# Deepwalk: Online Learning of Social Representations

**摘要：**

我们提出 DeepWalk，一种用于学习网络中顶点的潜在表示的新方法。这些潜在表示将社会关系编码到连续的向量空间中，编码到向量空间后的社会关系，很容易应用到统计模型中。 DeepWalk 将语言建模和无监督特征学习（或深度学习）的最新进展，从单词序列推广到图中。

DeepWalk 将随机游走得到的节点序列当做句子，从截断的随机游走序列中得到网络的局部信息，再通过局部信息来学习节点的潜在表示。为了展示 DeepWalk 得到的节点的潜在表示，我们对几个社交网络（BlogCatalog，Flickr 和 YouTube）进行了多标签分类任务。我们的研究结果显示，DeepWalk 能够对网络进行全局的观察，特别是在存在缺失信息的情况下。当已标记数据很少时，DeepWalk 的表示得到的 F1 分数比对比方法高出 10％。在一些实验中，当训练数据少于 60%时，DeepWalk 的表现能够胜过所有对比算法。

DeepWalk 也是可扩展的。DeepWalk 是可以建立有用的增量结果的在线学习算法，并且是平行的。这些特性使其适用于广泛的实际应用，如网络分类和异常检测。

**背景：**

网络表示法的稀疏性既是一种优势，也是一种弱点。稀疏性使设计高效的离散算法成为可能，但也使统计学习中的推广变得困难。

本文首次将自然语言处理成功的深度学习(无监督特征学习)技术引入到网络分析中。我们创造了一种算法(DeepWalk)，它通过对短随机游动流的建模来学习图的顶点的一般表示。社会表征是捕捉顶点的潜在特征。 e 邻域相似性和社区成员关系。这些潜在的表示在一个相对较少维数的连续向量空间中编码社会关系。DeepWalk 推广神经语言模型处理一种特殊的语言，由一组随机生成的游走组成.这些神经语言模型已经被用来捕捉人类语言的语义和句法结构甚至逻辑对比。

**问题定义：**

社交网络中节点之间的连接比较稀疏，是比较典型的信息比较少的网络，文章通过对社交网络中的节点进行分类这一问题，说明了问题定义。

图的表示：令 G =(V,E)，其中 V 表示网络的节点，E 是网络中的连接，E⊆(V×V)。

GL=(V,E,X,Y)是部分标记的社交网络。X 是各个节点的属性空间，$X\in R^{|V|\times S}$，其中 S 是每个节点的属性向量的特征空间的大小；$Y\in R^{|V|\times|Y|}$，Y 是标签的集合。

传统机器学习分类问题中，目标是学习一个假设 H，能够将 X 中的向量映射到 Y 中的标签（不考虑图的结构，根据每个节点的属性向量，将节点分类）。因为网络中的信息比较少，所以该方法不适用与网络。

在本文中，采用的方法是：先得到图的结构，然后利用图的结构实现分类，这种分类方式被称为关系分类,也叫做集体分类。

本文提出了新的、无监督的、独立于标签分布的（捕获结构信息时不考虑标

签）、捕获图结构信息的算法。算法目标是学习图的结构特征 $XE \in R^{|V| \times d}$，其中 d 是节点的潜在表示（向量形式）的维数。图结构特征可以用于任何分类算法。

将 $XE \in R^{|V| \times d}$ 与简单的机器学习算法集成，还可以用来实现很多其他问题。

**对算法的要求：**

适应性：社交网络是不断变化的，当网络发生变化时，不能对整个网络重新进行计算。

社区意识：节点在潜在表示的维度空间中的距离，应该表示网络中对应的成员的相似度，以此保证网络的同质性。

低维：当被标记的成员很少时，低维的模型一般表现的更好，并且收敛和推理速度更快。

连续性：需要通过图的潜在表示来对连续空间中的部分社区成员进行建模。除了提供对社区成员资格的细微视图之外，连续表示还可以使社区之间的决策界限平滑，从而实现更强大的分类。

为了满足这些要求，算法利用最初为了语言建模设计的优化技术(word2vec)，从短的随机游走序列中学习节点的表示，

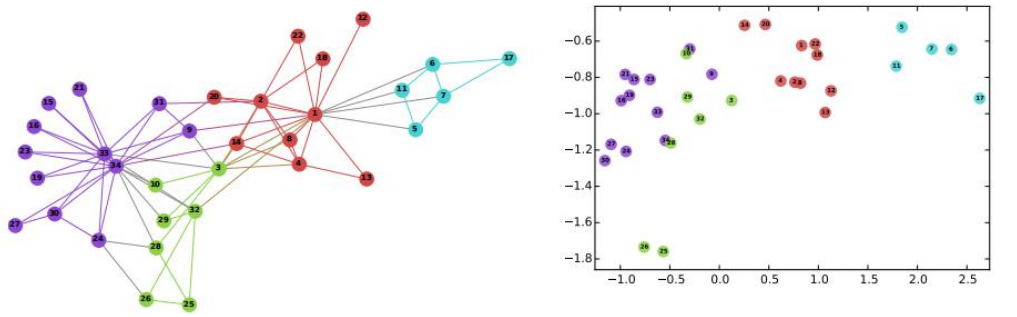
**算法思路：**

随机游走:

将从顶点 vi 开始的随机游走序列表示为 Wvi。Wvij 表示序列 Wvi 中的第 j 个点。其中，$Wvi^1$ 为 vi，$Wvi^{k+1}$ 是从 $Wvi^k$ 的邻居中随机选择的节点。

输入输出：

DeepWalk 将一个图作为输入，并产生一个潜在表示（将图中的每个节点表示为一个向量）作为输出。图 1 是将算法应用到空手道网络的结果：



(a) Input: Karate Graph　　(b) Output: Representation

**语言建模:**

1. 语言建模的目标是估计出现在语料库中的特定序列的可能性。即，给定 $W^n =（w_0，w_1，...，w_n）$的序列，其中 $w_i \in V$（V 是词汇表），我们想最大化 $Pr（w_n | w_0，w_1，...，w_{n-1}）$。在最近的工作中，语言建模扩展到使用概率神经网络来构建词语的一般表示。
   随机游走得到的序列可以被认为是一种特殊语言的短句，类比语言建模可以得到：在随机游走中给定迄今为止访问的所有先前顶点的情况下，下一

个顶点是 vi 的可能性可以表示为：

$$Pr(v_i|(v_1,v_2,\cdots,v_{i-1})) \quad (1)$$

2. 为了得到节点的潜在表示，引入映射函数Φ：$v \in V \rightarrow R^{|V|\times d}$，|V|×d 的矩阵Φ表示图中每个顶点的在 d 维空间中的潜在表示。这样公式(1)可以表示为：

$$Pr(v_i|\Phi(v_1),\Phi(v_2),\cdots,\Phi(v_{i-1})) \quad (2)$$

然而，随着步数的增长，不可能进行计算。

3. 对语言建模进行下面的 relaxation：

1)不是通过上下文预测单词，而是使用单词来预测上下文。

2)上下文由给定的单词左右两边的单词组成（原本只考虑左边的）。

3)不考虑句子中上下文出现的顺序，最大化出现在上下文中的所有单词的概率。

对于顶点表示建模，就产生了下面的优化问题：

$$\text{minimize} \quad -\log Pr(\{v_{i-w},\cdots,v_{i+w}\}\backslash v_i \mid \Phi(v_i)) \quad (3)$$

通过解决优化问题(3)就可以得到图中节点的向量表示形式。具有相同的上下文的节点的表示相似（使用 LINE 中的定义，即算法使用的是二阶相似度）。

**算法伪代码：**

DeepWalk：

**Algorithm 1** DEEPWALK($G$, $w$, $d$, $\gamma$, $t$)

**Input:** graph $G(V, E)$
window size $w$
embedding size $d$
walks per vertex $\gamma$
walk length $t$

**Output:** matrix of vertex representations $\Phi \in \mathbb{R}^{|V|\times d}$

1: Initialization: Sample $\Phi$ from $\mathcal{U}^{|V|\times d}$
2: Build a binary Tree $T$ from $V$
3: **for** $i = 0$ to $\gamma$ **do**
4: $\mathcal{O} = \text{Shuffle}(V)$
5: **for each** $v_i \in \mathcal{O}$ **do**
6: $\mathcal{W}_{v_i} = RandomWalk(G, v_i, t)$
7: SkipGram($\Phi$, $\mathcal{W}_{v_i}$, $w$)
8: **end for**
9: **end for**

SkipGram:

**Algorithm 2** SkipGram($\Phi$, $\mathcal{W}_{v_i}$, $w$)

1: **for each** $v_j \in \mathcal{W}_{v_i}$ **do**
2: **for each** $u_k \in \mathcal{W}_{v_i}[j - w : j + w]$ **do**
3: $J(\Phi) = -\log \Pr(u_k \mid \Phi(v_j))$
4: $\Phi = \Phi - \alpha * \frac{\partial J}{\partial \Phi}$
5: **end for**
6: **end for**