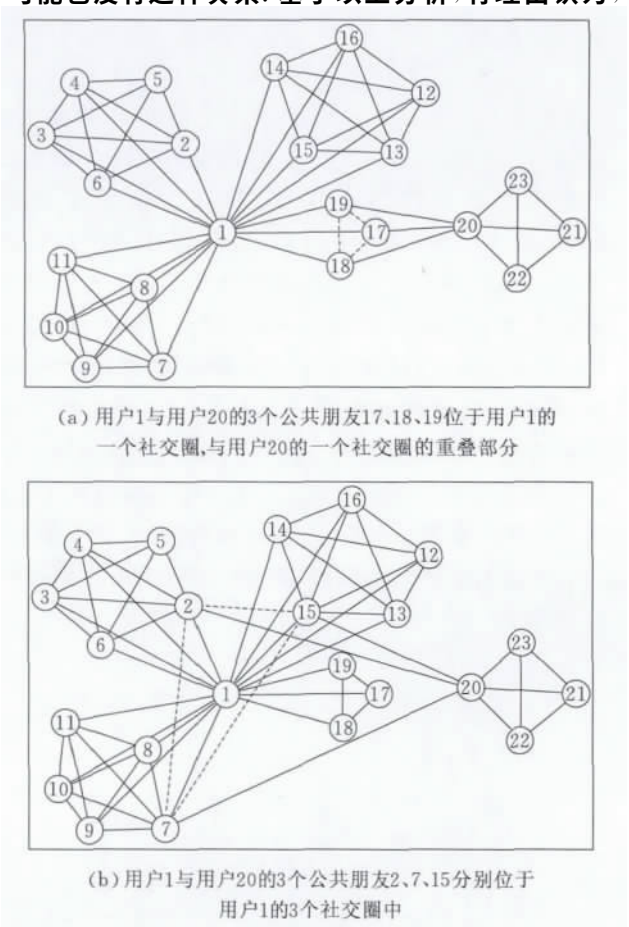

图 4 公共朋友关系示意图

图 4(a)中的顶点 1 与顶点 20 更可能建立朋友关系. 实验 1 进一步验证了该假设.

本文利用社交圈预测用户间的朋友关系，认为两个用户的重叠社交圈个数越多、社交圈重叠朋友越多、重叠圈子中的朋友连接越紧密，他们越可能成为朋友. 基于这个考虑，提出用户间的社交圈相似性. 将用户 u 的社交圈记为 $uc_1, uc_2, \cdots, uc_p$，用户 v 的社交圈记为 $vc_1, vc_2, \cdots, vc_q$，uc_i 与 vc_j 重叠顶点的导出子图记为 $overlap\_circle_{ij}$，n_{ij} 表示 $overlap\_circle_{ij}$ 中点的个数，m_{ij} 表示 $overlap\_circle_{ij}$ 中边的条数. 定义社交圈 uc_i 与 vc_j 的重叠度为 $n_{ij}\times(m_{ij}+1)$，用户 u 与用户 v 的相似性定义为 $s_{uv}=\sum_{i=1}^{p}\sum_{j=1}^{q}[n_{ij}\times(m_{ij}+1)]$.

利用该定义，给出朋友推荐算法，其步骤如算法 2 所示.

算法 2.　基于社交圈的朋友推荐算法.

输入：用户 u，用户 u 的社交圈 $uc_1, uc_2, \cdots, uc_p$，参数 k，朋友关系网络 $G=(V,E)$

输出：推荐给用户 u 的 k 个朋友及其所属社交圈

步骤：

1. 初始化集合 $candidate$，其中存储与 u 有公共邻居但无边相连的顶点；
2. FOR $i\in candidate$
3. 　检测 i 的社交圈；计算 s_{ui}；
4. END FOR
5. 将 $candidate$ 中与用户 u 相似性最大的 k 个顶点存储在集合 $protential\_friends$ 中；
6. FOR $j\in protential\_friends$
7. 　将 j 推荐给用户 u；
8. 　建议 j 应属于 u 的社交圈中与 j 的社交圈重叠度最高的圈子；
9. END FOR

利用在线社交网络中用户朋友能够分类成不同社交圈这一特性，推荐有相似社交圈的用户成为朋友，推荐算法不仅能预测朋友关系，还提供新朋友所属社交圈的建议.

4　实验验证与结果分析

4.1　验证实验

4.1.1　实验数据

本文使用 YouTube 用户数据集验证假设：有相似社交圈的用户更容易建立朋友关系. YouTube 是一个视频共享社交网络，用户可以相互建立朋友关系，创建并加入不同的兴趣小组，用户所加入的一个兴趣小组可以看作是用户的一个社交圈. 该数据来自美国东北大学 Alan Mislove 教授工作组(http://socialnetworks.mpi-sws.org)，包括 1 157 827 个用户、2 987 624 对朋友关系和 16 386 个兴趣小组.

4.1.2　实验结果

实验随机采样 2 000 000 对用户，统计在有 n 个公共邻居的条件下，一对用户有边相连的概率 p_{CN} 以及在有 n 个公共邻居且这 n 个公共邻居属于同一社交圈的条件下，一对用户有边相连的概率 p_{CS}. 实验结果如表 1 所示.

表 1　不同条件下一对顶点有边相连的概率比较

n	p_{CN}/%	p_{CS}/%	同比增长/%
3	4.07	7.04	73.0
4	5.70	8.50	49.1
5	7.32	10.05	37.3
6	8.92	11.60	30.0
7	10.47	13.12	25.3
8	12.02	14.63	21.7
9	13.52	16.10	19.1
10	14.87	17.51	17.8

从表 1 中可以看出，加上“公共邻居属于同一社交圈”的约束后，一对用户建立朋友关系的概率更大，也就是说，当一对用户有相似社交圈时，他们更容易成为朋友. 特别是当 n 较小时，连边概率显著增大. 实验结果很好地验证了本文假设.

4.2　社交圈检测实验

4.2.1　实验数据

社交圈检测算法的验证数据来自斯坦福大学 Leskovec 教授工作组(http://snap.stanford.edu)，包括 Facebook 中 10 组自我网络数据. 10 组自我网络中共包含 4039 个用户与 193 个社交圈.

4.2.2　评价标准

本文利用平衡误差率 BER(Balanced Error Rate)衡量一个预测社交圈 p 与一个真实社交圈 b 的相似程度[15]. BER 定义为

$$BER(b,p)=\frac{1}{2}\left(\frac{|p\backslash b|}{|p|}+\frac{|b\backslash p|}{|b|}\right).$$

通过比较预测社交圈集合 P 与真实社交圈集合 B 的相似性来验证社交圈检测算法的精确性. 预测社交圈集合 P 与真实社交圈集合 B 之间的相似性定义为

$$1-BER=\max_{f:P\to B}\frac{1}{|f|}\sum_{p\in dom(f)}(1-BER(p,f(p))),$$

这里 f 是集合 B 与集合 P 之间的部分映射，$|f|=\min(|B|,|P|)$. 当预测社交圈个数 $|P|$ 小于真实社

交圈的数目$|B|$时，对每一个$p\in P$都要找到一个对应的$b\in B$. 当$|P|>|B|$时，不惩罚多找到的社交圈，因为它们可能是还未被分组的社交圈. $1-BER$越大说明预测越准确.

4.2.3 实验结果

在自我网络数据集上，本文算法平均为每个网络检测到21个社交圈，社交圈的平均规模为23，非常接近真实情况.

实验比较了低秩嵌入[8]、概率模型[7]、边社团[10]与本文算法检测结果的精确性和算法在不同规模数据上的运行时间. 如图5所示，我们算法的$1-BER$值为0.79，比边社团算法高31.7%，比低秩嵌入算法高25.4%，比概率模型算法低6.3%. 虽然精确性比概率模型算法稍差，但运行时间远少于它.

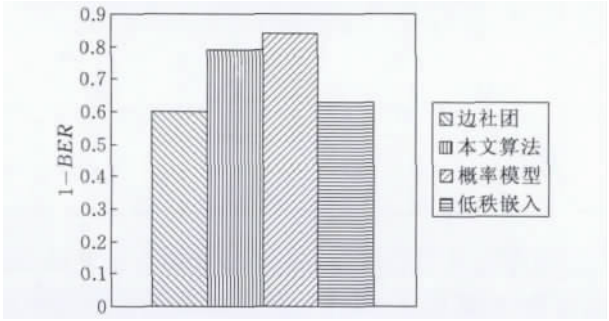

图5 4种算法的$1-BER$值比较

通过选择3个规模不等的自我网络，比较了4种算法在输入不同等级数据时的运行时间，实验结果如图6所示. 在顶点数为66的网络中，本文算法运行时间不到0.1 s，概率模型算法需要5 s；在顶点数为159的网络中，本文算法运行时间为1.8 s，概率模型算法需要76 s；在顶点数为1046的网络中，本文算法运行时间为240 s，而此时概率模型算法需要3 h. 由于对边的相似性定义更加复杂、对得到的社团过滤噪声，我们算法需要的运行时间大于边社团算法，在该实验的3个网络中，运行时间大约是边社团算法的2倍.

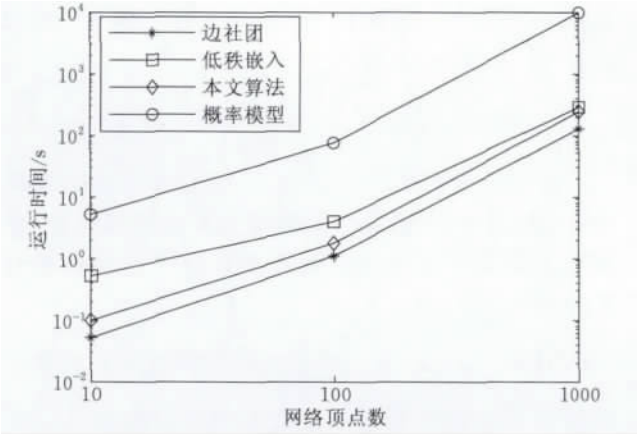

图6 4种算法运行时间的比较

实验结果表明，我们的算法可以很好地平衡精确性与有效性，能快速准确地识别用户社交圈.

4.3 朋友推荐实验

4.3.1 实验数据

朋友推荐算法的验证数据来自美国东北大学的Alan Mislove工作组(http://socialnetworks.mpi-sws.org)，数据采集了在2008年12月29日～2009年1月3日这一时间段中，新奥尔良地区的Facebook朋友关系，包括63731个用户和1545685个朋友关系. 本实验忽略了朋友关系的方向，将其简化为一个顶点数为63731，边数为817090的网络. 需要说明的是，该数据集没有顶点的特征信息，在社交圈检测时只利用了拓扑相似性.

4.3.2 评价标准

利用AUC与$Precision$指标评价朋友推荐算法.

为了验证朋友推荐算法的准确性，将网络中已知的连边集E分为训练集E^{T}和测试集E^{P}两部分，$E=E^{\mathrm{T}}\cup E^{P}$，$E^{\mathrm{T}}\cap E^{P}=\varnothing$. 在计算时只使用测试边的信息，并把不属于现有边集E的任意一对顶点之间的可能连边称为不存在的边. AUC从整体上衡量预测算法的精确度[16]. 每次随机从测试集中选取一条边与随机选择的不存在的边进行比较：如果测试集中的边的分数值大于不存在的边的分数值，就加1分，如果两个分数值相等就加0.5分. 这样独立比较n次，如果有n'次测试集中的边的分数值大于不存在的边的分数值，有n''次两个分数值相当，那么AUC定义为

$$AUC=\frac{n'+0.5n''}{n}.$$

如果所有分数都是随机产生的，那么$AUC=0.5$. 因此$AUC>0.5$的程度衡量了算法在多大程度上比随机选择的方法精确.

$Precision$只考虑排在前L位的边是否准确预测[17]. 如果排在前L位的边中有m个在测试集中，那么$Precision$定义为$Precision@L=\frac{m}{L}$. $Precision$越大说明算法越倾向于把真正的朋友关系排在前面.

4.3.3 实验结果

实验利用新奥尔良Facebook数据，通过10折交叉验证，比较了社交圈相似性、公共邻居相似性与Jaccard相似性在朋友推荐中的准确性. 3种相似性的AUC值分别为0.962(社交圈相似性)、0.938(公共邻居相似性)和0.934(Jaccard相似性)，社交圈

相似性的 AUC 值明显优于其他两种.

在使用 *Precision* 衡量预测精度时，实验分别取 $L=50$，$L=100$，$L=200$ 和 $L=500$，结果如表 2 所示，在每一种情况下，社交圈相似性都明显占优.

表 2　3 种相似性在不同 L 取值下的 *Precision* 值比较

Precison	公共邻居	Jaccard 相似性	社交圈相似性
$L=50$	0.532	0.482	0.628
$L=100$	0.485	0.437	0.576
$L=200$	0.468	0.422	0.495
$L=500$	0.314	0.340	0.402

实验结果表明，社交圈相似性可以更准确地预测朋友关系.

5　总结与展望

本文将用户的社交圈看作相似关系的集合，提出了基于关系的社交圈检测算法，算法结合用户特征信息与社交网络拓扑结构，定义了关系的相似性，通过聚类相似的关系，快速识别出重叠社交圈. 利用有相似社交圈的用户更可能成为朋友这一假设，定义了用户间的社交圈相似性，设计了朋友推荐算法，该算法能够对新朋友所属的社交圈给出建议，实现了朋友分组的自动更新. 最后用 Facebook 数据验证社交圈检测与朋友推荐算法的准确性，实验结果表明我们的方法能够快速检测社交圈、准确预测朋友关系.

在接下来的工作中，将考虑关系的权重与方向，将算法扩展到有权重的有向网络中.

致　谢　感谢 Mislove 教授，他为我们提供了用于实验验证的数据集！

参　考　文　献

[1] Yin Z, Gupta M, Weninger T, Han J. LINKREC: A unified framework for link recommendation with user attributes and graph structure//Proceedings of the 19th International Conference on World Wide Web. New York, USA, 2010: 1211-1212

[2] Dimicco J, et al. Motivations for social networking at work//Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work. California, USA, 2008: 711-720

[3] Guy I, Ronen I, Wilcox E. Do you know ? Recommending people to invite into your social network//Proceedings of the 14th International Conference on Intelligent User Interfaces. Florida, USA, 2009: 77-86

[4] Lv L, et al. Recommender systems. Physics Reports, 2012, 519(1): 1-49

[5] Lv L, Zhou T. Link prediction in complex networks: A survey. Physica A, 2011, 390(6): 1150-1170

[6] Liu Y B, et al. Simplifying friendlist management//Proceedings of the 21st International World Wide Web Conference. Lyon, France, 2012: 385-388

[7] McAuley J, Leskovec J. Learning to discover social circles in ego networks//Proceedings of the Neural Information Processing Systems. Lake Tahoe, USA, 2012: 548-556

[8] Yoshida T. Toward finding hidden communities based on user profiles//Proceedings of the IEEE International Conference on Data Mining Workshops. Sapporo, Japan, 2010: 380-387

[9] Mislove A, et al. You are who you know: Inferring user profiles in Online Social Networks//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York, USA, 2010: 251-260

[10] Ahn Y Y, Bagrow J P, Lehmann S. Link communities reveal multiscale complexity in networks. Nature, 2010, 466(7307): 761-764

[11] Bollobas B. Random Graphs. London: Academic Press, 2001

[12] Barthelemy M. Spatial networks. Physics Reports, 2011, 499(1): 1-101

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

[14] Liu Z, Zhang Q M, Lv L, Zhou T. Link prediction in complex networks: A local naive Bayes model. Europhysics Letters, 2011, 96(4): 48007

[15] Chen Y W, Lin C J. Combining SVMs with various feature selection strategies//Guyon I, Gunn S, Nikravesh M, Zadeh L eds. Feature Extraction. Springer Berlin Heidelberg, 2006: 315-324

[16] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters, 2006, 27(8): 861-874

[17] Herlocker J, et al. Evaluating collaborative filtering recommender systems. ACM Transactions on Information System, 2004, 22(1): 5-53

WANG Yu, born in 1980, Ph. D. candidate, lecturer. Her research interests include data mining, modularity analysis in complex network.

GAO Lin, born in 1964, Ph. D., professor, Ph. D. supervisor. Her research interests include bioinformatics, data mining in biological data, graph theory and intelligence computation.

Background

Online social networks (OSN) allow users to connect, communicate, and share information. In OSN, recommending friends to registered users is a crucial service since new friends provide new information source and satisfy their communication needs. OSN will recommend a friend to a user if they share some common attributes or neighbors. But the recommendation accuracy usually not so good since users' profile information may incomplete and the relationship between neighbors are ignored. However, too many friends lead to information overload, which makes categorizing users' friends a necessary task for OSN. One of the main mechanisms for users of social networking sites to organize their networks and the content generated by them is to categorize their friends into what we refer to as social circles. Once a user creates her circles, they can be used for content filtering, for privacy, and for sharing groups of users that others may wish to follow. Currently, users in online social networks identify their circles either manually, or by identifying friends sharing a common attribute. Neither approach is particularly satisfactory: the former is time consuming and does not update automatically as a user adds more friends, while the latter fails to capture individual aspects of users' communities, and may function poorly when profile information is missing or withheld.

In this paper, the authors first propose a relationship based social circle detection algorithm, which combines users' feature information with topological structure of the social network, forming friends with the same relationship into one social circle. Their method can detect overlapping social circles. Secondly, they define the users' similarity based on social circles and design a friend recommendation algorithm that provides suggestions about which social circle a new friend belongs to. Finally they apply our algorithms on Facebook data to validate its accuracy. Experiments show that compared with other methods, our algorithms can detect social circles effectively and predict relationship accurately.