社交网络中基于信息词频和节点相似度的影响最大化算法

胡启志,颜 娜,谢瑾奎

(华东师范大学 计算机科学技术系,上海 200241)
E-mail:jkxie@cs.ecnu.edu.cn

摘 要: 深入挖掘社交网络中传播力较强的个体,并利用其进行产品营销往往会达到事半功倍的效果,影响最大化问题就是在特定社交网络中寻找影响力较大的个体.为了更加准确的评估影响力,本文不仅从节点相似度方面进行改进,而且从信息内容本身出发,基于信息在社交网络中的传播,结合信息词频等信息自身特点来刻画节点的影响力,提出了基于信息词频和节点相似度的影响最大化算法(IMFS,Influence Maximization algorithm based on term Frequency and node Similarity).随后,在真实的社交网络中对该算法进行了实验,并与传统的影响最大化算法对比,实验结果表明由 IMFS 得到的集合的影响范围大于其他启发式算法的结果,同时算法的运行速度也有相应的提高,说明了本文提出的算法是解决影响最大化问题的有效算法.

关 键 词: 影响最大化;节点相似度;信息词频;EM 算法
中图分类号: TP393 **文献标识码:** A **文 章 编 号:** 1000-1220(2017)02-0259-05

Influence Maximization Algorithm Based on Term Frequency and Node Similarity in Social Networks

HU Qi-zhi,YAN Na,XIE Jin-kui
(Department of Computer Science and Technology,East China Normal University,Shanghai 200241,China)

Abstract: Deeply mining the individual who has strong propagation force in social network,and using it to promote products tends to achieve more with less. Influence maximization algorithm aims to solve this problem in the specified social network. In order to estimate the influencemore accurate,we not only improve the nodes similarity,but also consider the information content itself. Based on the transmission of information in a social network,combining with the term frequency and other information characteristics to depict the node's influence,we proposed a new algorithm based on the diffusion of information,which considered the perspective of term frequency and node similarity. We implement this algorithm in the real world data set. Compared to traditional influence maximization algorithms,the test result shows that our algorithm is better than other heuristic algorithms. Moreover,our algorithm also runs faster than others. It is verified that our algorithm is an effective solution to influence maximization problem.

Key words: influence maximization; node similarity; term frequency; EM algorithm

1 概 述

社交网络是指社会中个体与个体之间通过特定关系,如朋友关系、论文作者间文章引用关系等形成的复杂网络关系,这种复杂的网络关系成为信息传播和扩散的载体,对信息的传播有非常重要的影响,也为"病毒式营销"提供了途径.在社交网络中,信息所能影响的范围是指转发这条信息的节点数.不同的节点集合所能影响的范围有很大的差异,因此对影响力的估计在现实生活中有很重要的作用[1].例如,公司发布新产品时会通过电视、互联网等媒体发布广告来推广自己的产品,但这种方式成本较高,往往需要投入大量的人力物力.此外,还有些公司会提供产品给部分人免费试用,这种方式的花费明显比前一种少很多,但是却不一定能达到很好的效果.于是如何挑选合适的试用人员就显得愈加重要.通常可以选择一些关键性的个体让其试用,之后再推荐给其朋友,朋友之间的推荐效果往往会更好[2-3],这里就涉及到影响力问题.显然,对于固定数目的个体,其影响力越大,所能影响的客户越多,对产品的宣传也就越有利.影响力最大化算法就是针对这类问题,找出一部分节点构成集合,使得该集合所能影响的范围最大,影响的节点个数最多.影响最大化算法主要有基于独立级联模型和线性阈值模型两种.

Jung 等人[4]基于独立级联模型以及扩展的独立级联模型 IC-N 提出了改进复杂度的影响力排名算法 IRIE,加入贪心算法中对边际影响力的计算,然后运用影响力估计方法选择影响力最大的前 k 个节点.该方法在一定程度上具有较好的扩展性和鲁棒性,但同其他基于节点的度的算法一样没有考虑到信息对节点影响力的作用.

Gong 等人[5]利用 PageRank 算法计算影响力最大的前 k 个节点,并且考虑了节点的相关性和重要性.文中提出了一种基于 PageRank 的传播模型,并计算两个节点之间的权值,最后利用 top-K 贪心算法计算影响力最大的前 k 个节点.这种算法只考虑了节点的度和相关性,没有结合具体的信息计算

收稿日期:2015-12-07 收修改稿日期:2016-03-02 基金项目:国家自然科学基金项目(61502170)资助. 作者简介:胡启志,男,1989 年生,硕士研究生,研究方向为社会计算、推荐系统、聚类算法;颜 娜,女,1992 年生,硕士研究生,研究方向为社会计算、算法博弈论;谢瑾奎(通信作者),男,1975 年生,博士,副教授,研究方向为算法博弈论、社会计算与分布式计算.

节点的影响力.

Mao 等人[8] 在研究微博影响力最大化模型中,引入阅读次数、阅读时间以及转发数量等属性. 但是在微博网络中,仅仅使用阅读次数和转发量来衡量一个用户的影响力,而不考虑用户所在网络中邻居节点的分布这一对信息传播范围影响较大的因素,导致算法达不到理想效果.

Cao 等人[9] 利用 k-核来计算每个用户的影响力大小,提出了基于 k-核的核覆盖算法(CCA). CCA 算法按核数和度数作为第一、第二关键字排序,在排序后的列表中选取没有被覆盖的点作为下一个候选解,记为 u,然后将与 u 距离小于 d 的所有节点都标记为覆盖状态. 该算法能够较准确的计算出用户的影响力,文中最后实验也证明了该算法较其他传统的影响最大化算法有更好的结果.

Guo 等人[10] 基于网络中节点的信息偏好分析社交网络影响力. 在潜在语义索引模型中,将用户-主题矩阵 R 分为 U、Σ、V 三个矩阵,其中 U、V 为正交矩阵,Σ 为对角矩阵. 于是下式成立:

$$R = U\Sigma V = U\Sigma^{1/2}\Sigma^{1/2}V$$

记 $X = U\Sigma^{1/2}$,$Y = \Sigma^{1/2}V$,于是:

$$C(a,t) = C + X(a) * Y(t)$$

在基于向量空间模型的计算方法中,利用余弦相似度计算偏好,根据事先设置的阈值计算信息的易感染网络. 将得到的网络中的节点按重要度排序,排序后的列表中的前 k 个节点即为影响力最大的节点集合. 其中重要度 IP 的计算方法如下:

$$IP(v) = Activity(v) + (1 + v.outdegree)$$

$Activity(v)$ 由一段时间内用户发表的评论数来计算. 基于偏好的计算方法在一定程度上考虑了节点自身的属性,但是每个节点除偏好之外还有其他的属性. 除自身的属性之外,该文中并没有考虑节点的朋友关系对其影响力的大小.

上述传统的影响最大化算法往往只考虑节点的度、节点所在的社区等与社交网络的结构相关的因素,没有深入挖掘信息内容本身和节点相似度等对传播的影响,导致算法往往达不到理想效果. 基于以上问题,本文在考虑了信息的内容和节点的相似度这两种影响信息传播的重要因素的基础上,对信息在网络中的传播进行了更准确和更细致的刻画. 最后我们利用两种真实的数据集进行实验,得到最具影响力的 K 个节点,并与其它相似算法进行对比分析,验证了本文算法的有效性.

文章的整体思路如下,第一小节对影响最大化进行简要概述,并对相关工作进行具体介绍; 第二小节对信息词频和节点相似度进行定义,并对本文提出的算法做详细的描述; 第三小节在两种真实的数据集上进行实验,并与其他影响最大化算法对比,验证本文算法的有效性; 第四小节是对文章的总结和对未来工作的展望.

2 基于信息词频和节点相似度的影响最大化算法 IMFS

影响最大化问题是指在特定的社交网络中,选取 k 个节点组成种子集,使其影响的节点的数目最多,即影响范围最大. 本文考虑到信息词频这一影响传播过程和范围的重要因素,并结合节点之间的相似性提出了基于信息词频和节点相似度的影响最大化算法 IMFS(Influence Maximization algorithm based on term Frequency and node Similarity),更能准确的描述信息的传播. 因为信息从一个节点传播到下一个节点不仅与信息词频有关,还和当前节点和下一个节点的相似程度有关. 例如两个节点的兴趣爱好都包括篮球,那么当一个节点发布一些关于篮球比赛的信息时,另一个节点就很可能会转发其信息,继而传播到其他节点. 但是两个节点的兴趣爱好不可能完全相同,而且除了兴趣爱好以外的其他特性也会有差异,所以当一个节点发布一些其他信息时,另一个节点也有可能不会转发,可见信息词频也是需要考虑的因素. 本文通过加入信息词频和节点相似度这两个因素,根据观测到的数据,利用极大似然模型结合期望最大化(EM) 对参数进行估计,得到影响力最大的 k 个节点.

2.1 信息词频

向量空间模型是将文本的内容转化为向量,按照向量空间中的向量运算来处理文本信息,并且以空间上的相似度表达语义的相似度,直观易懂,运算方便[10]. 将文档作为向量来处理,就可以通过计算两个向量之间的余弦相似度来计算两个文档之间的相似性. 将文档转化为向量有词频率表、词逆向文档频率表(TF-IDF) 等方法[11].

本文采用词频率表法,首先提取出每条信息的内容,以不重复计入的方式加入到单词表. 对于每一条消息,计算相应分量上单词的词频率,继续做归一化处理之后生成对应的向量. 设单词表为 W,长度为 n,由信息 i 生成的向量为 v,信息 i 中对应于 W 的每个单词的词频为 n_i,则有:

$$v_i = \frac{n_i}{\sum_i n_i} \tag{1}$$

例如有 4 条信息,分别为 [A,E,B,C,A],[D,B,C],[E,D,C,D] 和 [C,B,D,E,A]. 按上述方法,生成的单词表 W = [A,B,C,D,E],于是 [A,E,B,C,A] 的词频率向量为 [2,1,1,0,1],继续做归一化处理之后得到相应的 [0.4,0.2,0.2,0,0.2]. 对其他的 3 条信息做同样的处理,生成对应的向量 [0,1/3,1/3,1/3,0],[0,0,0.25,0.5,0.25] 和 [0.2,0.2,0.2,0.2,0.2].

2.2 节点相似度

互联网的快速发展推动了各种形式的社交网络的形成. 每一种社交网络的形成都是由一定数量的个体通过某种方式,比如共同爱好、朋友关系等建立联系. 对于社交网络中的任意两个节点可以通过距离、相似度等不同的度量来描述他们之间的紧密程度,其中相似度方法应用最为广泛. 相似度刻画了两个个体之间的相似程度,用来度量两个节点之间的关系. 相似度的分类也有很多种,比如余弦相似度、皮尔逊相似度、距离相似度、Jaccard 相似度等,其中余弦相似度的应用最为广泛.

在特定的社交网络中,一个节点与其邻居节点的相似度不仅与社交网络的结构有关,还与两个节点自身的属性,比如爱好,地区等信息有很大的关系. 一般来说,信息在网络中从一个节点传播到相邻的其他节点和该节点与其他邻居节点的相似度存在正相关. 两个节点之间的相似度越大,信息从该节点传播到另一个节点的概率也会相应的增大. 例如,在微博的转发过程中,越相似的两个人更有可能转发同样的信息. 社交网络中,计算两个节点之间的相似度的方法有很多. 两个节点之间的相似度也与该节点的邻居节点的分布以及该节点本身

的一些属性如兴趣爱好、年龄、地区等因素有关. 将两个节点的相关属性用向量的形式表示出来,用 X,Y 表示. 下面分别介绍几种相似度的计算方法[12]:

1) 余弦相似度:

$$Sim(X,Y) = \cos(X,Y) = \frac{X \cdot Y}{\|X\| \cdot \|Y\|} \tag{2}$$

2) 皮尔逊相似度:

$$Sim(X,Y) = r_{xy} = \frac{n\sum XY - \sum X\sum Y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}} \tag{3}$$

3) Jaccard 相似度:

$$Sim(X,Y) = \frac{|\Gamma(X) \cap \Gamma(Y)|}{|\Gamma(X) \cup \Gamma(Y)|} \tag{4}$$

本文采用改进的 Jaccard 相似度结合节点自身的若干属性计算两个节点之间的相似度. 分别用 $\Gamma(X)$ 和 $\Gamma(Y)$ 表示节点 X,Y 的邻居节点,其中 $\Gamma(x)$ 和 $\Gamma(y)$ 都包括自身特性,同时提取节点相关的属性值,例如兴趣,地理位置等信息,综合计算两个节点之间的相似度.

2.3　影响最大化算法框架 IDFS

信息在特定社交网络中的传播过程和影响范围的大小不仅与信息的内容有关,还与所在的社交网络中各个节点的属性以及网络本身的结构有很大的关系. 首先,不同的信息在同一网络中传播的速度和影响的范围不一定相同; 其次,如果社交网络的结构发生变化,信息的传播也会随之发生改变,而且不同的人对同一种信息的处理方式也不一样. 例如,在微博的转发过程中,同一条微博,每个人转发与否与每个人的兴趣相关联.

本文基于信息在网络中传播的过程,选择影响力最大的 k 个节点. 与传统的影响最大化算法只考虑节点的度、节点所在的社区等与社交网络的结构相关的因素相比,本文提出的模型同时考虑信息词频和网络节点相似度对传播过程的影响. 设 M 表示信息的数量,N 表示社交网络中节点的个数,$\lambda_1,\lambda_2,\cdots,\lambda_M$ 分别表示信息经过预处理后对应的向量表示形式,$v_1,v_2,\cdots,v_N$ 分别表示社交网络中的节点,D_{ij} 表示信息内容对传播过程的影响的概率,s_{ij} 表示社交网络中 i 节点与 j 节点之间的相似度. 设定在信息的传播过程中 λ_i 服从参数为 α 的狄利克雷分布,即 $\lambda_i \sim Dir(\alpha)$. D_{ij} 服从参数为 λ_i 的多项分布,即 $D_{ij} \sim Mult(\lambda_i)$. 不同个体对信息传播过程的影响概率用 S_{ij} 表示. 设定 S_{ij} 服从于参数为 μ,σ 的正态分布,参数之间的关系如图 1 所示. 于是所有当前可观测概率的极大似然估计可以表示为:

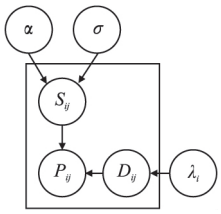

图 1　模型的图形表示

Fig. 1　Graphic representation of model

$$L = \prod_{i,j}\prod_k p(v_i \xrightarrow{\lambda_k} v_j, \forall i,j,k, 1 \leqslant i,j \leqslant N, 1 \leqslant k \leqslant M \mid D_{ij}, S_{ij}) \tag{5}$$

引入各参数的分布,于是式(5)可以转化为:

$$\begin{aligned} L &= \prod_{i,j}\prod_k p(v_i \xrightarrow{\lambda_k} v_j, \forall i,j,k, 1 \leqslant i,j \leqslant N, 1 \leqslant k \leqslant M \mid \alpha,\mu,\sigma^2) \\ &= \prod_{i,j}\int p(\lambda_k \mid \alpha)\, p(v_i \xrightarrow{\lambda_k} v_j, \forall i,j,k, 1 \leqslant i,j \leqslant N \mid \lambda_k,\mu,\sigma^2) \end{aligned} \tag{6}$$

其中 $P(v_i \xrightarrow{\lambda_k} v_j, \forall i,j,k, 1 \leqslant i,j \leqslant N \mid \lambda_k,\mu,\sigma^2) = s_{ij}^{pq}$. 通过最大化公式(6),对参数 α,μ,σ^2 进行估计得到:

$$Arg \max_{\alpha,\mu,\sigma^2} logL$$

下面通过期望最大化算法(EM)对参数 α,μ,σ^2 进行估计. 引入后验分布 $Q = Q(\lambda,Z) = Q(\lambda)Q(Z)$. 其中 $Q(Z_{ij}) \sim Mult(\pi_i)$, $Q(\lambda_i) \sim Dir(\gamma_i)$. 根据凸函数的性质:

$$\begin{aligned} logL \geqslant\ & \sum E_Q logP(\lambda_k \mid \alpha) + \sum\sum E_Q logP(D_{ij} \mid \lambda_k) + \\ & \sum\sum E_Q logP(v_i \xrightarrow{\lambda_k} v_j \mid D_{i,j},\mu,\sigma^2) + H(Q) \end{aligned} \tag{7}$$

其中 H 表示 Q 的熵.

$$E_Q logP(D_{i,j} \mid \lambda_k) = -log\frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)} + \sum_k(\alpha_k - 1)log\theta_{ik} \tag{8}$$

$$E_Q logP(D_{i,j} \mid \lambda_k) = log\frac{N!}{\pi_{i1}!\cdots\pi_{iN}!}\sum_k \pi_{ik}\theta_{ik} \tag{9}$$

$$\sum\sum E_Q logP(v_i \xrightarrow{\lambda_k} v_j \mid D_{i,j},\mu,\sigma^2) + H(Q) = \sum_{ij}\sum_k \pi_{ij}s_{ij} \tag{10}$$

$$H(Q) = -\sum E_Q logQ(\theta_i) + \sum E_Q logQ(D_{ij}) = log\frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)} - \sum_k(\alpha_k - 1)log\theta_{ik} - \sum\sum \pi_{ij}log\pi_{ij} \tag{11}$$

在 EM 算法的迭代过程中,E 步更新参数 π_{ij} 和 γ_i,M 步更新参数 α,μ,σ^2. 按下面的式子更新 π_{ij} 和 γ_{ik}:

$$\pi_{ij} \sim s_{ij}exp(logp_{ij} - ln\sum_k p_{ik}) \tag{12}$$

$$\gamma_{ik} = \alpha_k + \sum \pi_{ij} \tag{13}$$

在期望最大化算法的过程中,使用牛顿法沿着梯度的方向循环迭代.

$$\frac{\partial logL}{\partial \alpha_k} = \sum -\alpha_k + log(\sum \alpha_k) + p_{ij} - log(p_{ij}) \tag{14}$$

式(14)中,p_{ij} 表示当前节点将信息传播到邻居节点的概率. 迭代过程中,p_{ij} 的初始值设置为 $\frac{1}{n}$,其中 n 表示邻居节点的个数.

$$\frac{\partial logL}{\partial \mu} = \sum\sum \frac{\pi_{ij}p_{ij}ln(s_{ij})}{s_{ij}\sigma^2(\sum p_{ij})^2} \tag{15}$$

$$\frac{\partial logL}{\partial \sigma^2} = \sum\sum \frac{\pi_{ij}}{s_{ij}} \times \frac{p_{ij}(ln(s_{ij} - \mu))^2}{2(\sum p_{ij})^2} \tag{16}$$

使用 EM 算法迭代直到收敛时停止迭代.

算法 1 给出了利用 EM 算法结合信息词频和节点相似度估计 α,μ,σ^2 的框架. 按(12)(13)式更新参数 π_{ij} 和 γ_i,按(14)(15)(16)式更新参数 α,μ,σ^2. 整个迭代过程到 EM 算法收敛时停止. 算法 2 首先利用算法 1 中估计出的参数值计算每个节点将信息传播到邻居节点的概率,然后根据设定的阈值计算每个节点将信息传播到其邻居的概率大于阈值的邻居节点的个数,最后以影响的邻居节点的个数为关键字排序来选择影响力最大的前 k 个节点.

算法 1. 基于 EM 的参数估计算法

输入: 社交网络图 G,信息 I

输出: α,μ,σ^2

While Convergence criteria not met do

按式

$$\pi_{ij} \sim s_{ij} exp(logp_{ij} - ln\sum\nolimits_k p_{ik})$$

$$\gamma ik = \alpha_+ \Sigma \pi_{ij}$$

更新参数 π_{ij} 和 γ_i.

按式

$$\frac{\partial logL}{\partial \alpha_k} = \Sigma - \alpha_k + log(\Sigma \alpha_k) + p_{ij} - log(p_{ij})$$

$$\frac{\partial logL}{\partial \mu} = \sum\sum \frac{\pi_{ij} p_{ij} ln(s_{ij})}{s_{ij}\sigma^2(\Sigma p_{ij})^2}$$

$$\frac{\partial logL}{\partial \sigma^2} = \sum\sum \frac{\pi_{ij}}{s_{ij}} \times \frac{p_{ij}(ln(s_{ij} - \mu))^2}{2(\Sigma p_{ij})^2}$$

对参数 α, μ, σ^2 进行更新,最大化 logL.

End while.

算法 2: 基于信息词频和节点相似度的影响最大化算法 IMFS

输入: 社交网络图 G,信息 I,参数 k,阈值 δ

输出: 影响力最大的前 k 个节点

调用算法 1 训练参数,得到各个参数的训练值

计算每个节点将信息传播到邻居节点的概率

For 每个节点 v

If v 传播信息到邻居节点的概率 $> \delta$

v 的传播值加 1

End if

End for

选取传播值最大的前 k 个节点.

3 实 验

前面章节对本文提出的基于信息词频和节点相似度的影响力最大化算法作了详细的介绍,本节运用互联网上的真实数据作为数据集,对算法的运行效果进行评估,分析总结.

3.1 实验数据和实验环境

本文从互联网上选取两个典型的数据集作为测试数据集,分别是新浪微博和 Twitter 上用户关系的数据集合. 具体获取数据集的方法是以某个用户为初始点,利用 Python 的 urllib 包的网页提取 API,抓取其关注的用户和被关注的用户以及每个用户自身的属性信息,例如兴趣爱好等. 然后分别用该用户的关注列表和被关注列表作为初始集合,重复上述过程. 使用该方法,本文抓取了两份数据集作为算法的测试数据,分别为新浪微博 8461 个用户的相关信息,共有 56028 条边和 Twitter5218 个用户,共有 34871 条边,如表 1 所示.

表 1 两种测试数据集

Table 1 Two datasets for testing

	节点数	边数
新浪微博	8461	56028
Twitter	5218	34871

本文实验所采用的编程语言分别为Java1.7和Python3.4,运行在 Ubuntu14.04 上,处理器为 64 位 4GB 内存. 测试数据集运用 Python 的 urllib 库提供的网络相关函数编程获取,期望最大化(EM)算法使用 java 编程语言实现. 其中中文分词需要调用 IK Analyzer 库,该库是一个开源的,基于 Java 编程语言实现的轻量级的中文分词工具包.

对算法的评估包括算法的效率和以该算法得到的种子集的影响范围,即能够影响的节点的大小. 为了评估算法的效率,便于与其他影响最大化算法的实验结果做比较,本文使用以下评价指标:

1) 算法的运行时间: 用于评估算法的运行速度;

2) 平均差异: 用于对不同算法得出的种子集的影响范围做比较.

平均差异为文献 [2] 中提出的影响最大化算法的评价指标.

$$Diff(A,B,i) = \frac{\sigma(A,i) - \sigma(B,i)}{\sigma(B,i)}$$

其中 $\sigma(B,i)$ 表示算法 B 在种子集合大小为 i 时所能影响的节点个数. 影响最大化算法 A 和 B 的平均差异定义为:

$$\mathrm{f}\ Diffavg(A,B,i) = \frac{1}{k}\sum_{i=0}^{k} Diff(A,B,i)$$

其中 k 表示每次实验中种子集合的大小. 在本实验中 k 的范围为 10、20、30、40、50.

本文实验中所采用的对比算法为 Google 用来表示网页影响力的 PageRank 算法[7],基于影响力排名和影响力评估相结合的启发式算法 IRIE[8].

3.2 实验结果和分析

该小节给出了算法运行在新浪微博和 Twitter 两个测试数据集上的实验效果. 图2给出了在不同种子集合下新浪微博的实验结果,图 3 是运行在 Twitter 数据集上的实验结果. 两种数据集的运行结果总结于表 2 和表 3.

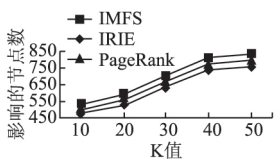

图 2 新浪微博数据集测试结果

Fig. 2 Test results on SinaMicroBlog

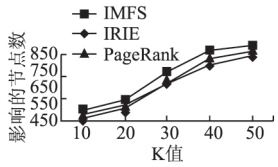

图 3 Twitter 数据集测试结果

Fig. 3 Test results on Twitter

表 2 新浪微博数据集测试结果(影响的节点数目)

Table 2 Test results on SinaMicroBlog(influenced nodes)

算法 \ K 值	10	20	30	40	50
IMFS	**526**	**587**	**703**	**808**	**831**
IRIE	481	527	628	738	761
PageRank	514	552	663	785	808

图 2 中显示了三种算法在新浪微博数据上运行结果的对比. 由图可以看出本文提出的 IMFS 算法所得的种子集的影响范围要比IRIE算法和PageRank算法都广,分别提高了大约 11.9% 和 6%. 图 3 显示了三种算法在 Twitter 数据集上运行结果的对比,IMFS 算法所得的种子集的影响范围较 IRIE 算法和 PageRank 算法提高了大约 11.2% 和 9.6%.

表 3 Twitter 数据集测试结果(影响的节点数目)

Table 3 Test results on Twitter (influenced nodes)

算法 \ K 值	10	20	30	40	50
IMFS	**289**	**340**	**467**	**574**	**594**
IRIE	253	292	397	502	547
PageRank	273	313	426	532	562

下面对上述实验结果做相应的分析. PageRank 算法是 Google 的网页排名算法,用于标示网页的等级. 在影响最大化中,PageRank 算法侧重分析社交网络中特定节点的邻居节点对其影响的大小,而忽略了信息内容本身对传播过程的影响. 而 IRIE 算法是一种启发式算法,用于对影响力进行排名和评估. 它仅仅考虑了节点边际影响力的增长,没有考虑节点自身的特点. 本文提出的 IMFS 算法基于社交网络的结构、节点自身的属性和信息内容,运用期望最大化算法(EM) ,对数据集的特点和社交网络的结构有更加精确的描述,综合对节点的影响力进行计算. 在影响信息传播的因素方面考虑的更加全面,从而对种子集影响力的估计有一定提升,增加了种子集最后的影响范围,得到了较好的实验结果.

算法在两种数据集上的运行时间如表 4 所示. 在最大化的过程中利用牛顿法迭代计算目标函数的最大值,由表 4 可以看出 IMFS 算法在效率上较 IRIE 算法和 PageRank 算法也有一定程度的提高.

表 4　不同算法在两种测试集上的运行时间(单位: 秒)

Table 4　Running time of algorithms on two datasets(unit: s)

	PageRank	IRIE	IMFS
新浪微博	10. 837	11. 759	10. 043
Twitter	7. 831	9. 053	7. 855

综合上述实验结果来看,本文提出的 IMFS 算法在实验效果和算法的运行效率上较 IRIE 算法和 PageRank 算法都有相应的提高,取得了较好的预测效果. 由此也说明信息在社交网络的传播不仅与节点的度和相似度有关,还与信息词频等相关,仅仅考虑单一因素都会影响计算结果的准确性.

4　结论和未来工作

影响力的计算在营销、广告等领域的重要性越来越明显,影响最大化算法在这些领域的运用对算法的有效性要求也越来越高. 本文分析了现有影响最大化算法的缺点,提出了基于信息词频和节点相似度的影响最大化算法,不仅考虑了节点周围的邻居对信息传播的影响,还结合节点自身的属性和信息词频更加全面的计算影响力,提高预测信息的影响范围. 最后运行新浪微博和 Twitter 两个数据集进行实验,实验结果也显示出本文算法较 IRIE 算法和 PageRank 算法有较为显著的提高. 同时也说明信息在社交网络中的传播与多种因素有关,只考虑某一方面的因素不能对信息影响的范围进行准确的估计,应该综合多种因素加以计算,提高算法的有效性. 在算法的效率方面,IMFS 算法比其他两种算法也有相应的提高.

本文在计算节点的属性值和节点之间的相似度之后直接假设其符合多项分布和正态分布,没有考虑其他可能的分布对计算结果的影响. 同时,对信息内容的处理过于粗糙,只简单使用了词频率表法,没有对信息进行预处理. 进一步的工作将基于以上两点,对节点信息及网络结构进行处理,使得对社会网络的建模更加贴合实际,提高模型的精准性.

References:

[1] Dominggos P,Richardson M. Mining the network value of customers [C] . Proc of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD) ,New York: ACM, 2001: 57-66.

[2] Granovetter M. The strength of weak ties [J] . American Journal of Sociology,1973,78(6) : 1360-1380.

[3] Burt R S. The social structure of competition [J] . Networks and Organizations: Structure,Form,and Action. Boston: Harvard Business School Press,1992: 57-91.

[4] Weng Jian-shu,Lim Ee-peng,Jiang Jinh,et al. Twitterrank: finding topic-sensitive influential twitterers [C] . Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM) ,New York,USA,2010: 261-270.

[5] Krackhardt D. The strength of strong ties: the importance of philos in organizations [J] . Networks and Organizations: Structure,Form, and Action. Boston: Harvard Business School Press,1992: 216-239.

[6] Jung K,Heo W,Chen W. IRIE: scalable and robust influence maximization in social networks [C] . Proceedings of the 12th International Conference on Data Mining (ICDM) ,Brussels,Belgium, 2012: 918-923.

[7] Gong Xiu-wen,Zhang Pei-yun. Research propagation model and algorithm for influence maximization in social network based on PageRank [J] . Computer Science,2013,40(6) : 136-140.

[8] Mao Jia-xin,Liu Yi-qun,Zhang Min,et al. Social influence analysis for micro-blog user based on user behavior [J] . Chinese Journal of Computers,2014,37(4) : 1-10.

[9] Cao Jiu-xin,Dong Dan,Xu Shan,et al. A k-core based algorithm for influence maximization in social networks [J] . Chinese Journal of Computers,2015,38(2) : 238-248.

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

[11] Berry M,Dumais S,Obrien G. Using linear algebra for intelligent information retrieval [J] . Siam Review,1995,37(4) : 573-595.

[12] Li Ai-ping,Di Peng,Duan Li-guo. Document sentiment orientation analysis based on sentence weighted algorithm [J] . Journal of Chinese Computer System (JCCS) ,2015,36(10) : 2252-2256.

附中文参考文献:

[7] 宫秀文,张佩云. 基于 PageRank 的社交网络影响最大化传播模型和算法研究 [J] . 计算机科学,2013,40(6) : 136-140.

[8] 毛佳昕,刘奕群,张　敏,马少平. 基于用户行为的微博用户社会影响力分析 [J] . 计算机学报,2014,37(4) : 1-10.

[9] 曹玖新,董　丹,徐　顺,等. 一种基于 k-核的社会网络影响最大化算法 [J] . 计算机学报,2015,38(2) : 238-248.

[10] 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.

[12] 李爱萍,邸　鹏,段利国. 基于句子情感加权算法的篇章情感分析 [J] . 小型微型计算机系统,2015,36(10) : 2252-2256.