

# 一种面向主题耦合的影响力最大化算法

吕文渊 周丽华 廖仁建

(云南大学信息学院 昆明 650000)

**摘要** 网络逐渐成为了人与人之间的主要社交工具,在网络中挖掘最有影响力的用户成为了非常值得关注的问题。在传统影响力最大化算法的基础上提出了一种面向主题耦合的影响力最大化算法,该算法首先分析网络中不同主题之间的耦合相似性,在综合考虑主题之间耦合相似性与用户对不同主题偏好的基础上扩展独立级联模型,并使用经典的贪心算法挖掘最具有影响力的用户。与不考虑主题耦合的影响力最大化算法相比,所提算法考虑了传播主题之间的耦合相似性,并且能够与用户偏好进行更为有效地结合。最后,实验表明,相比于经典的影响力最大化算法,该算法能够更为有效地挖掘在特定主题下最具有影响力的种子节点。

**关键词** 社会网络,影响力最大化,耦合相似度,主题

**中图分类号** TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.12.005

## Coupled Topic-oriented Influence Maximization Algorithm

LV Wen-yuan ZHOU Li-hua LIAO Ren-jian

(School of Information, Yunnan University, Kunming 650000, China)

**Abstract** As networks are main tool for communication in the modern society, digging the most influential network users has become a hot issue. This study proposed a coupled topic-oriented influence maximization algorithm. It analyzes couplings among topics and extends the independent cascade model by considering couple similarity among topics and users' preference on different topics. The classical greedy algorithm is used to dig the most influential users on the extended model. Compared with the influence maximization algorithm without the coupled topic, the proposed algorithm digs out more rational users who can affect more users in networks.

**Keywords** Social network, Influence maximization, Similarity of coupling, Topic

## 1 引言

市场经济是当今时代人们最关心的话题之一,如何在市场中减少营销成本,扩大收益,进而取得利益的最大化是一个永恒不变的追求。随着时代的进步,互联网思维已经紧紧地与市场结合在一起。近几年较为流行的“病毒营销”<sup>[1,2]</sup>和“口碑效应”<sup>[3,4]</sup>恰恰利用了人与人之间的互相影响。在互联网上利用多种多样的社交网络进行病毒式营销,往往会以极小的成本获得巨大影响,从而使产品销量大大提高,影响力最大化算法就是解决“病毒营销”问题的关键所在。

影响力最大化问题是“病毒营销”中的关键问题。谁是营销过程中最具有“病毒”传播能力的用户呢?找到这些最具传播能力的用户就如同在谷堆中找到最优良的种子一样既复杂又关键,只有找到了最优良的“种子”才能使影响力最大化问题取得最优解。因此,如何发现影响力最大化问题的最优“种子”成为一个十分重要的研究问题。在过去的研究中已经有学者提出面向主题的影响力最大化问题<sup>[9,12]</sup>,但是并没有考虑主题之间的耦合相似性,因此本文提出一种面向主题耦合

的影响力最大化算法 GACT(Greedy Algorithm Based on the Couped Topics)来解决该问题。

本文首先针对具体的问题选择主题,为传播过程中信息的不同主题建立描述属性的集合,并且通过这个集合分析不同主题之间的耦合关系,即不同主体(信息资源)建立在同一客体(信息资源)基础上所形成的相互之间的潜在联系,进而通过这种潜在的联系衡量主题之间的耦合相似性,然后在独立级联模型的基础上修改节点之间的激活概率,以用户对耦合主题的偏好重新定义激活概率,并以此建立一个面向主题耦合的影响力传播模型。最后通过实验将 GACT 算法与解决影响力最大问题的贪心算法进行比较。

## 2 相关工作

本节将对影响力最大化问题的研究背景和传播模型做相关介绍。

### 2.1 影响力最大化问题

影响力最大化问题最初由 Richardson 和 Domingos<sup>[5]</sup>引入社会网络领域,他们给出了社会网络上影响力最大化问题

到稿日期:2016-10-11 返修日期:2016-11-12 本文受国家自然科学基金项目(61262069,61472346,61762090),云南省自然科学基金项目(2015FB114,2016FA026),云南省创新团队,云南省高校科技创新团队(IRTSTYN),云南大学创新团队发展计划(XT412011)资助。

吕文渊(1988-),硕士,主要研究方向为社会网络分析;周丽华(1968-),博士,教授,主要研究方向为数据挖掘、社会网络分析,E-mail:lhzhou@ynu.edu.cn(通信作者);廖仁建(1991-),硕士生,主要研究方向为数据挖掘。

的详细定义和评价指标。Kempe 等<sup>[6]</sup>在影响力最大化问题引入社会网络后对该问题进行了详细的研究,提出使用贪心算法求解影响力最大化问题,并提出了两种经典的传播模型:独立级联模型 IC(Independent Cascade)<sup>[7]</sup>和线性阈值模型 LT(Linear Threshold)<sup>[8]</sup>。

在此之后,Tang 等人<sup>[9]</sup>提出了基于主题的社会网络影响力最大化问题。他们在网络中找到每个节点的主题偏好,并且划分出每个主题的网络子图,在这个子图中找到节点对于特定主题的影响力权值并在这样的模型下研究影响力最大化问题,通过实验发现不同主题的信息在网络上传播的效果也是不同的。但是 Tang 等人并没有研究网络上传播的信息所属的不同主题之间的相互关系。虽然主题之间可能因时间或空间的距离没有直接联系,但是如果主题与主题之间存在着某种耦合的关系,则主题作为主体通过相同的客体的一种联系纽带会在主题之间建立一种间接联系,这种间接联系可以体现两个事物之间共同的特征交集。比如,若两个导演共同聘请过同一个演员,则就形成了一种耦合关系,代表着两个导演之间的相似性。本文研究如何通过主题之间的耦合关系促进影响力最大化问题的求解。

### 2.2 影响力传播模型

独立级联模型是应用得最为广泛的传播模型,同时也是一种以发送者为中心的概率模型。独立级联模型的设计基于概率论中的相互粒子系统,是动态级联模型概念中最简化的一个模型。独立级联模型按照随机的过程展开。

(1)当节点  $u$  在  $t-1$  时刻被激活时,在  $t$  时刻节点  $u$  有唯一的一次机会激活它的邻居节点  $v$ ,激活概率为  $P_w$ 。 $P_w$  值越高,节点  $u$  越有可能对节点  $v$  产生影响。若在  $t$  时刻节点  $v$  有多个邻居节点对它进行激活,则  $v$  的邻居节点以任意顺序激活节点  $v$ 。

(2)完成节点  $u$  激活节点  $v$  的过程后,无论  $v$  是否被激活,在时间  $t$  以后的时刻  $u$  都不能再次对其邻居节点  $v$  进行激活。

当整个社交网络中再也没有新的节点被激活时,整个激活过程结束。

线性阈值模型是以接收者为核心的离散模型。在该模型中每个用户节点  $u$  都从  $(0,1)$  中选择一个被激活的阈值  $\theta_u$ ,节点  $u$  被它的邻居节点  $v \in N(u)$  以权重  $b_{uv}$  影响( $b_{uv} < 1$ )。在信息传播的过程中,给定节点  $v \in N(u)$  的权重值和  $u$  的激活阈值,传播过程为以下离散步骤:

(1)对于在  $t-1$  时刻已经成为激活状态的节点,其在  $t$  时刻保持激活状态不变。

(2)对于在  $t-1$  时刻不属于激活状态的节点  $u$ ,若在  $t$  时刻  $u$  节点所有的邻居节点的权重值之和大于  $u$  的激活阈值  $\theta_u$ ,则节点  $u$  被激活。

(3)节点  $u$  被激活以后,在下一个时刻将会影响他的邻居节点。当整个社交网络中再也没有新的节点被激活时,传播过程结束。

## 3 面向主题耦合的影响力最大化算法

### 3.1 面向主题耦合的影响力最大化问题描述

社会网络图  $G(V, E)$  中,  $V$  代表社会网络中用户节点的

集合,  $E$  代表社会网络中用户之间联系的边的集合,传播过程中种子节点集合为  $Z(Z \subseteq V)$ ,  $RS(Z)$  为种子节点集合最终在网络中取得的影响。网络中包含  $n$  个主题,  $t_i (i \leq n)$  代表一个主题,主题与主题之间的耦合程度记作  $\delta(t_i, t_j)$ , 每一个用户  $u$  都有一个针对特定主题  $t_i$  的主题偏好,记作  $M_{ut_i}$ 。 $M_{ut_i}$  的值越大说明节点  $u$  对主题  $t_i$  越感兴趣。在综合了主题耦合因素之后节点对于主题  $t_i$  感兴趣的程度定义为  $C_{ut_i}$ ,  $C_{ut_i}$  的值越大说明用户  $u$  在综合了主题耦合的情况下对主题  $t_i$  的兴趣度越高。 $m$  与  $n$  两个用户之间激活的概率表示为与两个用户对主题感兴趣程度的值所相关的函数,即  $P_{mn} = q(C_{mt_i}, C_{nt_i})$ 。

社会网络中信息传播的过程使节点分成两个状态:激活状态与未激活状态。激活状态是指节点已经接受了新的信息,未激活状态是指节点还未接受新的信息。在社会网络图  $G(V, E)$  中,  $v$  为图中已激活的节点,  $N(v)$  表示与节点  $v$  直接相连的邻居节点的集合。若节点  $u \in N(v)$ ,使得节点  $u$  从未激活状态转化到激活状态的过程称为节点  $v$  对  $u$  的激活,在传播模型中节点一旦被激活就一直保持激活状态不再发生变化。

因此,我们将面向主题耦合的影响力最大化问题定义为:给定一个社会网络图  $G$  和种子节点数目  $k$ ,在图中每个用户节点都有耦合偏好  $C_{ut_i}$ ,图中每条边都会对应一个激活概率  $P_{mn}$ ,  $P_{mn}$  是与用户  $m$  和  $n$  的主题偏好  $C_{mt_i}$  和  $C_{nt_i}$  相关的函数。面向主题耦合的影响力最大化问题就是要找到一个包含着  $k$  个种子节点的初始集合  $Z$ ,使得  $RS(Z)$  达到最大,影响节点的数目记作  $\sigma(Z) = RS(Z)$ 。

### 3.2 主题耦合分析

用一个信息表的形式将需要进行耦合相似分析的数据组织在一起,可以表述为  $S = \langle U, T, V, f \rangle$ 。 $U = \{u_1, \dots, u_m\}$  表示由数据对象组成的非空有限集,  $A = \{t_1, \dots, t_n\}$  表示由属性组成的非空有限集,  $V = \bigcup_{j=1}^n v_j$  是由特征值的集合  $v_j$  组成的并集,  $v_j$  是属性  $a_j (1 \leq j \leq n)$  对应的特征值组成的集合,映射  $f_j$  的集合  $f = \bigcup_{j=1}^n f_j$ ,其中映射  $f_j: U \rightarrow V_j$  是将数据对象映射到属性  $a_j$  的特征值  $v_j$  上。

表 1 是由 6 个数据对象  $\{u_1, \dots, u_6\}$  和 3 个属性  $\{a_1, a_2, a_3\}$  构成的项目信息表。从表 1 可以看出,对象  $u_1$  在属性  $a_2$  上对应的特征值为  $f_2(u_1) = b_1$ ,属性  $a_2$  特征值的集合为  $V_2 = \{b_1, b_2, b_3\}$ 。对象  $u_x$  在属性  $a_j$  下的特征值可以表示为  $V_j^x$ 。

表 1 项目信息表

属性	$a_1$	$a_2$	$a_3$
$u_1$	$t_1$	$b_1$	$c_1$
$u_2$	$t_2$	$b_1$	$c_1$
$u_3$	$t_2$	$b_2$	$c_2$
$u_4$	$t_3$	$b_3$	$c_2$
$u_5$	$t_4$	$b_3$	$c_3$
$u_6$	$t_4$	$b_2$	$c_3$

属性值与对象之间相互转换的公式  $SIF_1$  (Set Information Functions)<sup>[10]</sup> 为:

$$F_j(U') = \{f_j(u_x) | u_x \in U'\} \tag{1}$$

$$G_j(V') = \{u_i | f_j(u_i) \in V_j'\} \tag{2}$$

其中,  $1 \leq j \leq n, 1 \leq i \leq m, U' \subseteq U$  且  $V_j' \subseteq V_j$ 。

这一对 SIF<sub>0</sub> 公式用集合的形式描述了对象与特征值之间的关系,比如在属性  $a_2$  上对象  $u_1, u_2$  和  $u_3$  所对应的特征值可以表示为  $F_2(\{u_1, u_2, u_3\}) = \{b_1, b_2\}$ , 特征值  $b_1, b_2$  所对应的对象可以表示为  $G_2(\{b_1, b_2\}) = \{u_1, u_2, u_3, u_6\}$ 。

特征值之间相互对应的公式 IIF(Inter-information Function)<sup>[10]</sup> 可以表示为:

$$\psi_{j \rightarrow k}: \psi_{j \rightarrow k}(v_j) = F_k(G_j(\{v_j\})) \quad (3)$$

IIF 公式由函数  $F_k$  和  $G_j$  组成,映射  $\psi$  的下标  $j \rightarrow k$  表示从属性  $a_j$  映射到属性  $a_k$ , 直观地说,  $\psi_{j \rightarrow k}(v_j)$  计算了属性  $a_j$  下的特征值  $v_j$  在属性  $a_k$  下所对应的特征值的集合。例如  $\psi_{2 \rightarrow 1}(B_2) = \{t_1, t_2\}$ , 表明属性  $a_2$  下的特征值  $b_1$  与属性  $a_1$  下的特征值  $t_1$  和  $t_2$  相对于对象  $u_1$  和  $u_2$  是相关联的。

相互关联的特征值的条件概率公式 ICP(Information Conditional Probably)<sup>[10]</sup> 可以定义为:

$$P_{k|j}(\{V_k' | v_j\}) = \frac{|G_k(V_k') \cap G_j(\{v_j\})|}{|G_j(\{v_j\})|} \quad (4)$$

其中,  $V_k'$  表示与特征值  $v_j$  相关联的特征值的集合, ICP 公式表示特征值  $v_j$  所对应的对象与特征值的集合  $V_k'$  所对应的对象发生重合的部分在特征值  $v_j$  所对应的对象中所占的比率, 例如  $P_{1|2}(\{t_1 | b_1\}) = 0.5$ 。

### 3.3 耦合相似度计算

(1) 内耦合频率分布相似度的计算<sup>[10]</sup>

$$\delta_j^{i_a}(v_j^x | v_j^y) = \frac{|G_j(\{v_j^x\})| \cdot |G_j(\{v_j^y\})|}{|G_j(\{v_j^x\})| + |G_j(\{v_j^y\})| + |G_j(\{v_j^x\})| \cdot |G_j(\{v_j^y\})|} \quad (5)$$

内耦合相似度的计算只考虑了一个属性内部特征值的频率分布相似度, 并没有考虑属性之间的依赖关系对特征值之间相似度的影响, 比如仅在属性  $a_2$  中考虑属性值  $b_1$  和  $b_2$  之间因耦合而产生的相似度  $\delta_2^{i_a}(b_1 | b_2) = 0.5$ 。内耦合相似度的计算存在一个问题:  $v_j^x$  与  $v_j^y$  出现的频率相同时  $\delta_j^{i_a}(v_j^x | v_j^y) = \delta_j^{i_a}(v_j^y | v_j^x)$ 。但是这个问题会通过间耦合导致的相似性解决, 因为在间耦合相似度的计算过程中,  $v_j^x$  和  $v_j^y$  的相似度会与  $v_j^x$  和  $v_j^y$  的相似度拉开很大的距离, 我们最终对属性值相似度的计算是综合了内耦合频率分布相似度和间耦合特征值依赖聚合相似度后得到的结果。

(2) 间耦合特征值依赖聚合相似度的计算<sup>[10]</sup>

属性  $a_j$  中的属性值  $v_j^x$  和  $v_j^y$  关于属性  $a_k$  间耦合的相似度计算公式为:

$$\delta_j^{i_a}(v_j^x, v_j^y, \{v_k\}_{k \neq j}) = \sum_{k=1, k \neq j}^n \alpha_k \delta_{j|k}(v_j^x, v_j^y, V_k) \quad (6)$$

(3) 耦合相似度的计算<sup>[10]</sup>

式(5)与式(6)分别计算了特征值在属性内部因内耦合而产生的频率分布相似度以及间耦合导致的特征值依赖聚合相似度。因此, 式(7)综合了内耦合和间耦合的耦合相似度后进行计算:

$$\delta_j^A(v_j^x, v_j^y, \{v_k\}_{k=1}^n) = \delta_j^{i_a}(v_j^x, v_j^y) \cdot \delta_j^{i_e}(v_j^x, v_j^y, \{v_k\}_{k \neq j}) \quad (7)$$

### 3.4 用户偏好计算

在社会网络中, 每个人都有自己感兴趣的主题。通过对

用户发布的信息中的内容进行衡量, 可以判断其对不同主题的喜好程度; 再通过协同过滤的潜在语义索引方法就可以找出用户的兴趣偏好。

首先, 获取不同用户在不同主题之下发布的信息记录, 以此构成一个评分矩阵, 共有  $m$  个用户  $n$  个主题组成一个 user-topic 矩阵  $A$ , 大小为  $m$  行  $n$  列。

$$A_{m \times n} = [a_{ij}] = (\text{topic}_1, \text{topic}_2, \dots, \text{topic}_n) \times (\text{user}_1, \text{user}_2, \dots, \text{user}_m)^T$$

其中,  $a_{ij}$  表示用户  $i$  在主题  $j$  下发表的信息数量。在这个矩阵中有的用户会比较活跃, 其发表的信息数目会非常大, 而有的用户发表的信息数目比较少, 为了避免数目上存在较大的差异性, 使用五分制的方法将这个矩阵转化为一个较为简单的评分矩阵, 转化规则如表 2 所列。

表 2 转化规则

主题下发表的信息数目/发表信息总数	评分值
0	0
(0, 0.03]	1
(0.03, 0.05]	2
(0.06, 0.1]	3
(0.1, 0.3]	4
(0.3, 1]	5

user-topic 矩阵转化为五分制的评分矩阵之后,  $a_{ij} \in \{0, 1, 2, 3, 4, 5\}$ 。随后对 user-topic 矩阵进行奇异值分解以替代原始的 user-topic 矩阵。

假设要在主题集合  $T$  中选取  $t_i (t_i \in T)$  主题的信息进行传播。在面向主题的影响力最大化算法中只考虑了用户对于主题  $t_i$  的感兴趣程度, 并没有考虑到用户感兴趣的其他主题与主题  $t_i$  的耦合度。因此, 本文在面向主题耦合的影响力最大化算法中, 除了考虑用户对于主题  $t_i$  的偏好, 还考虑了其他主题  $t_j (t_j \in T, i \neq j)$  与主题  $t_i$  之间的耦合相似度, 即:

$$C_{w_i} = M_{w_i} + \frac{\sum_{t_j \in T, i \neq j} M_{w_j} \delta^A(t_i, t_j)}{|T - \{t_i\}|} \quad (8)$$

其中,  $\delta^A(t_i, t_j)$  相当于用  $t_i$  和  $t_j$  替代了式(7)中  $v_j^x$  和  $v_j^y$  之后计算得出的结果,  $M_{w_i}$  是节点用户  $w$  对传播的特定主题  $t_i$  的固有偏好, 式(8)在用户对主题  $t_i$  固有偏好的基础之上附加了用户的耦合偏好。耦合偏好是用户对网络中除了  $t_i$  之外的所有主题的偏好与主题  $t_i$  耦合相似度的乘积的平均值。这样既考虑了对主题  $t_i$  本身感兴趣的用户群体, 也考虑了即使不对主题  $t_i$  感兴趣但是对于主题  $t_i$  相似的其他主题  $t_j$  感兴趣的用户群体, 使得用户偏好  $C_{w_i}$  有着更高的精度。

在经典的独立级联模型中, 通过参数来设定节点之间的激活概率。Wang 等人<sup>[11]</sup> 提出了一种基于联系频率的独立级联模型, 使模型中的激活概率与节点间联系的频繁程度相关。但是在实际的社交网络中, 激活概率不仅与联系的频繁程度相关, 还与用户的偏好相关; 用户  $v$  与用户  $w$  对某个传播的主题越感兴趣,  $v$  与  $w$  之间的激活概率就越高。基于这样的考虑, 本文重新定义节点之间的激活概率:

$$p_{vw} = p \cdot (C_{v t_i} \cdot C_{w t_i})^2 \quad (9)$$

其中,  $C_{v t_i}, C_{w t_i}$  分别通过式(8)计算得出。式(9)中  $p$  是设定的一个激活概率参数, 若  $p$  值过高, 则网络中所有的节点都

会被激活;过低则导致过程终止较快,无法达到预期效果。根据实验, $p$  值取 0.05 较为合适。当  $v$  与  $w$  之间不止一条边时,每条边的激活概率都为  $P_{vw}$ 。我们将引入了基于主题耦合的用户偏好的独立级联模型定义为扩展的独立级联模型(Extended Independent Cascade Model, ECIM)。

### 3.5 面向主题耦合的影响力最大化算法

#### 3.5.1 影响力计算方法

在社交网络  $G(V, E)$  中任意节点发生一次激活行为的结果图为  $G_r$ , 若图中节点  $v, w \in G_r, w \neq v$ , 且存在  $v$  到达  $w$  的路径, 则称  $w$  为  $v$  的可达节点。 $R(v, G_r)$  为节点  $v$  的可达节点集合。对于图  $G_r$  中节点集合的子集  $A$ , 其可达节点集合为  $R(A, G_r) = \bigcup_{v \in A} R(v, G_r)$ 。种子节点集合  $Z$  在社交网络最终的影响力计算公式为:

$$\sigma(Z) = \frac{1}{N} \sum_{r=1}^N |R(Z, G_r)| \quad (10)$$

其中,  $N$  为重复影响力传播过程的次数, 取其平均值计算得出最终的影响结果。种子集合  $Z$  基于传播模型进行影响力传播, 最终影响结果算法 TotalInf 的流程如算法 1 所示。

#### 算法 1 TotalInf

输入:  $G(V, E), Z$ , 模拟传播的次数  $N$

输出: 最终影响结果  $\sigma(Z)$

初始化: Build Models // 构建影响力传播模型

For  $i=1$  to  $N$  do

AnalogModels // 模拟影响力传播过程

$I(G_r) = |R(Z, G_r)| + I(G_{r-1})$  //  $N$  次传播累计的影响节点总计数目

End for

$\sigma(Z) = \frac{1}{N} I(G_r)$  // 计算  $N$  次随机过程之后的平均值

Return  $\sigma(Z)$

#### 3.5.2 GACT(Greedy Algorithm based on Couped Topics)

在扩展的独立级联模型之上, 我们使用贪心策略选取最具有影响力的  $k$  个用户。种子节点集合记为  $Z$ , 集合  $Z$  之外的图中节点标记为非活跃节点;  $RS(Z)$  为种子节点的影响结果,  $|RS(Z)|$  为影响的节点个数。

贪心算法的输入为社交网络  $G$ , 种子节点数目  $k$  以及面向主题耦合的用户偏好  $C_{ut}$ , 输出为一个大小为  $k$  的集合  $Z$ , 使得集合  $Z$  中的节点影响范围达到最大。算法 2 给出了整个贪心算法的详细流程。

#### 算法 2 GACT

输入: 社交网络  $G$ , 种子集合大小  $k$ , 面向主题耦合的用户偏好  $C_{ut}$

输出: 大小为  $k$  的种子集合  $Z$

初始化:  $Z = \emptyset, R = 3000$  // 初始化一个有 3000 个节点的社会网络

For every edge  $(v, w)$  do

$p_{vw} = p \cdot (C_{vt} \cdot C_{wt})^2$  // 为每条边赋予新的激活概率

End for

For  $i=1$  to  $k$  do

For  $j=1$  to  $R-i$  do

$A \leftarrow Z_{i-1} \cup \{a_j\}$  // 在集合  $Z$  中尝试加入新的节点

$\sigma(A) \leftarrow \text{TotalInf}(R, A)$  // 在  $R$  个节点的社交网络中计算影响结果

结果

If  $\sigma(A)$  is Max // 找到使得种子集合影响力最大的节点

End for

$Z \leftarrow A$

End for

Return  $Z$  // 返回种子节点集合  $Z$

在算法 2 中, 首先计算基于主题耦合的用户偏好  $C_{ut}$ , 以此为基础计算每条边的激活概率  $p_{vw}$ , 然后, 在新的独立级联模型上使用贪心算法挖掘最具有影响力的  $k$  个节点, 在每一轮  $i$  找到一个新的节点加入集合  $Z$  使得集合  $Z$  在加入新节点后的边际影响值达到最大; 最后输出的集合  $Z$  即为需要挖掘的种子节点集合。

## 4 实验分析

### 4.1 实验数据

“豆瓣网”是一个以评论为中心的著名社交网络, 其内容包含图书、电影、音乐、旅行等年轻人较为感兴趣的版块, 其中豆瓣电影成为国内主流的电影评分网站, 有着较高的权威度。因此本文选取豆瓣网上的数据作为实验数据集来构建一个电影交流网, 以更好地模拟现实中的社交网络。

本文首先使用爬虫程序从豆瓣电影板块下“中国”的类目中下载所有的电影信息(5177 部), 爬虫程序下载的主要信息包含: 电影名称、导演、主演、年代、类型、评分, 将这些信息以表的形式进行存储。在本文构建的电影交流网中, 导演成为网络传播中的一种主题, 下载的电影信息主要用于主题耦合的分析。

为了使实验数据更为有效, 依然采用爬虫程序从豆瓣网上的“云南大学”社区内下载用户数据集(2491 人)。选择同一社区用户的好处是用户之间单方面关注或者互相关注的比例较大, 即在抽象的网络图  $G(V, E)$  中边的数目较多, 能够更好地对比不同策略的影响力最大化算法。爬虫程序下载的主要信息有用户名、电影的评论、想看的电影、喜欢的电影(在豆瓣用户的信息中包含有这些内容)。在这些内容上使用潜在语义索引的方法分析用户对不同导演的偏好, 形成评分矩阵。

### 4.2 实验标准

在以往对影响力最大化问题的研究中, 算法的研究主要聚焦于时间复杂度, 而本文则更加关注算法的精确度。传统的影响力最大化算法在精确度上的衡量标准是种子集合  $Z$  的影响范围, 即:

$$IS(Z) = |IS(Z)| \quad (11)$$

$IS(Z)$  是在独立级联模型上最终激活的节点集合。在面向耦合主题的影响力最大化算法研究中, 不仅要考虑种子集合最终影响范围的广度, 还需要考虑种子集合最终影响范围的有效程度。例如在特定主题  $t_i$  的传播过程中, 激活一个对主题  $t_i$  感兴趣的人的实际效果远超过激活一个对主题  $t_i$  不感兴趣的人。因此我们在激活概率的定义的基础上重新定义一个新的评价标准  $ISCT(S)$ :

$$ISCT(Z) = \sum_{u \in RS(Z)} u C_{ut}^z \quad (12)$$

$ISCT(Z)$  是在传播特定主题  $t_i$  时的影响范围, 这个衡量标准是考虑用户面向主题耦合的偏好之后在指定主题  $t_i$  上产生的影响力。其中  $RS(Z)$  是在独立级联模型上最终激活

的节点集合。实验中将分别采用  $IS(Z)$  和  $ISCT(Z)$  两种标准进行比较,以观察各种策略的算法最终在影响力最大化问题上产生的效果如何。

### 4.3 实验结果

本文在实验中选取近期较为流行的电影《澳门风云3》作为实验中进行宣传的电影,《澳门风云3》的导演为“刘伟强”,在电影社交网中“刘伟强”就成为传播的特定主题,本文在该特定主题下对比不同算法影响力最大化的效果。

首先在模拟的电影社交网中对比贪心算法 GA(Greedy Algorithm)、面向主题的影响力最大化算法 GAT(Greedy Algorithm based on Topic)、RA 影响力最大化问题中的随机选择算法 RA(Random Algorithm)与面向主题耦合的影响力最大化算法 GACT 产生的种子集合  $Z$  的  $IS(Z)$  值。实验分别控制种子集合的大小为 50, 100, 150。实验结果如图 1 所示。

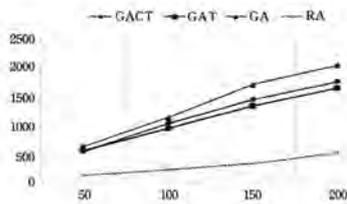


图 1  $IS(Z)$  效果对比

从图 1 中可以看出,由于计算  $IS(S)$  的值时未考虑最终影响的用户节点对主题的偏好,因此经典的贪心算法在影响力最大化问题上的表现最为优秀,GACT 算法在 GAT 算法的基础上考虑了主题耦合相似度,最终结果比 GAT 算法高出约 10%;表现最差的算法为随机算法,其结果远远劣于其他 3 种算法。基于这样的考虑,我们使用  $ISCT(Z)$  标准再次对几种算法进行比较,实验结果如图 2 所示。

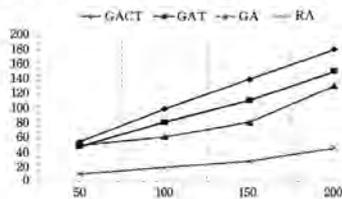


图 2  $ISCT(Z)$  效果对比

从图 2 中可以看出,在  $ISCT(S)$  的标准下,表现最为优秀的算法是 GACT 算法,虽然实验结果中考虑了用户对主题的偏好,但是加入了主题耦合相似度的 GACT 算法的效果依然优于不考虑主题耦合相似度的 GAT 算法。在此标准下,GACT 算法的效果高出经典的贪心算法约 30%,因为通过经典的贪心算法可能会激活数量更多的用户节点,但是它激活的节点中的一部分并不关注传播的特定主题,在传播结果中不具有影响力。

**结束语** 随着网络的普及,社交网络成为了人与人社交的主要途径,信息在人群中更加方便快捷地传播,使得影响力最大化问题的研究具有更大的价值和意义,但同时其也面临

着严峻的挑战。本文通过分析网络中传播的信息具有的耦合性质以及信息在社交网络中传播的特性,提出了新的解决思路,主要工作总结如下:

首先,介绍影响力最大化问题在学术领域内的研究现状以及相关理论,例如影响力最大化问题在社交网络中的定义,两种主流的传播模型:线性阈值模型、独立级联模型。

然后,主要介绍了数据之间的耦合关系,详细说明了如何计算数据之间的耦合相似度,基于该理论计算用户对网络中传播的某一特定主题的偏好程度。在此基础上改变传播模型中的激活概率,使之与用户对主题的偏好相关,形成一个扩展的独立级联模型。

最后,对面向主题耦合的影响力最大化问题进行定义,提出解决该问题的 GACT 算法,在扩展的独立级联模型上用该算法解决影响力最大化问题,并通过实验对比 GACT 算法与其他影响力最大化算法。

### 参考文献

- [1] BASS F. A new product growth model for consumer durables [J]. *Management Science*, 1969, 15(1): 215-227.
- [2] MAHAJAN V, MULLER E, BASS F. New product diffusion models in marketing: a review and directions for research [J]. *Journal of Marketing*, 1990, 54(1): 1-26.
- [3] BROWN J, REINEGEN P. Social ties and word-of-mouth referral behavior [J]. *Journal of Consumer Research*, 1987, 14(3): 350-362.
- [4] MA H, YANG H, LYU M R. Mining social networks using heat diffusion processes for marketing candidates selection [C] // *CIKM*. 2008: 233-242.
- [5] RICHARDSON M, DOMINGOS P. Mining Knowledge-Sharing Sites for Viral Marketing [C] // *Eighth Intl. Conf. on Knowledge Discovery and Data Mining*. 2002.
- [6] KEMPE D, KLEINBERG J, TARDOS E. Influential nodes in a diffusion model for social networks [J]. *Languages and Programming*, 2005, 32: 34-40.
- [7] MACY M. Chains of Cooperation: Threshold Effects in Collective Action [J]. *American Sociological Review*, 1991, 56: 78-83.
- [8] GOLDENBERG J, LIBAI B, MULLER E. Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth [J]. *Marketing Letters*, 2001, 12: 44-50.
- [9] TANG J, SUN J, WANG C. *Social influence analysis in large-scale networks* [M]. London: Cambridge University Press, 2009.
- [10] WANG C, CAO L. Coupled Attribute Similarity Learning on Categorical Data [J]. *IEEE Transactions on Neural Networks & Learning Systems*, 2015, 26(4): 781-797.
- [11] CHEN W, WANG Y J, YANG S Y. Efficient influence maximization in social networks [C] // *ACM SIGKDD*. 2009: 199-208.
- [12] 甘紫文. 学术网络主题影响力最大化研究 [D]. 杭州: 浙江大学, 2015.