主题关注模型下的影响最大化算法研究

郭景峰[1,2],范超智[1,2],陈 晓[1,2,3]

[1](燕山大学 信息科学与工程学院,河北 秦皇岛 066004)
[2](河北省虚拟技术与系统集成重点实验室,河北 秦皇岛 066004)
[3](华北理工大学 迁安学院,河北 唐山 064400)
E-mail:jfguo@ ysu. edu. cn

摘 要:目前,社会网络大多以社交关系为基础进行信息传播,影响最大化是信息传播领域挖掘有影响力的顶点集的热点. 随着大型社交网络的兴起,将主题偏好作为网络中实体的属性进行研究的影响最大化研究越来越多,较为缺少主题关注模型(融合社交关系和主题关注关系的新型社交网络模型)上的影响最大化研究. 针对这种情况,本文在此模型基础上,首先,在集对联系度基础上,结合随机游走计算各步内顶点主题偏好度,得到候选种子集;其次,在候选种子集上,基于贪心策略挖掘有影响力顶点;最后,在豆瓣数据集上,实现算法 TA_CELF,L_GAUP 和 CELF,从 ISST,ISRT,ISRNT 三个指标评价实验结果,实验结果表明,基于主题关注模型下进行的算法 TA_CELF 影响范围有较好的表现.

关 键 词:主题偏好;主题关注模型;信息传播;影响最大化
中图分类号:TP311 **文献标识码**:A **文 章 编 号**:1000-1220(2017)09-2113-06

Influence Maximization Based on Topic-attention Model

GUO Jing-feng[1,2],FAN Chao-zhi[1,2],CHEN Xiao[1,2,3]

[1](College of Information Science and Engineering,Yanshan University,Qinhuangdao 066004,China)
[2](Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province,Qinhuangdao 066004,China)
[3](School of North China University of Science and Technology,Hebei 064400,China)

Abstract:At present,most of the social network is based on social relations for information dissemination,influence maximization plays an important role in information diffusion in mining influential vertexes. With the rise of large social networks,research on the topic preference as the attributes of the entities in the network is studied more and more,lacking of influence maximization research on topic-attention model (a new model of social network fusion of social relationships and the topic attention relationship). In view of this situation,based on this model,first,combine the set pair connection degree and random walk to calculate the preference degree of the vertexes in each step,and get alternative seed set;then,in the alternative seed set,based on the greedy strategy to find the top-k vertexes;in the end,we conduct algorithm TA_CELF,L_GAUP and CELF on douban dataset,and evaluate experiment results from three metrics ISST,ISRT and ISRNT. Experiment results show that algorithm TA_CELF has the good performance on influence scope.

Key words:topic preference;topic-attention model;information diffusion;influence maximization

1 引 言

近年来,国内外受到移动社交网络传播而引发的重大社会事件倍增,例如,天津港爆炸事件、南海仲裁案等,给国家的公共安全和社会稳定带来了较大压力,这些庞大的数据量以及隐藏的信息为研究带来了新的机遇和挑战. 因此,信息传播及影响最大化问题成为当前的研究热点之一. 其中,影响最大化最早应用于“病毒营销”中,通过寻找多个有影响力的促销对象,从而达到最好的促销效果. 随后,在挖掘最有影响力用户、广告投放和产品推荐等不同领域都得到了广泛应用.

现有传统关系型社会网络的影响最大化问题研究的出发点是,在整个网络中寻找 k 个种子顶点,通过单一的用户实体间拓扑关系,挖掘种子顶点,目标是使整个网络中顶点最大化的被激活. 然而,现实生活中,消息或产品等传播的速度和范围,不仅和传播者与被传播者之间的自然关系有关,而且与消息或产品等的类型有重要的关系;如果被传播者对该类消息或产品与传播者有共同的感兴趣或认同,他会很快接收并继续传播下去,否则将不继续任何传播行为,在该网络中,用户与消息(或者产品)对传播都起到了至关重要的作用. 因此,可将消息或产品等作为网络的另一类关键实体——主题实体,将同时具有用户实体和主题实体的网络称之为主题关注网络,如图 1(a)所示,该网络不同于传统的单一用户实体关

收稿日期:2016-07-19 收修改稿日期:2016-10-14 基金项目:国家自然科学基金项目(61472340)资助;国家青年科学基金项目(61602401)资助;河北省自然科学基金面上项目(G2015203242)资助. 作者简介:**郭景峰**,男,1962 年生,博士,教授,博士生导师,CCF 会员,研究方向为数据库、社会网络;**范超智**,女,1992 年生,硕士研究生,研究方向为社会网络、数据挖掘;**陈 晓**,女,1983 年生,博士研究生,研究方向为数据库、社会网络

系网络,因此,更具有现实意义和研究价值.

当前以主题共享为目的的影响最大化研究相对较少. 本文在主题关注模型下进行影响最大化研究,问题的出发点是,在与待传播主题联系紧密的顶点集中寻找 k 个顶点,目标是使整个网络中顶点最大化的被激活的基础上,达到使网络中没有直接关注主题的顶点被激活的效果最佳. 例如,在图 1(a)所示的一个包含 7 个用户顶点和 3 个主题顶点的主题关注网络中,以主题 t_1 作为信息传播;首先,基于主题关注模型挖掘种子顶点,得到 3 个种子顶点,分别为 u_2,u_3 和 u_6,如图 1(b)所示;然后,将这些种子顶点应用于选定好的传播模型,进行最大化传播;最后,得到最大化的影响范围,7 个用户顶点均被激活,并且没有直接关注主题 t_1 的用户顶点 u_1,u_4,u_5 和 u_7 也均被激活,如图 1(c)所示. 本文在主题关注网络上提出一种新的影响最大化算法,侧重于更加精确挖掘种子顶点,使主题的传播范围达到最大化,本文主要贡献如下:

1)根据用户与主题之间的拓扑结构关系提出一种新的用户主题偏好度和主题影响力的度量方法,并将主题影响力作为种子顶点候选集的筛选阈值;

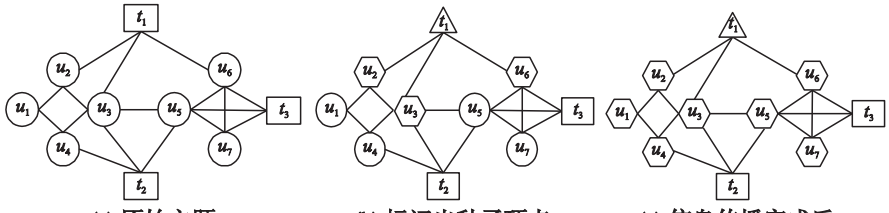

图 1 基于主题关注模型的影响最大化研究

Fig. 1 Influence maximization on the topic-attention network

2)在 IC 的主题关注模型(Topic Attention-independent Cascade model,TA_IC)基础上,基于贪婪算法提出了 TA_CELF 算法;

3)通过 TA_CELF 算法进行影响顶点的挖掘,获得了较为精确的小规模种子顶点集,同时,被传播的主题获得了较大影响范围.

2 相关工作

近年来影响最大化问题受到了许多学者的关注,现已存在许多关于社会网络影响最大化问题的研究[1-4],国内外的学者针对该问题的研究成果已有很多,具体研究现状如下.

Domingos[5] 等人最先考虑社会网中最具影响力的顶点选择问题. 2003 年,Kempe[6] 首次提出了影响最大化问题,证明了影响最大化问题在独立级联模型和线性阈值模型上都为 NP-hard 问题,并且设计出具有(1-1/e)近似比的贪心算法. 贪心算法虽然简单,但是由于在每次迭代选择种子顶点的过程中都需要进行大量的蒙特卡洛模拟来估计影响范围,导致贪心算法的效率较低. Leskovec[7] 提出了一个 CELF 优化的方法来改进贪心算法的效率,在保持了和简单贪心算法一致的近似比的前提下,比贪心算法快 700 倍.

存在许多研究通过设计高效的启发式算法来改进大规模网络上影响最大化问题的计算效率. 例如,Kimura[8] 为了更高效的估计影响范围提出了两个基于最短路径的影响级联模型. Goyal[9] 通过计算邻居顶点的简单路径估计影响范围. 启发式算法通常没有近似比保证,但具有很高的效率.

Guo[10] 提出了局部影响最大化问题,给定一个目标顶点 w,选择 k 个对 w 影响程度最大的节点,是具有近似比保证的局部影响最大化算法. 然而 Guo 的研究中并没有考虑针对目标顶点集合 T,如何选择最具影响力的 k 个顶点.

上述影响最大化和局部影响最大化的研究中都忽略了主题因素. Saito 和 Kimura[11,12] 所做的实验结果显示,信息所属的主题不同,在网络上的传播也是不同的. 网络中不同的用户具有不同的偏好,对同一个主题的偏好程度也不一样,那么该用户在针对不同主题的信息传播过程中所发挥的作用也都是不同的,正如 Saito 等人的实验中显示的结果一样. 然而,虽然 Saito 将主题与社会影响分析结合起来,得到针对不同主题会产生不同影响结果的结论,但并没有在此基础上做影响最大化问题的研究.

之后,学者们受到启发,纷纷投入到主题为核心的影响最大化研究中. Liu[13] 提出了一个概率模型用来学习主题分布以及主题感知的影响力度. Barbicri[14] 扩展了传统 IC 模型. 提出了主题感知的独立级联模型(Topic-aware Influence Cascade,TIC). Zhou 等人[15] 在研究中,首次提出基于不同话题来分别寻找 top-K 最有影响力顶点的算法. 随后,Guo 等人[16] 又对 Zhou 的算法存在的问题进行改进,提出了基于主题偏好的影响最大化算法 L_GAUP. 以上的成果都是适用于传统网络,缺乏主题关注网络下的影响最大化算法的研究. 本文以 Leskovec[7] 提出的 CELF 算法和 Guo 等人[16] 提出的 L_GAUP 算法进行对比,进行相关实验研究.

3 TA_IC 模型及影响最大化问题

3.1 TA_IC 模型

目前,许多学者都针对传统 IC 模型的不足给出了改进模型,主要介绍 2 种与本文相关的改进模型,如下.

1)Zhou 等人[11] 加入对某一主题 t_k 偏好的因素,修改了顶点对(u_i,u_j)之间影响概率,如式(1)所示.

$$p(u_i,u_j)=p\times F(C_{u_it_k},C_{u_jt_k}) \tag{1}$$

其中,$F(C_{u_it_k},C_{u_jt_k})=(C_{u_it_k}\times C_{u_jt_k})^2$,$C_{u_it_k}$和 $C_{u_jt_k}$分别表示用户 u_i 与 u_j 对主题 t_k 的偏好度. 但是,该方法存在一个问题,即,对于 $C_{u_it_k}$和 $C_{u_jt_k}$中任意一个为 0,那么 $p(u_i,u_j)$的结果就为 0,这是不符合逻辑的. 某一特定主题,当 p 足够大时,即

使 $C_{u_i t_k}$ 较小,顶点 u_i 也可以激活 u_j.

2)吕加国等人[12]通过引入调和因子 a,解决了上述问题,并提出了改进后的 E_IC 模型,a 用来表示该主题下用户 u_i 对用户 u_j 的总体影响力,如式(2)所示.

$$p(u_i,u_j,t_k)=a\times p(u_i,u_j)+(1-a)\times F(C_{u_i t_k},C_{u_j t_k}) \quad (2)$$

上述方法虽然考虑了主题对传播的影响,但仍不能直接将其应用于主题关注网络中. 因为主题关注网络涉及到 2 类实体,所以用户间的激活概率受到两部分的影响,一部分是结构上,即广义范围上的用户邻居;另一部分是内容上,即用户对主题的偏好度. 在此基础上提出新的传播模型,即 TA_IC 模型. 用户激活概率记为 $p(u_i,u_j,t_k)$,如式(3)所示.

$$p(u_i,u_j,t_k)=w_1\times Inf(u_i,u_j)+w_2\times F(Per(u_i,t_k),Per(u_j,t_k)) \quad (3)$$

其中,$Inf(u_i,u_j)$ 表示用户 u_i 对用户 u_j 结构上的影响力,如式(4)所示,即表示为 u_i 与 u_j 共同一级主题邻居与 u_i 与 u_j 所有一级主题邻居的比值;$F(Per(u_i,t_k),Per(u_j,t_k))$ 表示用户对于主题 t_k 的偏好度,如式(5)所示,即为 u_i 与 u_j 对主题 t_k 的偏好度的平均值;对于两部分,分别给予不同的权值 w_1 和 w_2,并且希望主题的影响因素较大,所以实验中设定 $w_2>w_1$.

$$Inf(u_i,u_j)=ST(u_i,u_j)_1/|NT(u_i)_1\cup NT(u_j)_1| \quad (4)$$

$$F(Per(u_i,t_k),Per(u_j,t_k))=(Per(u_i,t_k)+Per(u_j,t_k))/2 \quad (5)$$

3.2　TA_IC 模型的性质

基于 IC 模型下及其改进的模型挖掘种子顶点,是一个 NP-hard 问题. 因此,寻找最优解是十分困难的. 然而,可以通过一个贪心的爬山算法得到近似解,达到的精确度为(1-1/e). 由于 TA_IC 模型是对原始 IC 模型的边权值的改进,所以,Kempe 等人[3]给出的影响函数子模性证明过程同样适用于 TA_IC 模型:即:假设 $f(\cdot)$ 将任一有限子集 U 映射成为非负实数,当 f 满足边际收益递减性质时,称之为子模,即对于任意的集合 $S\subseteq T$ 满足下式:

$$f(S\cup\{v\})-f(S)\geqslant f(T\cup\{v\})-f(T) \quad (6)$$

3.3　基于 TA_IC 模型的影响最大化问题

不同于传统网络下的影响最大化问题,基于 TA_IC 模型的影响最大化问题是,对于给定的主题关注网络 G,用户 u_i 对于主题 t_k 的偏好矩阵 MT,给定种子集的规模 k,寻找 G 中规模为 k 的节点集 S^*,使得,在 TA_IC 模型下,以 S^* 为激活顶点集进行传播时,使之前没有直接关注主题的顶点被激活,并达到影响顶点集最大化.

在 TA_IC 模型下,计算得到偏好矩阵 MT,以 S 为激活顶点集在主题关注网络 G 中传播主题 t_k 时,传播过程结束后所得的最终影响集(G 中没有直接关注主题 t_k)的顶点个数记为 $N(G,MT,S,t_k)$. 因此,基于主题关注模型的影响最大化问题的形式化定义如下:

给定主题关注网络 G、偏好矩阵 MT、主题 t_k 和种子集的规模 k,在 G 中寻找一个规模为 k 的顶点集 S^*,使得,对于 G 中任意一个规模为 k 的顶点集 S,都有:$N(G,MT,S^*,t_k)\geqslant N(G,MT,S,t_k)$.

4　算法设计

下面,从主题偏好计算和基于主题关注模型的影响种子顶点挖掘 2 方面详细介绍 TA_CELF 算法的实现过程.

4.1　主题偏好的计算

对于某一主题,不同的用户对该主题会产生不同的关注行为(直接关注或间接关注),这样不同的行为往往表示用户对主题的不同偏好度. 如图 2 所示,用户 u_3 直接关注主题 T_1,然后用户 u_5 却通过用户 u_3 关注主题 T_1,由此可见,通过不同的路径,用户对主题的偏好程度是不同的,因此,在本文的偏好计算中,考虑不同步长情况下的偏好度计算.

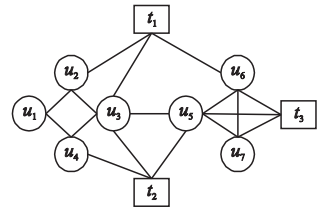

图 2　主题关注模型

Fig. 2　Topic-attention model

定义 3.1.(主题 t_k 的 m 步可达点)给定主题关注网络 $G=(V,E,W)$,$\forall t_k\in T$,如果存在用户顶点 u_q 与主题 t_k 直接相连接,则称 u_q 为 t_k 的一步可达点,如果用户顶点 u_q 与主题 t_k 间存在长为 $m(m>1)$ 的最短路径,则 u_q 称为 t_k 的 $m(m>1)$ 步可达点,表示为与主题直接相连接的用户顶点 u_i 的 $m-1$ 级用户邻居 u_q 的集合,主题 t_k 的 m 步可达点集记为 $RN(t_k)_m$,如式(7)所示.

$$RN_m=\begin{cases}\{u_q|(u_q,t_k)\in UET\} & m=1\\ \{u_q|(u_i,t_k)\in UET,u_q\in NU(u_i)_{m-1}\} & m>1\end{cases} \quad (7)$$

定义 3.2.(主题 t_k 的 m 步可达点偏好度)给定主题关注网络 $G=(V,E,W)$,$\forall t_k\in T$,如果 u_q 为 t_k 的一步可达点,则记 $Per_1(u_q,t_k)$ 为 t_k 的一步可达点偏好度,如果 u_q 为 t_k 的 $m(m>1)$ 步可达点,则记 $Per_m(u_q,t_k)$ 为 t_k 的 m 步可达点偏好度,如式(8)所示.

$$Per_m(u_q,t_k)=\begin{cases}\mu(u_q)\times\dfrac{1}{\sum_{i=1}^{n}(u_q,t_k)} & m=1\\ \dfrac{\sum_{i=1}^{|NU(u_q)_1\cap RN_{m-1}|}Per_{m-1}(u_i,t_k)}{|NU(u_q)_1\cap RN_{m-1}|}\times\dfrac{\mu(u_q)}{|NU(u_q)_1|} & m>1\end{cases} \quad (8)$$

定义 3.3.(主题影响力)给定主题关注网络 $G=(V,E,W)$ 对于 $\forall t_k\in T$,存在 $u_j\in NU(u_i)_1$,$u_q\in NU(u_i)_2$,$u_p\in NU(u_i)_L$,经计算,网络达到亚稳态时,步长为 L,并且得到 M 个一步可达点,N 个两步可达点,P 个三步可达点…Q 个 L 步可达点,则最后的主题影响力为所有可达点偏好度的和,记为 $Per(t_k)$,如式(9)所示.

$$Per(t_k)=\frac{\sum_1^M Per_1(u_i)+\sum_1^N Per_2(u_j)+\sum_1^P Per_3(u_q)+\cdots+\sum_1^Q Per_L(u_p)}{M+N+P+\cdots+Q} \quad (9)$$

4.2　基于主题关注模型的影响最大化算法

和多数挖掘种子顶点算法相同,将种子集 S 的影响集规

模 $|influenceSet(S)|$ 做为目标函数,利用经过优化的 CELF 算法挖掘种子顶点的算法 TA_CELF 如下.

算法 1. TA_CELF(G,AU,UAT,t_k,k)

输入:主题关注网络 G、用户-用户矩阵 AU、用户-主题矩阵 UAT、信息类型 t_k、种子集规模 k;

输出:种子集 S.

Begin

(1) Initialize $S=\varnothing$,$Q=\varnothing$,$R=1000$;

(2) For each edge $(u_i,u_j)\in EU$ do

(3) $p(u_i,u_j,t_k)=w_1\times Inf(u_i,u_j)+w_2\times F(Per(u_i,t_k),Per(u_j,t_k))$

(4) End for

(5) $S_{sub}=getSubSeeds(G,AU,UAT)$;

(6) For each $u_i\in U$ do

(7) $u_i.ins=0$;

(8) For $j=1$ to R do

(9) $u_i.ins+=|influenceSet(S_{sub},\{u_i\})|$;

(10) End for

(11) $u_i.ins+=u_i.ins/R$;

(12) $u_i.flag=0$;

(13) Add u_i to Q by $u_i.ins$ in descending order;

(14) End for

(15) While $|S|<k$ do{

(16) $u_i=Q[top]$;

(17) If $u_i.flag==|S|$ then

(18) $S=S+\{u_i\}$;

(19) $Q=Q-\{u_i\}$;

(20) Else

(21) $u_i.ins=0$;

(22) For $j=1$ to R do

(23) $u_i.ins+=|\inf luenceSet(G_{res_{t_k}},S+\{u_i\})|$;

(24) End for

(25) $u_i.ins=u_i.ins/R-|\inf luenceSet(G_{res_{t_k}},S)|$;

(26) $u_i.flag==|S|$;

(27) Resorted Q by $u_i.ins$ in descending order;

(28) End if

(29) End While

(30) Return S.

End

算法 1 主要描述了挖掘种子顶点的细节,(1)-(4)行描述了计算主题关注网络中用户之间的传播概率;(5)行描述了计算种子顶点候选集合,(6)-(14)行描述了计算每一个顶点的影响范围;(15)-(30)行描述了迭代挖掘种子顶点.

其中,第(5)行函数 $getSubSeeds(G,AU,UAT)$ 的功能是,对给定的主题关注网络 G、用户-用户矩阵 AU 和用户-主题矩阵 UAT 得到网络中各个用户对于主题 t_k 的偏好度,根据式(8)计算主题网络中任意顶点 u_i 的偏好度,根据公式(9)计算网络中该主题的影响力,将该影响力作为阈值,进而得到种子顶点候选集合 S_{sub},具体实现过程如算法 2 所示.

算法 2. $getSubSeeds(G,AU,UAT)$

输入:主题关注网络 G、用户-用户矩阵 AU 和用户-主题矩阵 UAT

输出:种子顶点候选集合 S_{sub}

(1) For each vertex $u_i\in U$ do

(2) $u_i.preference=Per_m(u_i,t_k)$;

(3) End for

(4) $Per(t_k)=\dfrac{\sum_1^M Per_1(u_i)+\sum_1^N Per_2(u_j)+\sum_1^P Per_3(u_q)+\cdots+\sum_1^Q Per_m(u_p)}{M+N+P+\cdots+Q}$;

(5) $S_{sub}=\varnothing$;

(6) For each vertex $u_i\in U$ do

(7) $S_{sub}=\{u_i|u_i\in U,Per(u_i,t_k)\geqslant Per(t_k)\}$

(8) End for

(9) Return S_{sub}.

5 实验分析

5.1 实验设置

5.1.1 数据集介绍

通过爬虫程序,在豆瓣网上获得了用户间互相关注以及电影喜好和评论等相关信息的数据集. 对于用户的行为数据,收集了关于用户看过的电影和书写的电影评论两部分. 数据经过整理,共得到带有评分数据的电影记录为 563173 条,带有评分数据的电影评论为 119766 条,用户为 2253 人. 其中涉及的电影为 29009 部,共包含电影类别为 36 个,如表 1 和表 2 所示.

表 1 不同类型顶点数量统计

Table 1 Number statistics for different types of vertexes

顶点类型	user	topic	总计
顶点个数	2253	36	2289

表 2 不同类型边数量统计

Table 2 Number statistics for different types of edges

参与实体	边条数
user-user	27988
user-topic	34818
总计	62806

5.1.2 传播模型与参数设置

在豆瓣数据集上实现了基于 TA_IC 模型的 CELF 算法,并且,在实验中采用蒙特卡洛模拟的方法,将 TA_IC 模型中的传播过程模拟 1000 次. 经过多次实验,将公式(3)中的权重 w1 和 w2 分别设为 0.4 和 0.6.

5.1.3 度量指标

为了评价算法的性能,将本文模型的特点与文献[16]评价指标相结合,提出如下 3 个度量指标:ISRT(influence spread result on a specific topic)、ISST(influence spread of a seed set on a specific topic)和 ISRNT(influence spread result on a specific topic not direct to the topic).

①$ISRT(S,t)$:指种子集对于特定主题 t 的传播最终得到的影响范围. $ISRT$ 的定义如下:

$$ISRT(S,t_k)=\sum_{u_i\in seedSet(S)}|\inf luenceSet(u_i)| \quad (10)$$

②$ISST(S,t)$:与文献[11]类似,指种子集对于特定主题 t 的影响传播. 该指标是在考虑用户偏好的基础上评估种子集 S 对特定主题类型的影响. *ISST* 的定义如下:

$$ISST(S,t_k)=\sum_{u_i\in seedSet(S)}Per(u_i,t_k) \tag{11}$$

③$ISRNT(S,t)$:指种子集对于特定主题 t 的传播最终得到的影响范围,该范围不包括直接与 t 相关联的顶点. *ISRNT* 的定义如下:

$$ISRNT(S,t_k)=\sum_{u_i\in seedSet(S)}|\inf luenceSet(u_i)|-NU(t_k) \tag{12}$$

5.1.4　实验环境

实验的硬件环境是 Intel(R) Core(TM) i7-4760HQ 四核的 CPU,内存 8GB;软件环境是 Windows 8 系统,Java JDK 1.7,Eclipse 4.3.

5.2　实验结果与分析

主题类别共计 36 种,经过统计与各个主题直接相关的用户数如图 3 所示,用户数≥1000 的主题共计 17 个,500≤用户数<1000 的主题数共计 7 个,用户数<500 的主题数共计 12 个.

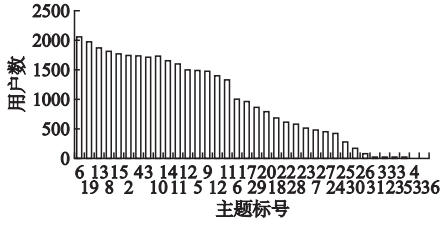

图 3　与主题直接相关用户

Fig. 3　Users directly relate to topic

希望通过实验观察是否更多的与主题间接关注的用户被影响,所以在选择待传播主题向量时,不考虑与主题直接相关的用户数≥1000 的主题,并且用户数<500 的主题传播效果不明显,所以,将待传播主题向量定为 500≤用户数<1000 的主题,在这 7 个主题中经过挑选,选择 3 个,分别如表 3 所示.

表 3　主题类型及与其直接相关用户数

Table 3　Topic type and numbers of directly related users

主题类型	歌舞	纪录片	西部
主题序号	29	20	18
直接用户数	848	790	660

本次实验,在 *ISST*、*ISRT* 和 *ISRNT* 三个指标上对算法 TA_CELF 与 CELF 和 L_GAUP 进行比较,同时给予不同主题,考察主题对算法 TA_CELF 的影响.

1)TA_CELF 算法与 CELF 算法和 L_GAUP 算法的 *ISST* 的比较. 在实验中,将种子集的规模依次从 10 取到 50,以主题 20 为传播向量进行传播,得到的 *ISST* 随种子集规模变化的关系如图 4 所示. 由图 4 可见,种子集 S 对三种算法的影响是不同的,随着 k 值的增大,三种算法下的种子集影响规模都逐渐在增加,同时,算法 TA_CELF 和 L_GAUP 相对 CELF 效果要好. 算法 TA_CELF 和 L_GAUP 影响范围相接近,主要是两种算法均是针对主题,挖掘种子顶点的算法都是 CELF,进而得到的种子顶点较为相似,因此,算法 TA_CELF 和 L_GAUP 效果较好,且优于算法 CELF.

2)TA_CELF 算法与 CELF 算法和 L_GAUP 算法的 *ISRT* 的比较. 在实验中,将种子集的规模依次从 10 取到 50,以主题 20 为传播向量进行传播,得到三种算法的 *ISRT* 随种子集规模变化的关系如图5所示. 由图5可见,随着k值的增大,

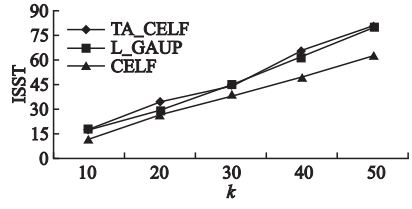

图 4　ISST 随 k 的变化

Fig. 4　ISST vs. k

三种算法的种子集最终影响范围都逐渐在增大,并且 TA_CELF 算法的最终影响效果明显优于 L_GAUP 和 CELF. 这是因为在网络中挖掘种子顶点集的过程中,算法的目标函数是 $|influenceSet(S)|$,即寻找被影响顶点增量个数最大的顶点,而 L_GAUP 的目标函数是 $|influenceIP(S)|$,即寻找被影响顶点的重要度较大的顶点,因此,在同一个主题下,ISST 随 k 的变化都是上升趋势,且 TA_CELF 算法效果最好.

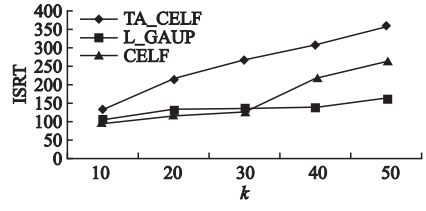

图 5　ISRT 随 k 的变化

Fig. 5　ISRT vs. k

3)不同主题下的 *ISRNT* 的比较. 在实验中,将种子集的规模依次从 10 取到 50,以主题 29,20 和 18 分别为传播向量进行传播,得到的 *ISRNT* 随种子集规模变化的关系如图 6 所示. *ISRNT* 这个指标考虑的是,与主题没有直接相关的顶点被影响的范围. 由图6可见,随着k值的增大,不同主题下的种

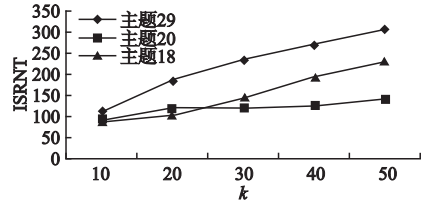

图 6　ISRNT 随 k 的变化

Fig. 6　ISRNT vs. k

子集最终影响范围都逐渐在增大,这是因为在网络中挖掘种子顶点集的过程中,计算过程如式(11),与 *ISRT* 的 $|influenceSet(S)|$ 是成正比例变化的,即寻找被影响顶点增量中与主题不直接相关个数最大的顶点,所以,致使任何一个主题下,ISRNT 随 k 的变化都是上升趋势.

6 结 论

本文针对现有的影响最大化算法都是应用于传统的社会网络的问题,提出一种基于主题关注模型的影响最大化算法.在该模型下,采用集对联系度与随机游走相结合的方法,对豆瓣数据采集来的用户主题数据进行偏好计算,得到备选种子集;然后,在备选种子集中挖掘种子顶点;最后,给出 ISST,ISRT,ISRNT 三个指标来评价实验结果. 实验结果表明,不同主题下,种子集的影响范围是不同的;同时,对于同一主题,随着种子集的增加,影响范围也逐渐增加,对于没有直接关注主题的用户也被影响,ISST,ISRT,ISRNT 三个指标均呈现出增长的趋势.

由于随机游走需要进行矩阵运算,在单机运行程序时,效率较低. 所以,下一步工作:

1)将程序改进,可以进行并行计算.

2)将在新浪微博、Twitter 和人人网等数据集上验证其影响范围.

References:

[1] Chen Wei. Time-critical influence maximization in social networks with time-delayed diffusion process[C]. In Association for the Advancement of Artificial Intelligence,2012:592-598.

[2] Li Yan-hua,Chen Wei,Wang Ya-jun. Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships[C]. Web Search and Data Mining,2013:657-666.

[3] Aslay C, Baribieri N, Barbieri N. Online topic-aware influence maximization queries[C]. International Conference on Extending Database Technology,2014:279-291.

[4] Kutzkov K, Bifet A, Bonchi F. Strip: stream learning of influence probabilities[C]. Proc of the 19th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2013: 275-283.

[5] Domingos P, Richardson M. Mining the network value of customers [C]. Proc of the 7th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2001:57-66.

[6] Kempe D, Kleinberg J, Tardos E. Maximizing the spread of influence through a social network[C]. Proc of the 9th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2003:137-146.

[7] Leskovec J, Krause A, Guestrin C, et al. Cost-effective outbreak detection in networks[C]. Proc of the 13th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2007:420-429.

[8] Kimura M, Saito K. Tractable. Models for information diffusion in social networks[C]. Proc of the Knowledge Discovery in Database (PKDD 2006), Berlin: Springer, 2006:259-271.

[9] Goyal A, Bonchi F, Lakshmanan L. Learning influence probabilities in social networks[C]. Proc of the 20th Int Conf Companion on World Wide Web. New York: ACM, 2010:241-250.

[10] Guo Jing, Zhang Peng, Zhou Chuan. Personalized influence maximization on social networks[C]. International Conference on Information and Knowledge Management, 2013:199-208.

[11] Saito K, Kimura M, Ohara K. Motoda H: Learning continuous-time information diffusion model for social behavior data analysis[C]. In Machine Learning-first Asian Conference on Machine Learning, 2009:322-337.

[12] Saito K, Kimura M, Ohara K. Motoda H: behavioral analysis of information diffusion models by observed data of social network [C]. In Social Computing-third International Conference on Social Computing, Bchavioral Modeling, and Prediction, 2010:149-158.

[13] Liu Lu, Tang Jie, Han Jia-wei, et al. Mining topic-level influence in heterogeneous networks[C]. In International Conference on Information and Knowledge Management, 2010:545-576.

[14] Barbieri, Bonchi F, Manco G. Topic-aware social influence propagation models[C]. In International Conference on Information and Knowledge Management, 2012:81-90.

[15] Zhou J, Zhang Y, Cheng J. Preference-based mining of top-K influential nodes in social networks[J]. Future Generation Computer Systems, 2014, 31(1):40-47.

[16] Guo Jing-feng, Lv Jia-guo. Influence maximization based on information preference[J]. Journal of Computer Research and Development, 2015, 52(2):533-541.

附中文参考文献:

[16] 郭景峰,吕加国. 基于信息偏好的影响最大化算法研究[J]. 计算机研究与发展,2015,52(2):533-541.