·人工智能及识别技术· 文章编号：1000-3428(2018)11-0215-07 文献标志码：A 中图分类号：TP393

基于遍历约束与交互信息增强的社交网络表征算法

石立鹏[a]，王 莉[b]

（太原理工大学 a. 信息与计算机学院；b. 大数据学院，山西 晋中 030600）

摘 要：传统网络表征方法将网络的拓扑结构转换为邻接矩阵以进行特征提取，在准确率和效率上存在局限性。为此，通过分析自然语言处理模型 word2vec 和多种网络表征算法，结合社交网络的特征，提出一种基于遍历约束和交互信息的社交网络表征算法。对社交网络遍历进行约束以提高算法的时间效率，利用用户交互信息修改 word2vec 模型，提高社交网络表征的准确率。在 BlogCatalog 和新浪微博 2 个社交网络数据集上进行的实验结果表明，相对 DeepWalk、Line 算法，该算法在时间效率上提高约 20%，在准确率上提高约 12%。

关键词：特征学习；网络遍历；自然语言处理；交互信息；社交网络；网络表征

中文引用格式：石立鹏，王 莉. 基于遍历约束与交互信息增强的社交网络表征算法[J]. 计算机工程，2018，44(11)：215-221.

英文引用格式：SHI Lipeng，WANG Li. An enhanced social network representation algorithm based on traversal constraint and interactive information[J]. Computer Engineering，2018，44(11)：215-221.

An Enhanced Social Network Representation Algorithm Based on Traversal Constraint and Interactive Information

SHI Lipeng[a]，WANG Li[b]

（a. College of Information and Computer；b. College of Data Science，
Taiyuan University of Technology，Jinzhong，Shanxi 030600，China）

【Abstract】Traditional network representation methods transform network topology into adjacency matrix for feature extraction，which has limitations in accuracy and efficiency. Therefore，by analyzing the natural language processing model word2vec and a variety of network representation algorithms，combined with the characteristics of social networks，a social network representation algorithm based on traversal constraint and interactive information is proposed. Constraints are applied to the traversal of social networks to improve the time efficiency of the algorithm. word2vec model is modified by user interaction information to improve the accuracy of social network representation. Experimental results on two social network datasets，BlogCatalog and Sina Weibo，show that compared with DeepWalk and Line algorithm，this algorithm improves the time efficiency by about 20% and the accuracy by about 12%.

【Key words】feature learning；network traversal；natural language processing；interactive information；social network；network representation

DOI：10.19678/j.issn.1000-3428.0048755

0 概述

网络技术的迅猛发展，使社交网络成为覆盖用户最广、传播影响力最大的互联网发展产物之一，其蕴含着巨大的商业价值。

社交网络的主体是网络中的用户，也即网络中的节点。网络规模的不断壮大对社交网络表征算法提出了新的挑战。如何对社交网络进行准确、高效地表征，成为一个重要的研究方向。一个性能良好的表征算法便于计算用户间的相似度、实现用户间的链预测和用户社区划分等。传统社交网络表征算法多数通过人为特征定义、特征提取、矩阵运算以进行用户向量表示。但随着网络节点（社交网络用户）数量的增加，传统方法在准确率和效率上表现出一定的不足。

近年来，无监督算法在很多研究领域取得了很好的效果。在自然语言处理方面，word2vec 模型被提出。受该模型的启发，在网络学习中，出现了

基金项目：国家高技术研究发展计划(2014AA015204)；山西省自然科学基金(2014011022-1)。

作者简介：石立鹏(1990—)，男，硕士研究生，主研方向为社交网络、数据挖掘；王 莉，教授、博士。

收稿日期：2017-09-21 **修回日期：**2017-12-01 **E-mail：**270866242@qq.com

DeepWalk、Line、node2vec 及 ComEmbed[1] 网络表征算法。这些算法能够避免传统方法中人为提取特征和大量矩阵运算的问题,但都只利用网络的拓扑结构,并未涉及网络的交互信息。而用户间的交互行为恰是社交网络的重要组成部分,交互信息对网络表征具有十分积极的作用。

针对上述算法存在的不足,在现有网络表征算法和 word2vec 模型的基础上,本文依据社交网络特性,提出一种改进的基于遍历约束与交互信息的网络表征算法。该算法分析社交网络单个节点的特征,通过增加遍历规则来提高学习效率,利用网络用户间的交互信息改进自然语言模型 word2vec,以提高结果准确率。

1 相关工作

1.1 自然语言处理模型

文献[2]提出 word2vec 模型后,文献[3-4]对单词表征算法进行了改进,利用单个单词与其上下文的关系,将大量句子集合为语料库,通过对语料库进行训练并使用三层神经网络对词构造矢量表征,然后把单词映射到低维向量空间。虽然文献[5]指出单词的向量表征缺乏解释性,但该模型在自然语言处理中已经取得了很好的效果。

1.2 网络表征算法

网络表征问题一直受到许多研究者的关注,现有特征表示方法主要有:

1)传统的谱方法:文献[6-9]将网络转换成矩阵,分别利用邻接矩阵和拉普拉斯矩阵,通过矩阵运算、特征分解、降维等方法得到网络表征。但这种方法对较大型的网络并不适用,随着网络中节点和关系数目的增加,矩阵计算将耗费大量时间及计算机资源,且最终得到的向量表示也并不理想。

2)利用自然语言模型的方法:受自然语言处理模型的影响,近年来,很多研究者将网络与自然语言模型相结合,提出很多网络表征算法。文献[10]在 DeepWalk 中通过网络随机游走产生节点序列,类似自然语言中的句子,从而利用自然语言处理模型学习网络节点的特征表示。文献[11]将深度优先遍历策略和广度优先遍历策略相结合,对语料库进行优化。文献[12]提出 node2vec 模型,采用带有偏置的随机游走策略进一步提高网络表征的准确性。与传统的网络表征算法相比,这些算法利用机器学习的设计思想,在降低计算复杂度的同时能取得较好的网络表征效果。

在自然语言中,句子是满足特定语法规则的,单词间的关系是线性的,每个单词都存在特定的上下文。文献[13]认为,网络的结构是非线性的,通过不同的遍历方式产生的序列直接影响网络节点表征的结果。较普遍的遍历策略有深度优先、广度优先和随机遍历,不同遍历策略对结果会产生不同的影响,但目前仍然缺少普适性的遍历策略来提高网络表征的准确性。

1.3 社交网络表征

对于社交网络,文献[14-15]针对用户行为以及网络拓扑结构特征,对网络用户进行矢量化表示。文献[16]通过用户兴趣和拓扑结构实现社交网络的好友推荐。文献[17]基于社交网络拓扑结构,利用网络的连通性进行网络用户表征,最终实现好友推荐。文献[18]结合网络拓扑结构和用户信息对网络用户进行表征。这些方法都基于大量的矩阵运算,并不适合表征大型社交网络。

在利用自然语言模型对社交网络进行表征时,文献[19]将节点信息加入网络表征,可以减少矩阵运算,但其表征效果并没有得到明显提高。

现有表征算法只考虑网络本身,并未有效利用社交网络交互信息。而对于社交网络,其本质是一个用户交互平台,因此,网络交互信息对表征起着重要作用。本文根据社交网络邻居数量来约束网络遍历,同时利用网络交互信息对网络表征的学习过程进行优化。

2 改进的社交网络表征算法

在网络 $G=(V,E)$ 中,V 表示网络中的节点集合,E 表示网络中的边集合。网络表征的目的是将 V 中的各点映射到一个低维向量空间中,即存在这样一个映射 $f:V\rightarrow\mathbb{R}^d$,使 $v_i\in V$ 得到对应的 d 维向量表示。

2.1 网络表征学习模型

机器学习定义目标函数 f,利用合适的优化策略对函数进行优化,以得到较好的结果。对于网络表征算法,一般采用对数似然函数。设 $\boldsymbol{f}(w)$ 为节点 w 的向量表示,$Ns(w)$ 为遍历策略 s 下 w 的邻居。式(1)为网络表征的优化目标:通过遍历过程中的邻居节点以最大化网络节点 w 的对数概率。

$$L=\max\sum_{w\in V}\lg p(\boldsymbol{f}(w)\mid Ns(w)) \tag{1}$$

从 G 中各顶点开始,根据给定的遍历策略 s 生成语料库,将遍历序列作为自然语言中的“句子”,序列中各节点类比自然语言中的“单词”,根据节点出现频率构建 Huffman 树,G 中节点作为叶子,非叶子节点 θ 作为需要优化的参数。从根到叶子的路径为多次二分类,利用 Huffman 编码可以得到由 0 和 1 构成的序列 h,将该序列与路径上的二分类结果对应。对于节点 w,根据其当前遍历序列求邻居节点向量的和 $\boldsymbol{x}_w$。每次分类均对应一次逻辑回归,其编码为 1 或 0 的概率分别按照式(2)、

式(3)计算:

$$p_1 = \frac{1}{1 + e^{-x_w^T \theta}} \tag{2}$$

$$p_0 = 1 - \frac{1}{1 + e^{-x_w^T \theta}} \tag{3}$$

则从根到叶子节点 w 的路径选择概率对应多次二分类的概率乘积:

$$p(\boldsymbol{f}(w) \mid Ns(w)) = \prod_{j=1}^{l^w} p(h_j^w \mid \boldsymbol{x}_w, \boldsymbol{\theta}_j^w) \tag{4}$$

其中,h_j^w 表示节点 w 的 Huffman 编码中的第 j 位,l^w 表示 w 的 Huffman 编码长度,$\boldsymbol{\theta}_j^w$ 为该路径中非叶子节点的向量表示。最终的优化函数为:

$$L = \sum_{w \in V} \lg \prod_{j=1}^{l^w} \left\{ \left[\frac{1}{1 + e^{-x_w^T \boldsymbol{\theta}_j^w}} \right]^{h_j^w} \cdot \left[1 - \frac{1}{1 + e^{-x_w^T \boldsymbol{\theta}_j^w}} \right]^{1 - h_j^w} \right\} \tag{5}$$

根据遍历序列,采用随机梯度上升方法对参数 θ 及节点 w 的邻居进行优化,并更新各向量。

2.2　基于网络结构的遍历优化

在通常情况下,为获得足够大的语料库,可从网络中各节点起始,进行相同次数的遍历,产生庞大的语料库后进行训练。然而,通过对 BlogCatalog 和新浪微博数据集中节点度的统计,按照好友关系数可以得到一条递减的曲线。在曲线的头部,由于社交网络中名人和粉丝量巨大博主的存在,其好友关系数较多,随着博主影响力的降低,曲线会显著下降,尾部曲线会贴近于横轴,社交网络节点好友关系符合长尾分布。在采用相同遍历策略产生语料库的同时,会加入大量重复的句子,这不仅使遍历时间加长,增加训练时间成本,而且不能提高最终的表达效果,因此,根据节点分布约束遍历次数,更符合社交网络的特点。

如图 1 所示,与节点 x_3 关联的只有节点 v,在以节点 x_3 为起始节点的遍历序列中,前两跳总是固定的(x_3,v,…)。而在以节点 v 为起始节点的遍历序列中,遍历序列的第 2 跳可以看作是以 x_3 为起始节点的第 3 跳。根据节点特征的优化算法,以 x_3 为起始点的多次遍历显然是没有必要的。

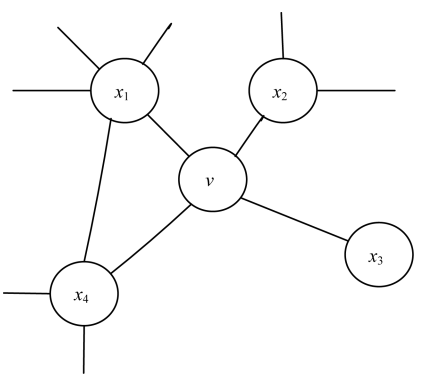


图 1　节点遍历约束示意图

按照长尾分布,将所有节点平均度数作为“头”和“尾”的分割。对于“头”中的节点,其好友关系复杂,根据算法给出的最大游走次数遍历生成语料库,对于“尾”中的节点,根据节点与均值的比值约束游走次数。各节点游走次数计算如下:

$$count(v_i) = \begin{cases} \dfrac{deg(v_i)}{avg} \times walknum, deg(v_i) < avg \\ walknum, deg(v_i) > avg \end{cases} \tag{6}$$

其中,v_i 代表网络中的第 i 个节点,$deg(v_i)$ 表示节点 v_i 的度数,$walknum$ 为最大游走次数,avg 为所有节点的平均度数,$count(v_i)$ 计算节点 v_i 的遍历次数,并作为遍历约束条件。

本文算法只限制以该节点为起点的随机游走次数,该节点仍然会在其他节点的随机游走序列中出现。根据 word2vec 原理,对于度数较低的点,其向量表示同样可以根据包含该节点的其他游走序列进行优化。

2.3　基于交互行为的算法改进

网络表征算法及 word2vec 模型的目的是使相似节点或单词具有更加接近的特征表示。在 word2vec 模型中,单词初始化向量是随机的,向量表示的距离也是随机的,但通过大量的语料训练可使具有相似含义的单词更加接近。这是因为自然语言中的句子是线性的,可以通过大量的训练得到较准确的表达。现有网络表征算法大都聚焦于调整网络遍历策略,然后直接利用 word2vec 模型得到节点表征。由于网络的非线性特性,当网络表征准确率趋于稳定后,通过扩充训练集来提高算法表征效果将变得十分困难。

社交网络作为用户交互平台,存在大量的用户交互信息,这些信息对社交网络表征具有积极的意义。通过对社交网络的分析发现,用户在一段时间内存在交互行为的好友,极有可能属于同一个或有限几个社交圈。因此,可以根据社交网络的这一特点对 word2vec 模型进行改进,虽然这无法完全避免由网络非线性结构带来的困扰,但却可以使模型更适合社交平台。

在 word2vec 中,各维向量缺乏解释性[5],虽然向量中的某些维度可能反映着不同信息,比如在词向量中,某些维度可能包含了性别这类信息,但现有模型无法准确指出每一维向量的具体意义,这也使得该模型在自然语言处理中存在一定的缺陷。然而,对网络表征而言,可以利用这部分信息与网络交互集合来修改 word2vec 模型的初始化阶段,使不同集合间的差异性更突出,然后通过训练使整个网络表征更准确。

2.3.1 交互集合选择

社交网络用户的交互行为很频繁,多数用户都会存在数量不等的交互信息。为避免平均化,应选择交互量较大用户的交互好友作为单个优化集合,选择交集较小的不同用户的交互集合作为优化对象。

交互集合作为算法优化的依据,需满足如下条件:1)应选择交互关系数较大的多个用户的交互集合,以减少集合个数,同时避免优化结果的平均化;2)单个交互集合内元素与同一用户存在交互行为;3)不同交互集合的交集尽可能小,避免对同一用户进行多次优化从而降低区分度。

2.3.2 表征算法优化

利用交互集合对社交网络表征算法进行优化,具体过程如下:

1)减小同一集合中的元素距离,提高元素间的相似度。在使用 word2vec 对向量初始化后,根据欧式距离求集合内各节点 u_i 的向量中心 *centre*,取集合中心与原节点来计算节点新的初始化结果:

$$\boldsymbol{f}(u_i)=\frac{\boldsymbol{f}(u_i)+centre}{2} \tag{7}$$

其中,$\boldsymbol{f}(u_i)$ 为用户 u_i 的向量表示,*centre* 表示用户 u_i 所在交互集合的中心。

2)增加不同集合间的距离,提高类别间的辨识度。随机对同一集合所有节点表征中的 m 维向量进行优化,使其在空间上属于另一个区间。随机产生小于节点向量维度的 m 个随机数 $ran=\{r_1,r_2,\cdots,r_m\}$。根据集合 *ran* 对向量进行优化:

$$[v_1,v_2,\cdots,v_i,\cdots,v_d]\times[sp_{ij}]_{d\times d} \tag{8}$$

其中,$[sp_{ij}]_{d\times d}$ 为对角矩阵,对角元素为初始化向量中每一维向量的跨度,$v_i\in[0,g]$,$g\in\mathbb{N}^+$,$\mathbb{N}^+$ 表示自然数,g 表示向量分量在更改过程中的范围,其在一定程度上反映类别辨识度的强弱。当 $i=r_j$ 时,$v_i=g$,其余为 0。根据式(7)、式(8)得同一集合的向量初始化结果为:

$$\boldsymbol{f}(u_i)=\frac{\boldsymbol{f}(u_i)+centre}{2}+[v_1,v_2,\cdots,v_i,\cdots,v_d]\times[sp_{ij}]_{d\times d} \tag{9}$$

利用用户的交互信息,优化 word2vec 的初始化,使不同集合内的用户与其他集合在特征表示上有明显差异,在不断的学习中,使与其相似的节点也表现出同样的特性。如果在选择的 n 个优化集合中出现同一类节点,在相同的优化策略下,其效果等同于选择 $n\times m$ 维对某些节点进行优化。当 m 较小时并不影响整体学习效果。

改进后的社交网络表征算法具体步骤为:

步骤 1 根据网络节点分布,计算各节点游走次数,生成相应集合,按照节点划分约束随机游走过程,生成语料库。

步骤 2 初始化节点向量并根据交互关系对向量进行修改。

步骤 3 训练得到各节点的向量表示。

对整个算法而言,随机游走和根据游走序列训练节点向量表示是消耗时间最多的步骤,引入节点划分策略,使不同节点的游走次数得到不同的控制,能让总游走次数降低、语料库减小,训练时间随之减少,同时利用交互集合可以提高最终表征的准确度。

2.4 算法描述

本文基于遍历约束和交互信息的社交网络表征算法描述如下:

算法 1 社交网络表征算法

输入 网络拓扑 $G=(V,E,W)$,向量维度 d,最大游走次数 r,游走长度 l,窗口大小 k,交互关系 *follows*,改变的维度 m,改变的大小 g

输出 d 维的节点向量表征

```
1. function learnFeatures(G,d,r,l,k,follows,m,g)
2. walks = restrictedWalk(G,l,r)//根据网络拓扑及节点
//度数遍历网络,构建网络遍历集合 walks
3. initvec = initvec(k,d,walks,follows,m,g)//根据遍
//历结果及交互信息对节点表征进行初始化
4. vec = train(walks,initvec)//利用自然语言模型对遍
//历结果进行训练,得到最终表征结果
5. return vec
6. function restrictedWalk(G,l,r)//构建网络遍历集合
7. removeSet = buildDifSetAccordNode(G)//根据网
//络节点度数设定不同的遍历次数
8. for iter = 1 to r do
9. for nodes u∈V do
10. walks = buildWalkAccording(G)//遍历节点构建
//walks
11. end for
12. V remove nodes which in removeSet[iter]//移除遍历
//次数达到阈值的节点
13. end for
14. return walks
15. function initvec(k,d,walks,follows,m,g)//对节点表
//征进行初始化
16. initialize_vec(walks)
17. for perFollows in follows
18. for node in perFollows
19. initvec(node) = (centre + vec(node))/2 + [vi]1×d
×[spanij]d×d//根据交互集合对极有可能属于同一集合的
//节点表征进行向量优化
20. end for
21. end for
22. return initvec
```

以上伪代码对算法流程进行了简要表述:

1)函数 *restrictedWalk*( )根据网络节点分布建立次数不等的遍历集合,在之后的遍历过程中不断删除遍历次数达到指定阈值的节点集合。

2)函数 *stochasticGradientDescent*( )首先利用 word2vec 对各节点进行初始化,然后根据交互集合对初始化后的节点实现优化,最后通过大量训练集来表征学习整个网络节点。

3　实验结果与分析

3.1　实验设计

为验证本文算法的效率和准确率,采用 BlogCatalog 和新浪微博数据集进行实验。对相同的实验任务,将本文算法与以下 4 种主流网络表征算法进行比较:

1)DeepWalk 算法:根据网络拓扑结构,采用相同次数的随机游走方式遍历网络,生成语料库,直接使用 word2vec 模型进行向量表征。

2)node2vec 算法:与 DeepWalk 相比,该算法通过参数调节达到带有偏置的随机游走,游走可能趋于深度优先或广度优先,同样直接使用 word2vec 模型进行向量表征。

3)Line 算法:将深度优先和广度优先策略结合,利用 word2vec 进行表征。

4)ComEmbed 算法:对社区和节点共同利用 word2vec 进行优化。

3.2　数据集及实验环境

实验所用数据集信息如下:

1)BlogCatalog:社交网络数据集,数据从 BlogCatalog 网站爬取,多数主流表征算法采用该数据集进行对比实验。该数据集包含 10 312 个节点、333 983 条边和交互信息,节点表示不同的微博用户,边代表 2 个微博用户之间存在着好友关系。数据集将 10 312 个用户分成 39 类。

2)新浪微博:该数据集为爬取的新浪微博部分数据,包含 1 701 个节点(微博用户)、29 439 条边(好友关系)、90 962 条交互行为,数据集将用户分为 8 类。

本次实验在个人计算机上进行,实验环境如下:处理器为 Intel(R) Core(TM) i5-2450M CPU 2.5 GHz双核;内存为 8 GB;操作系统为 Windows10(64 位)。

实验 1　分别利用 DeepWalk 算法、node2vec 算法、Line 算法、ComEmbed 算法和本文算法在 2 个数据集上进行实验。实验结果采用准确率作为衡量指标,同时验证算法的时间效率。

为避免由参数设置带来的误差,本实验中节点最大遍历次数与其他算法遍历次数相同,结果使用相同的分类算法进行对比验证。实验参数设置为:向量维度 $d=128$,游走步长 *walklength* = 80,窗口大小 *winsize* = 10,游走次数 *walknum* = 40。实验结果如表 1、表 2 所示。

表 1　实验 1 中各算法准确率结果

算法	BlogCatalog 数据集	新浪微博数据集
DeepWalk 算法	0.283	0.327
Line 算法	0.274	0.297
node2vec 算法	0.302	0.316
ComEmbed 算法	0.313	0.314
本文算法	0.421	0.452

表 2　实验 1 中各算法运行时间结果　　min

算法	BlogCatalog 数据集	新浪微博数据集
DeepWalk 算法	521	91
Line 算法	526	91
node2vec 算法	536	93
ComEmbed 算法	553	96
本文算法	415	71

5 种算法对网络节点进行维数为 128 的向量表征,采用 50% 带标签的节点作为训练集,剩余部分作为测试集。从表 1 可以看出,本文算法在准确率上优于流行的网络表征算法,在 BlogCatalog 数据集上达到 0.421,比 node2vec 和 ComEmbed 分别提高 39% 和 35% 。在新浪微博数据集中,本文算法相对其他算法,在准确率上也有明显提高。从表 2 可以看出,对比其他算法,本文算法运行时间下降均高于 20% 。

实验 2　在 BlogCatalog 数据集上,分别使用准确率较高的 node2vec 算法和本文算法,研究游走次数 *walknum* 对算法性能的影响。实验参数设置为:向量维度 $d=128$,游走步长 *walklength* = 80,窗口大小 *winsize* = 10,游走次数 *walknum* 不断增加。采用准确率和F1 值作为衡量指标,实验结果如表 3、表 4 所示。

表 3　实验 2 中 2 种算法准确率结果

walknum 值	准确率	
	node2vec 算法	本文算法
10	0.221	0.253
20	0.243	0.286
30	0.276	0.361
40	0.302	0.421
50	0.303	0.424

表 4　实验 2 中 2 种算法 F1 值结果

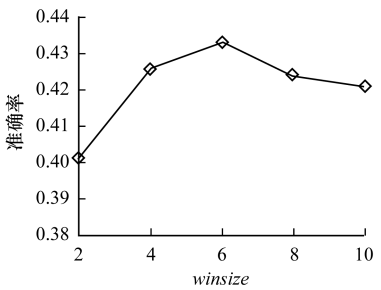
walknum 值	F1 值	
	node2vec 算法	本文算法
10	0.231	0.250
20	0.244	0.272
30	0.270	0.344
40	0.304	0.409
50	0.309	0.410

从表 3、表 4 可以看出,算法的准确率和 F1 值均随着游走次数 *walknum* 的增加而提高,当 *walknum* 达到 40 后算法结果趋于平稳。在准确率上,本文算法在 *walknum* = 30 时,其结果已经比 *walknum* = 40 时 node2vec 的结果高出 20%,当 *walknum* = 40 时,本文算法 F1 值比 node2vec 算法提高了 35%。

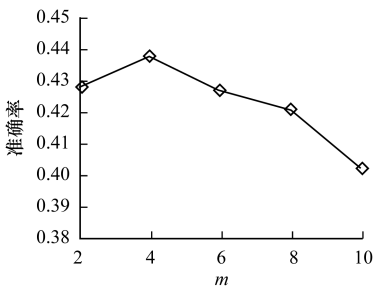
3.3 参数分析

本文算法包含若干参数,其中多数为目前网络表征中都会使用的参数(如游走步长 *walklength*、向量维度 *d*、游走次数 *walknum*、窗口大小 *winsize*),除此之外,本文算法还引入参数 *m*、*g*,*m* 为随机优化向量维数,*g* 为向量分量的大小范围。

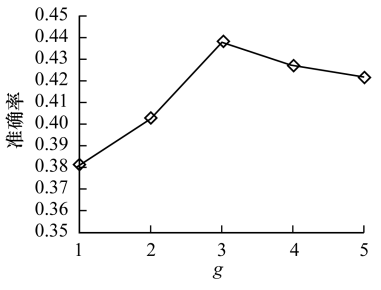
在已有的网络表征算法中,已经说明游走步长 *walklength* 对模型的影响。本次实验主要研究窗口大小 *winsize*、*m*、*g* 对算法性能的影响,实验结果如图 2所示。



(a)*winsize*对算法准确率的影响



(b)*m*对算法准确率的影响



(c)*g*对算法准确率的影响

图 2 各参数对算法性能的影响

从图 2 可以看出:

1)*winsize* 的设置和传统自然语言的表征算法有较大区别,当 *winsize* ∈ [4,8] 时,算法效果较好,当 *winsize* 超过 10 以后,算法效果会趋于稳定。造成该结果的原因可能是,自然语言处理在一定程度上符合大数定理,而社交网络的处理离不开“六度分离”理论。

2)*m* 的选择受向量维度 *d* 的影响,已有算法证明当 *d* 达到 100 之后算法表征效果趋于稳定。本文选取 128 作为向量表征维度,研究 *m* 对算法性能的影响。可以看出,维数的多少对算法有明显的影响,当维数较大时,算法的准确性会降低,原因是选择维数太大不仅不会增加类之间的区分度,还会使所有维度的修改趋于平均化。

3)分析向量改变大小 *g* 对算法性能的影响,通过实验结果可以看出,当 *g* = 3 时算法效果最好,当 *g* 超过 5 以后算法效果会趋于稳定。

4 结束语

本文分析社交网络中节点自身好友关系的数量,提出一种改进的社交网络表征算法。根据好友分布不均衡的特性控制网络遍历次数,指导随机游走后生成较小的语料库,同时根据交互关系优化节点向量。实验结果表明,该算法在准确率和效率上具有优势。社交网络包含丰富的信息,如何充分利用这些信息构造更准确的节点表征模型,将是下一步的研究方向。

参考文献

[1] ZHENG V W, CAVALLARI S, CAI H, et al. From node embedding to community embedding[EB/OL]. [2017-09-10]. https://www.researchgate.net/publication/309572690_From_Node_Embedding_To_Community_Embedding.

[2] MIKOLOVB T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2017-09-10]. http://ling.snu.ac.kr/class/AI_Agent/lecture/07-3-EfficientEstimationofWordRepresentationinVectorSpace.pdf.

[3] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distri-buted representations of words and phrases and their compositionality[C]//Proceedings of International Conference on Neural Information Processing Systems. [S.l.]: Curran Associates Inc., 2013: 3111-3119.

[4] MIKOLOV T, YIH W T, ZWEIG G. Linguistic regularities in continuous space word representations[C]//Proceedings of 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S. l.]: Association for Computational Linguistics, 2013: 746-751.

[5] SUN F, GUO J, LAN Y, et al. Sparse word embeddings using l1 regularized online learning[C]//Proceedings of International Joint Conference on Artificial Intelligence. [S.l.]: AAAI Press, 2016: 2915-2921.

[6] TANG L, LIU H. Leveraging social media networks for

classification[J]. Data Mining and Knowledge Discovery, 2011,23(3):447-478.
[7] YAN S,XU D,ZHANG B,et al. Graph embedding and extensions: a general framework for dimensionality reduction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2007,29(1):40-48.
[8] BELKIN M, NIYOGI P. Laplacian eigenmaps and spectral techniques for embedding and clustering[J]. Advances in Neural Information Processing Systems, 2002,14(6):585-591.
[9] GALLAGHER B, ELIASSIRAD T. Leveraging label-independent features for classification in sparsely labeled networks: an empirical study [C]//Proceedings of International Conference on Advances in Social Network Mining and Analysis. Berlin,Germany:Springer,2008:1-19.
[10] PEROZZI B, ALRFOU R, SKIENA S. DeepWalk: online learning of social representations[EB/OL]. [2017-09-10]. http://www. perozzi. net/publications/14_kdd_deepwalk-slides. pdf.
[11] TANG J, QU M, WANG M, et al. LINE: large-scale information network embedding [C]//Proceedings of International World Wide Web Conferences Steering Committee. New York, USA: ACM Press, 2015: 1067-1077.
[12] GROVER A, LESKOVEC J. node2vec: scalable feature learning for networks[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York,USA:ACM Press,2016:855-864.
[13] WANG D, CUI P, ZHU W. Structural deep network embedding[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York,USA:ACM Press,2016:1225-1234.
[14] 蔡波斯,陈　翔. 基于行为相似度的微博社区发现研究[J]. 计算机工程,2013,39(8):55-59.
[15] CAO S S, LU W, XU Q K. GreRep: learning graph representations with global structural information[C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management. New York, USA:ACM Press,2015:891-906.
[16] 何　静,潘善亮,韩　露. 基于双边兴趣的社交网好友推荐方法研究[J]. 计算机工程与应用,2015,51(6):108-113.
[17] 周芝民,龙　华,杜庆志,等. 基于连通性和随机游走的好友推荐算法[J]. 信息技术,2016(8):67-70.
[18] 张中军,张文娟,于来行,等. 基于网络距离和内容相似度的微博社交网络社区划分方法[J]. 山东大学学报(理学版),2017,52(7):97-103.
[19] YANG C,ZHAO D,ZHAO D,et al. Network representation learning with rich text information [C]//Proceedings of International Conference on Artificial Intelligence. [S.l.]:AAAI Press,2015:2111-2117.

编辑　吴云芳

(上接第 214 页)

[7] BO P,LEE L,VAITHYANATHAN S. Thumbs up?:sentiment classification using machine learning techniques [C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics,2002:79-86.
[8] BADER B W, KEGELMEYER W P, CHEW P A. Multilingual sentiment analysis using latent semantic indexing and machine learning [C]//Proceedings of IEEE International Conference on Data Mining Workshops. Washington D. C., USA: IEEE Computer Society,2011:45-52.
[9] LIN C,HE Y. Joint sentiment/topic model for sentiment analysis [C]//Proceedings of ACM Conference on Information and Knowledge Management. New York, USA:ACM Press,2009:375-384.
[10] WANG S, MANNING C D. Baselines and bigrams: simple, good sentiment and topic classification [C]//Proceedings of Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics,2012:90-94.
[11] BENGIO Y,SCHWENK H,SENÉCAL J S,et al. Neural probabilistic language models [M]//HOLMES D E, JAIN L C. Innovations in Machine Learning. Berlin, Germany:Springer,2006:1137-1155.
[12] LE Q V, MIKOLOV T. Distributed representations of sentences and documents [EB/OL]. (2014-05-22) [2017-05-24]. https://arxiv. org/abs/1405. 4053.
[13] 王盛玉,曾碧卿,胡翩翩. 基于卷积神经网络参数优化的中文情感分析[J]. 计算机工程,2017,43(8):200-207,214.
[14] 梁　斌,刘　全,徐　进,等. 基于多注意力卷积神经网络的特定目标情感分析[J]. 计算机研究与发展,2017,54(8):1724-1735.
[15] TRASK A, MICHALAK P, LIU J. Sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings[EB/OL]. (2015-12-19) [2017-05-24]. https://arxiv. org/abs/1511. 06388.
[16] BO P,LEE L. Seeing stars:exploiting class relationships for sentiment categorization with respect to rating scales[C]//Proceedings of the Meeting of Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2005: 115-124.

编辑　金胡考