复杂网络半监督的社区发现算法研究*

王静红[1,2]，于雅智[3]

(1. 河北师范大学 信息技术学院，石家庄 050024；2. University of Illinois at Champaign，Urbana Illinois 61801，USA；3. 石家庄理工职业学院，石家庄 050024)

摘 要：为提高社区发现算法的运行效率，提出了一种基于节点相似度的半监督社区发现算法——SSGN 算法。充分利用先验知识 must-link、cannot-link 约束集合，将先验信息通过衍生规则进行扩展，并对扩展的信息通过基于距离度量的方式加以验证。采用人工网络在 UCI 数据集和大型真实数据集上与真实网络进行验证，实验结果表明，基于节点相似度的半监督社区发现算法较其他半监督聚类算法更准确，也更高效。

关键词：广义社区发现；半监督聚类；社会网络分析；相似度；Girvan-Newman(GN)

中图分类号：TP301 **文献标志码**：A **文章编号**：1001-3695(2018)06-1663-05

doi:10.3969/j.issn.1001-3695.2018.06.014

Research algorithm of semi-supervised general community detection on complex networks

Wang Jinghong[1,2]，Yu Yazhi[3]

(1. *College of Information Technology*，*Hebei Normal University*，*Shijiazhuang* 050024，*China*；2. *University of Illinois at Champaign*，*Urbana Illinois* 61801，*USA*；3. *Shijiazhuang Institute of Technology*，*Shijiazhuang* 050024，*China*)

Abstract：Based on the similarity of the community detection methods GN algorithm has fast and accurate but has higher time complexity. In order to overcome the deficiency of GN efficiency，this paper presented a semi-supervised GN algorithm based on node similarity，took full advantage of the known node，must-link and cannot-link constraints，a priori information combined with the similarity information between nodes，and validated using artificial and real networks. It proves that the proposed algorithm reduces the GN algorithm's time complexity and improves the efficiency.

Key words：general community detection；semi-supervised clustering；social network analysis；similarity；GN

0 引言

随着信息时代的发展，复杂网络在大数据时代占据了主导地位。复杂网络将自然界中的每个实体抽象为网络中的节点，实体之间的关系抽象为网络中的边，对于社交网络即将网络中每个用户抽象为网络中的节点，用户之间的联系抽象为网络中的边。所谓社区即网络中存在着的簇结构，具有社区内部连接紧密、社区之间连接稀疏的特点。社区发现是将一个网络根据各节点之间的某种关系，将节点划分成若干簇的一个过程，其方法分为谱平分法[1]、基于图分割的社区发现方法、基于层次聚类或其他聚类的社区发现方法[2]、基于启发式思想的社区发现算法[3]、基于重叠社区的社区发现方法以及社区发现的其他方法[4]。

GN 算法是经典的网络社区发现方法，是不断地从网络中移除介数最大(full-betweenness)的边，边介数定义为网络中经过每条边的最短路径数目，算法的复杂度非常高[5]。针对计算边介值(betweenness)造成的算法时间复杂高的问题，Tyler 等人[6]采用蒙特卡洛方法对部分连边的边介值进行估算，一定程度上提高了算法的运行效率，但其准确性有所降低。在对真实社交网络进行社区划分过程中，大多数情况下真实网络的社区数量是未知的。为更好地确定社区划分状态，2004 年 Newman 等人[7,8]提出了一种基于贪婪法思想的凝聚算法 fast-Newman，该算法每次社区合并都是在使模块度增多最大和减小最少的方向进行。该算法总的复杂度增高，对于较为稀疏的复杂网络其时间复杂度降低，其中包含网络中节点的个数，以及网络中边的条数。同年 Radicchi 等人[9]提出了 self-contained GN algorithm，该算法以强社区结构和弱社区结构为理论基础，制定两者的衡量标准并提出边聚集系数概念，确保在不降低社区划分的准确性的同时，提升算法的运行效率。2008 年，Gregory[10]对传统 GN 算法进行了改进，采用对复杂网络的局部进行边介值计算方式划分社区，其算法在提高效率的同时降低了社区划分的准确性。为解决 GN 算法运行效率低的问题，朱小虎等人[11]2010 年提出了 MEA 算法，该算法将 modularity 增量作为社团结构的度量标准，基于 modularity 极值近似再使用贪心算法得到最优解。为解决串行社区发现方法速度慢、效率低以及不能很好地划分数据量大、信息种类多的复杂网络的问题，2012 年杨立文提出了基于传统 GN 算法的并行化社区发现方法，并从计算边介值的角度，将并行化 GN 社区发

收稿日期：2017-01-26；**修回日期**：2017-04-21 **基金项目**：河北省自然基金资助项目(F2013205192)；国家自然科学基金及河北省科技厅、河北省教育厅重点项目(61672206，14214504D，ZD2018023，KSZX201433，202213)

作者简介：王静红(1967-)，女，教授，博士，主要研究方向为机器学习、复杂网络等(wangjinghong6301@163.com)；于雅智(1990-)，女，硕士，主要研究方向为数据挖掘、社区网络.

现算法分为粗粒度并行策略和细粒度并行策略,仍属于无监督范畴。2013 年徐杨等人[12]提出了适用于微博社区的识别方法,该算法以微博用户和用户之间的关系为基础,且构建微博网络社区模型,并在此基础上利用 GN 算法对微博用户进行社区划分。2015 年,高庆一等人通过使用从属度的概念描述节点对不同社区的紧密程度,结合 MapReduce 模型改进算法,采用并行方式分析,从而实现社区划分。

以上社区发现算法都仅仅能发现网络中的社区结构(社区内节点链接紧密、社区间节点链接稀疏),在网络中没有社区或存在其他类型结构属于无监督范畴,无法处理事先给定的半监督信息(先验知识),采用无监督学习对无类标签样例进行标记将耗费大量的人力物力且耗时长,最终的标记结果准确性较差。半监督学习利用的先验知识是无噪声干扰的数据,根据先验知识对无类标签样例进行标记,大大缩减了标记成本同时提高了运行效率。

针对上述问题,本文提出一种基于节点相似度的半监督的社区发现算法。分析和比较已知不同数量的先验知识在社区发现时的性能差异,将半监督和相似度计算方法相结合,以取代传统 GN 算法中影响时间复杂度的关键因素——边介值的计算,从而降低 GN 算法的时间复杂度,提高算法的运行效率。在线社交网络每天产生庞大、繁杂的网络数据,挖掘和分析其潜在结构可为决策者提供强有力的决策依据。

1 相似度的构造

基于网络相似度的社区发现方法分为基于边的相似度以及基于节点的相似度,本文利用节点的相似度,考虑节点与其邻居节点的关系,社区内节点相似性高,社区间节点相似性低。

定义 1 假设网络中的两个节点有着相同或者相近的邻居节点,那么这两个节点被认为是相似的。这里要考虑节点的邻居信息,同时增加参数,通过参数的扰动来解决网络节点聚类受噪声链接影响易形成不平衡簇的问题。

节点相似度定义的四种形式如式(1) ~(4)所示,其中假设 τ_i 表示节点 i 的邻居集合,$|\tau_i|$表示该集合的势即集合元素的个数,$|\tau_i \cap \tau_j|$表示节点 i 和 j 共有的邻居个数。

$$S_j(i,j) = \frac{|\tau_i \cap \tau_j|}{|\tau_i \cup \tau_j|} \tag{1}$$

$$S_c(i,j) = \frac{|\tau_i \cap \tau_j|}{\sqrt{|\tau_i||\tau_j|}} \tag{2}$$

$$S_m(i,j) = \frac{|\tau_i \cap \tau_j|}{\sqrt{\min(|\tau_i|,|\tau_j|)}} \tag{3}$$

$$S_{nj} = \begin{cases} S_j(i,j) & i,j \in E \\ S_j(i,j) - \sigma & \text{otherwise} \end{cases} \tag{4}$$

在复杂网络中,两个节点的共同邻居节点越多,这两个节点越相似即归属为一个社区。式(2)主要应用于文献索引网络;式(4)表示如果网络中两个节点之间不存在连边,则在 S_{nj} 相似度构造的基础上再减去一个惩罚项 σ,由此可以构造出一种新的相似度构造方法[13,14]。

本文相似度的计算是利用数据集中的每一个数据点为网络中的一个节点,将距离近的或者相似度大的两个数据点 i 和 j 之间添加一条边,充分利用先验知识的约束集合,将先验信息结合节点间的相似度,构造网络节点的相似度矩阵。那么聚类问题转换为社区划分问题,可以利用任意一种社区划分方法对网络进行划分从而得到对应的聚类结果。

2 基于相似度的半监督 SSGN 算法

随着监督学习与无监督学习向半监督学习(SSL)的转型,充分利用先验信息指导聚类过程已成为半监督聚类的主导[15~21]。传统 GN 与 SGN 都属于无监督学习范畴,无法处理事先给定的半监督知识[17,22,23],本文提出了基于相似度的半监督聚类新算法,记为 SSGN(Givern-Newman based on similarity and semi-supervised)。采用半监督学习方法,对信息的处理过程是通过添加 must-link 节点、cannot-link 节点,使得网络趋于明显,从而提高了算法的精确度[24~26]。

针对传统 GN 算法的高复杂度,提出的 SSGN 算法的基本思想是利用计算顶点间的相似度代替计算边介值,从而改善时间复杂度。如果网络中两个节点之间存在连边,那么用这两个节点间的相似度替代边介值作为该条边的指标再对网络进行分裂得到 SSGN。SSGN 算法利用已知的和学习到的先验信息中节点对的关系,根据不同的相似度构造方法修改网络初始矩阵中对应的值。通过迭代计算节点相似度值,删掉相似度值最小的边至划分成最佳社区状态。在人工网络和真实网络中,SSGN 算法降低了算法的时间复杂度,又进一步提高了运行效率和性能。

定义 2 给定一个数据集 $X = \{x_1, x_2, \cdots, x_n\}$ 和一个点对约束集合 $C = C_= \cup C_{\neq}$,其中,$C_=$ 为必连约束集合,它的元素 $c = (x_i, x_j)$ 表示 x_i 和 x_j 属于同一个簇;$C_{\neq}$ 为不连约束集合,它的元素 $c \neq (x_i, x_j)$ 表示 x_i 和 x_j 属于不同的簇。

定义 3 网络结构图 $G(V,E)$,V 表示网络中所有节点的集合($v \in V, u \in V$),E 表示网络中所有边的集合,设网络中共有 n 个节点,m 条边。

网络中社区表示为 $C = \{C_1, C_2, \cdots, C_K\}$,$\bigcup_{i=1}^{k} C_i = V$ 且 $C_i \cap C_j = \varnothing, i \neq j, i = 1,2,\cdots,k$。

定义 4 划分的社区满足社区间联系稀疏,社区内联系紧密,表示为

$$\sum_{l=1}^{k} |\{(u,v) | (u,v) \in E, u \in C_l, v \in C_l\}|$$

$$\sum_{\substack{i=1 \\ j=1}}^{k} |\{(u,v) | (u,v) \in E, u \in C_i, v \in C_j, i \neq j\}|$$

定义 5 Must-link 约束,$C_{\mathrm{ML}}: \forall v_i, v_j \in V, (v_i, v_j) \in C_{\mathrm{ML}}$,保证 v_i 和 v_j 属于同一社区内;cannot-link 约束,$C_{\mathrm{CL}}: \forall v_i, v_j \in V, (v_i, v_j) \in C_{\mathrm{CL}}$,保证 v_i 和 v_j 不属于同一社区内。

本文的测试网络为具有先验信息的网络,根据先验信息中属性的划分确定 must-link 约束节点对以及 cannot-link 约束对。例如在含有先验信息的网络中,将某信息的创造者以及粉丝量大的用户定义为 must-link 约束节点,表示在一个集合,将与不含该信息的用户定义为 cannot-link 约束节点,然后根据衍生关系规则,扩展节点标记,由节点相似度进行社区划分。

SSGN 算法是基于节点相似度思想的一种层次分裂算法,基本过程为不断地删除网络中具有相对于所有源节点的相似度最小的边,再重新计算网络中剩余边端点之间的相似度,重

复此过程至所有边都被删除。其中充分利用先验知识 must-link 约束、cannot-link 约束集合,采用宽度优先搜索,将先验信息结合节点间的相似度,重新构造网络节点的相似度矩阵。

算法流程如下:

a)输入:要进行社团检测的网络 $G(V,E)$、已知 must-link 约束集合、cannot-link 约束集合。

b)根据衍生关系规则对其他未标记节点进行标记,标记规则如下:

$$(v_i,v_j)\in C_{ML}\text{且}(v_j,v_k)\in C_{ML}\Rightarrow(v_i,v_k)\in C_{ML}$$

$$(v_i,v_j)\in C_{CL}\text{且}(v_i,v_k)\in C_{ML}\Rightarrow(v_k,v_j)\in C_{CL}$$

至无节点可作标记。

c)选择 C_{CL} 集合中节点,删除两节点连接的边,至所有标记节点都被删掉。

d)计算网络中每一条边两个端点之间的相似度(添加),选择相似度最小的边删除。

e)重复步骤 d),计算剩余边端点之间的相似度至所有边被删除[27,28]。

流程如图 1 所示。

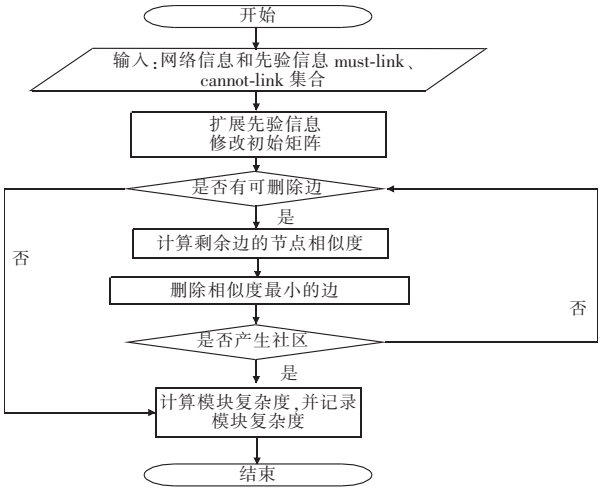

图 1　算法流程图

3　实验分析

实验运行环境:硬件配置为 Intel 电脑 2.0 GHz 处理器,2 GB内存;操作系统为 Windows 7;编程环境为 MATLAB 7.9.0.592。

3.1　数据集测试

人工数据采用 LFR 标准生成[18,29],规则如下:

网络 1。132 个节点,262 条边,4 个社区,最终社区划分结果如图 2 所示。

网络 2。1 000 个节点,10 000 条边,18 个社区。

社区内部节点按照概率 P_{in} 随机添加,社区之间节点按照概率 P_{out} 随机添加,两个概率的取值保证 $Z_{in}+Z_{out}=18$。其中,Z_{in} 表示节点与社区内部节点连边的平均值,Z_{out} 表示节点与社区外部节点连边的平均值。因此,Z_{in} 越大,表示社区结构越明显,Z_{out} 越大,表示社区结构越模糊。由于程序生成已知的社区结构,实验以最终划分结果相比真实结果的准确率(accuracy)作为评价指标,算法准确度量标准为 NMI,NMI 值越大表示算法准确率越高。算法运行的时间为 time,该值越小则表示算法的运行效率越高。

为了比较 SSGN 算法与 GN、SGN 算法在具有不同清晰度的传统社区网络上的准确率和运行效率,本实验分为两个部分:首先采用人工数据对 SSGN 算法进行训练,测试算法随着已知 must-link、cannot-link 约束条件数量变化时准确率和运行时间的变化规律。验证算法参数定义为:假设网络的节点个数为 n,边数为 m,聚类次数为 k,算法的迭代次数为 t,cannot-link、must-link 约束集合元素的数量分别为 X、Y,总数为 N。人工实验中假设两者数量等同,表 1、2 中 SSGN 算法默认 $N=8$。

实验结果如表 1 ~ 4 所示。

表 1　SSGN 算法与其他算法在人工网络 1 上的时间复杂度对比

对比项	GN	SGN	SSGN
NMI	1	1	1
time/s	2 354.06	188.45	134.32

表 2　SSGN 算法与其他算法在人工网络 2 上的时间复杂度对比

对比项	GN	SGN	SSGN
NMI	1	1	1
time/s	25 672.3	2 365.2	1 543.4

表 3　SSGN 算法给定不同数量的 CML、CCL 集合在人工网络 1 上的时间复杂度对比

对比项	$N=8$	$N=10$	$N=20$
NMI	1	1	1
time/s	134.32	55.68	8.23

表 4　SSGN 算法给定不同数量的 CML、CCL 集合在人工网络 2 上的时间复杂度对比

对比项	$N=8$	$N=10$	$N=20$
NMI	1	1	1
time/s	1 543.4	1 037.79	589.3

由结果对比表可知,所有算法在精度一致的情况下,GN 算法的时间复杂度最高,SGN 算法的时间复杂度较短于 GN 算法,SSGN 算法的运行时间明显短于 GN 算法,较 SGN 算法稍有提高。

图 2 表示在人工网络一定且 SSGN 算法及与 SGN 算法选取同样的(4%)先验信息的情况下,对比五种算法的准确率。实验得知,GN 算法准确率始终优于其他算法,改进的 SSGN 算法优于 SGN 算法,且两者均优于 fast-Newman 和 CNM 算法。由结果对比表可知,所有算法在精度一致的情况下,随着 cannot-link 节点对于 must-link 节点对个数的增多,SGN 算法运行时间随之缩短。值得注意的是,在选取 must-link、cannot-link 约束集合中,所选取的数据具有代表其运行效率越高,在实际应用中,选取确定的属性越多则分类越明显。

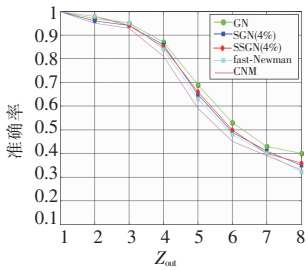

图 2　各种算法准确率对比

3.2　真实网络分析

真实网络数据采用经典网络 dolpins(海豚关系网)、karate

(Zachary 空手道俱乐部网络[14])、football(美国大学足球赛网络)和 books on politics(美国政论著作网络)(数据采集参照 http://www.orgnet.com/)。所选用的经典网络都有清晰的社区结构,先介绍如下:

a) Dolpins[16] 由 2 个社区、62 个节点、382 条边组成,其中每个节点表示海豚个体,社区表示海豚群体,节点之间的连边表示两个海豚之间接触频繁。网络初始化结构如图 3 所示。

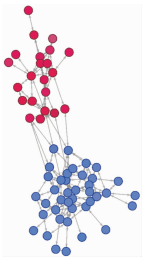

图 3 dolpins 网络初始化网络结构

b) Karate 由 34 个节点、192 条边组成,每个节点表示一个俱乐部的成员,两节点之间的边表示两成员之间的社会交往关系,该俱乐部主要分为两个社区,一个为俱乐部主管,另一个为校长,因此俱乐部成员就以俱乐部主管和校长为中心分成了两个社区,初始化结构如图 4 所示。

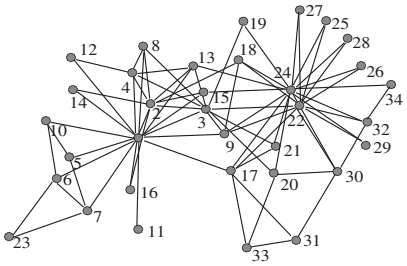

图 4 karate 网络初始化网络结构

c) Football 由 132 个节点、1 343 条边组成,每个节点表示一支队伍,两节点之间表示连接此边的两个队伍正在比赛,全部队伍被分为 12 个联盟,即 12 个社区,与联盟内球队进行比赛的概率大于与联盟外球队比赛的概率。

d) Books on politics 由 42 个节点、127 条边构成,每个节点表示一本书,每本书均为亚马逊网站出售的关于美国政论的著作,两节点间的边表示顾客同时购买了两本书,整个网络根据政治见地分成三派(三种社区),即自由派、中立派和保守派。网络初始化结构如图 5 所示。

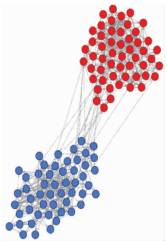

图 5 books on politics 网络初始化网络结构

针对各网络采用人工网络的分析方法,计算各算法的准确性结果如表 5、6 所示。

表 5 SSGN 算法与其他算法在各真实网络上的 NMI 对比

算法	dolpins	karate	football	books on politics
GN	1.000 0	0.960 1	0.900 1	0.798 0
SGN	1.000 0	1.000 0	0.987 8	0.879 2
SSGN	1.000 0	1.000 0	1.000 0	0.932 3

表 6 SSGN 算法给定不同数量的 CML、CCL 集合在各真实网络上的 NMI 对比

N	dolpins	karate	football	books on politics
$N=8$	1.000 0	1.000 0	1.000 0	0.932 3
$N=10$	1.000 0	1.000 0	1.000 0	0.956 7
$N=20$	1.000 0	1.000 0	1.000 0	1.000 0

通过实验表明,在精度条件相同时,当选取的关键节点越多,运行时间随之缩短,但是当选取的关键节点数量增加到一定量时,则算法的运行时间不会缩短,运行时间没有很大的变化,表示把关键节点都筛选一遍。总之,对于复杂网络社区划分问题,SSGN 算法在准确性一致的情况下较传统 GN 以及 SGN 算法的运行效率以及精确性上均有所提高,在选取 CML、CCL 集合数量上,选取越多的 CML、CCL 集合算法的时间复杂度越低,精确度越高。

4 结束语

本文根据目前的相似度构造方法,提出基于节点相似度的半监督 SSGN 算法,充分利用先验知识 must-link、cannot-link 约束集合,将先验信息结合节点间的相似度,重新构造网络节点的相似度矩阵,降低 GN 算法重复计算边介值引起的时间复杂度高的问题且提高了准确性。在真实网络中,对于 must-link、cannot-link 约束集合元素的数量以及质量对 SSGN 算法有决定性作用,对复杂网络定义越清晰,属性关系确定得越科学,在一定程度上进一步提高了算法的运行效率和准确性。

本文在测试中应用的数据主要为非重叠复杂网络,对重叠网络本算法处理方法是:根据划分的社区结果,重新计算社区内较为模糊节点的相似度,根据相似度值重新划分社区。但对重叠网络的边界节点划分的准确度有待提高。下一步将深入研究在重叠网络中 must-link 以及 cannot-link 约束节点,提高在重叠网络中的社区划分的准确度。

参考文献:

[1] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: analysis and an algorithm[C]//Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2001: 849-856.

[2] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 2004, 69(Pt2): 066133.

[3] 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.

[4] Sun Penggang, Gao Lin, Han Shanshan. Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks[J]. Information Sciences, 2011, 181(6): 1060-1071.

[5] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12): 7821-7826.

[6] Tyler J R, Wilkinson D M, B. Huberman A. Email as spectroscopy: automated discovery of community structure within organizations[C]// Proc of Communities and Technologies. Netherlands: Kluwer, 2003: 81-96.

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

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

[9] 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.

[10] Gregory S. Local betweenness for finding communities in networks, Technical Report CSTR-08-004 [R]. Bristol: University of Bristol, 2008.

[11] 朱小虎,宋文军,王崇骏,等. 用于社团发现的 Girvan-Newman 改进算法[J]. 计算机科学与探索,2010,4(12):1101-1108.

[12] 徐杨,蒙祖强. 基于 GN 算法的微博社区识别方法[J]. 广西大学学报:自然科学版,2013,38(6):1413-1417.

[13] Yang Liang, Cao Xiaochun, Jin Di, *et al*. A unified semi-supervised community detection framework using latent space graph regularization [J]. IEEE Trans on Cybernetics, 2015, 45(11): 2585-2598.

[14] Liu Zhiyuan, Li Peng, Zheng Yabin, *et al*. Community detection by affinity propagation [EB/OL]. 2008. http://nlp. csai. tsinghua. edu. cn/ ~ lzy/techreports/TR001_thunlp_community_detection. pdf.

[15] Ver Steeg G, Galstyan A, Allahverdyan A E. Statistical mechanics of semi-supervised clustering in sparse graphs[J]. Journal of Statistical Mechanics Theory and Experiment, 2011 (8): DOI: 10. 1088/1742-5468/2011/08/P08009.

[16] Ma Xiaoke, Gao Lin, Yong Xuerong, *et al*. Semi-supervised clustering algorithm for community detection in complex networks [J]. Physica A: Statistical Mechanics and Its Applications, 2010, 389(1): 187-197.

[17] Zhang Zhongyuan. Community structure detection in complex networks with partial background information[J]. Europhysics Letters, 2013, 101(4): DOI: 10. 1209/0295-5075/101/48005.

[18] Zhang Zhongyuan, Sun Kaidi, Wang Siqi. Enhanced community structure detection in complex networks with partial background information [J]. Scientific Reports, 2013, 3(11): DOI: 10. 1038/srep3241.

[19] Karthik S, Aggarwal C C, Jaideep S, *et al*. Community detection with prior knowledge[C]//Proc of the SIAM International Conference on Data Mining. Philadelphia, PA: SIAM, 2013.

[20] Cheng Jianjun, Leng Mingwei, Li Longjie, *et al*. Active semi-supervised community detection based on must-link and cannot-link constraints [J]. Plos One, 2014, 9(10): e110088.

[21] Liu Dong, Bai Hongyu, Li Huijia, *et al*. Semi-supervised community detection using label propagation[J]. International Journal of Modern Physics B, 2014, 28(29): 1450208.

[22] Kingma D, Rezende D, Mohamed S, *et al*. Semi-supervised learning with deep generative models[C]//Advances in Neural Information Processing Systems. 2014.

[23] 姜雅文. 复杂网络社区发现若干问题研究 [D]. 北京:北京交通大学,2014.

[24] Chai Bianfang, Yu Jian, Jia Caiyan, *et al*. Combining a popularity-productivity stochastic block model with a discriminative-content model for general structure detection [J]. Physical Review E, 2013, 88(1): 012807.

[25] 柴变芳,于剑,贾彩燕,等. 一种基于随机块模型的快速广义社区发现算法[J]. 软件学报,2013,24(11):2699-2709.

[26] 柴变芳,贾彩燕,于剑. 基于概率模型的大规模网络结构发现方法[J]. 软件学报,2014,25(12):2753-2766.

[27] 柴变芳,赵晓鹏,贾彩燕,等. 内容网络广义社区发现有效算法[J]. 计算机科学与探索,2014,8(9):1076-1084.

[28] 柴变芳,赵晓鹏,贾彩燕,等. 大规模网络的三角形模体社区发现模型[J]. 南京大学学报:自然科学版,2014,50(4):466-473.

[29] Guerra L, Bielza C, Robles V. Semi-supervised projected model-based clustering[J]. Data Mining and Knowledge Discovery, 2014, 28(4): 882-917.

(上接第 1630 页)

3　结束语

推荐多样性正日益成为评价推荐系统性能的重要指标。针对现有推荐算法缺少对推荐总体多样性考虑现象,提出一种基于二分图网络的推荐多样性增强算法。通过对现有推荐算法进行二次优化,与现有算法进行实验对比,结果表明本文算法能有效提高系统总体多样性且保持较高推荐准确率。本文算法的不足之处在于推荐增广路的选取上缺少对算法准确率的考虑,没有考虑推荐的个体多样性。如何同时提高系统个体多样性与总体多样性,将是下一步的研究方向。

参考文献:

[1] Hill W, Stead L, Rosenstein M, *et al*. Recommending and evaluating choices in a virtual community of use[C]//Proc of ACM SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 1995: 194-201.

[2] Resnick P, Varian H R. Recommender systems[J]. Communications of the ACM, 1997, 40(3): 56-58.

[3] Kim J E, Kim H G. Music recommendation method with respect to message service: US, US8410347[P]. 2013.

[4] 陈克寒,韩盼盼,吴健. 基于用户聚类的异构社交网络推荐算法[J]. 计算机学报,2013,36(2):349-359.

[5] 安维,刘启华,张李义. 个性化推荐系统的多样性研究进展[J]. 图书情报工作,2013,57(20):127-135.

[6] Ziegler C N, Mcnee S M, Konstan J A, *et al*. Improving recommendation lists through topic diversification [M]//A History of the Comstock Silver Lode & Mines. [S. l.]: Promontory Press, 1974: 22-32.

[7] Niemann K, Wolpers M. A new collaborative filtering approach for increasing the aggregate diversity of recommender systems[C]//Proc of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2013: 955-963.

[8] Adomavicius G, Kwon Y O. Maximizing aggregate recommendation diversity: a graph-theoretic approach[C]//Proc of the 1st International Workshop on Novelty and Diversity in Recommender Systems. 2011: 3-10.

[9] Adomavicius G, Kwon Y O. Improving aggregate recommendation diversity using ranking-based techniques[J]. IEEE Trans on Knowledge & Data Engineering, 2012, 24(5): 896-911.

[10] 刘慧婷,岳可诚. 可提高多样性的基于推荐期望的 top-*N* 推荐方法[J]. 计算机科学,2014,41(7):270-274.