社交网络中的网络表示学习技术

介绍

传统的网络表示一般使用高维的稀疏向量。但是高维稀疏的表示也成为了人们使用统计学习方法时的局限所在，因为高维的向量将会花费更多的运行时间和计算空间。随着表示学习技术的发展，研究者转而探索将网络中的节点表示为低维稠密的向量表示方法，有效融合网络结构与节点外部信息，形成更具区分性的网络表示成为挑战。

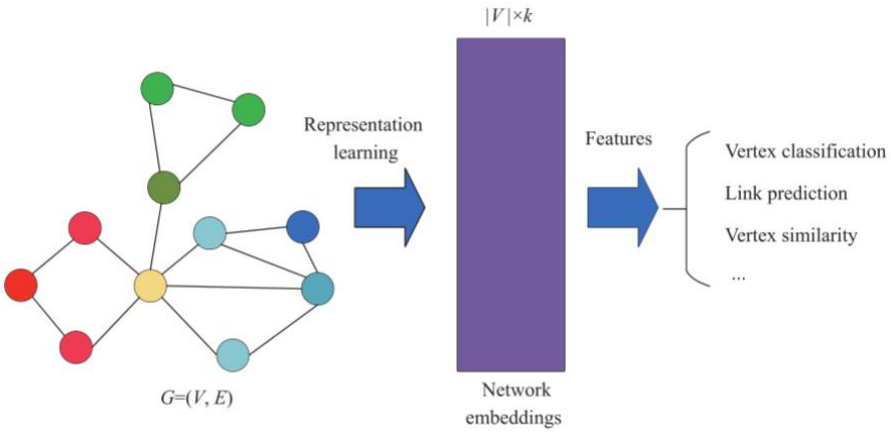


图 1 网络表示学习流程图

主要应用场景包括

节点分类

在进行网络数据的分析时，一个最常见的场景就是对网络中的节点进行合理的划分。在社交网络分析的应用中，不同的用户可以根据他们的兴趣爱好不同而分为不同的类别。
但是，实际数据中具有类别标签的信息是十分稀少的，所以需要设计算法利用节点间的连接关系以及少量的已标注分类信息，对大量的为标注节点的分类情况进行标注。类似的任务场景还有商品推荐，运用节点分类把用户进行分类，根据算法预测对未进行购买行为的用户推荐其感兴趣的产品。

连接预测

连接预测旨在预测网络中丢失的边，或者未来可能会出现的边。在进行链接预测时，需要对所有不在训练数据中的点对打分。在表示学习中，一般使用一对节点表示的内积或余弦相似度来计算得分。我们一般用 AUC 值来评价链接预测任务的结果。AUC 值代表了一条未观测到的点对的得分比一条不存在的点对得分高的概率。

社区发现

社区发现问题旨在对网络中的节点进行无监督的聚类，从而将网络中相似的节点归为同一个社区。与节点分类任务相比，社区发现最主要的不同就是社区发现任务是无监督的，即没有任何已标定的数据。在实际应用层面上，社区发现算法可以用来为社交网络中的用户自动划分好友的分组。

定义

网络表示学习是衔接网络原始数据和应用任务的桥梁。网络表示学习算法负责从网络数据中学习得到网络中每个节点的向量表示，之后这些节点表示就可以作为节点的特征应用于后续的网络应用任务，如节点分类、链接预测和社区发现等。

网络表示学习的目标就是对每个节点学习一个实数向量，其中向量的维度 k 远小于节点的数量。网络表示学习的过程可以是无监督或者半监督的。通过优化算法自动得到而不需要特征工程的节点表示可以进一步用于后续的网络应用任务而不必再去考虑原本的网络结构。

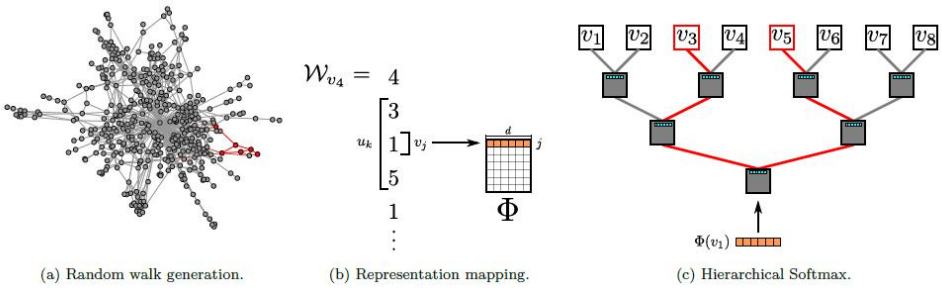
主要算法

1、DeepWalk

DeepWalk 算法第一次将深度学习中的技术引入到网络表示学习领域。DeepWalk 算法充分利用了网络结构中的随机游走序列信息。使用随机游走序列的信息有两点好处：

（1）随机游走序列只依赖于局部信息，所以可适用于分布式和在线系统，而使用邻接矩阵就必须把所有信息存储于内存中处理，面临着较高的计算时间和空间消耗。

（2）对随机游走序列进行建模可以降低建模 0-1 二值邻接矩阵的方差和不确定性。



(a) Random walk generation. (b) Representation mapping. (c) Hierarchical Softmax.

图 2 DeepWalk 算法流程图

2、LINE

LINE 也是一种基于邻域相似假设的方法，只不过与 DeepWalk 使用 DFS 构造邻域不同的是，LINE 可以看作是一种使用 BFS 构造邻域的算法。

LINE 定义了新的相似度度量方法：

一阶相似度：

一阶相似度用于描述图中成对顶点之间的局部相似度，形式化描述为若 u，v 之间存在直连边，则边权w_{uv}即为两个顶点的相似度，若不存在直连边，则一阶相似度为 0。

如下图，6 和 7 之间存在直连边，且边权较大，则认为两者相似且 1 阶相似度较高，而 5 和 6 之间不存在直连边，则两者间 1 阶相似度为 0。

二阶相似度：

如下图，虽然 5 和 6 之间不存在直连边，但是他们有很多相同的邻居顶点(1,2,3,4)，这其实也可以表明 5 和 6 是相似的，而 2 阶相似度就是用来描述这种关系的。

形式化定义为令$p_u = (w_{u,1},\cdots,w_{u,|V|})$表示顶点 u 与所有其他顶点间的一阶相似度，则 u 与 v 的二阶相似度可以通过p_u和p_v的相似度表示。若 u 与 v 之间不存在相同的邻居顶点，则二阶相似度为 0。

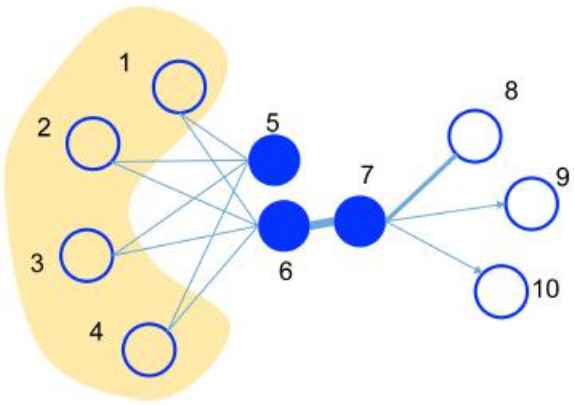


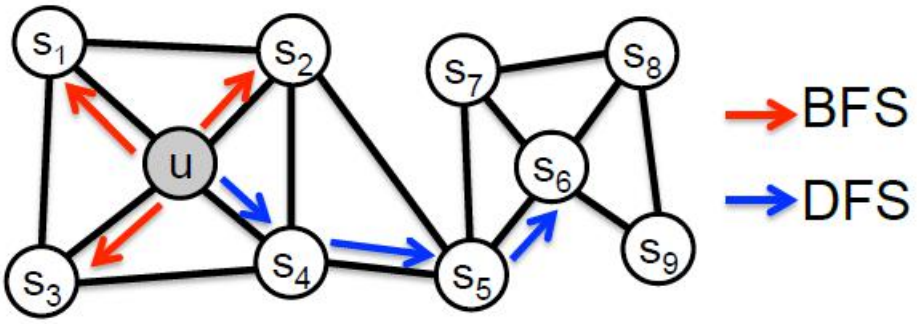
图 3 一个简单的网络结构图

该算法的优化目标结合一阶相似度和二阶相似度，采用分别训练一阶相似度模型和二阶相似度模型，然后将学习的两个向量表示连接成一个更长的向量。更适合的方法是共同训练一阶相似度和二阶相似度的目标函数，比较复杂，文章中没有实现。

3、Node2vec

node2vec 通过改变随机游走序列生成的方式进一步扩展了 DeepWalk 算法。DeepWalk 选取随机游走序列中下一个节点的方式是均匀随机分布的，而 node2vec 通过引入两个参数 p 和 q，将宽度优先搜索和深度优先搜索引入随机游走序列的生成过程。

宽度优先搜索(BFS)注重临近的节点，并刻画了相对局部的一种网络表示，宽度优先中的节点一般会出现很多次，从而降低刻画中心节点的邻居节点的方差；深度优先搜索(BFS)反应了更高层面上的节点间的同质性。如下图所示：



node2vec 中的两个参数 p 和 q 控制随机游走序列的跳转概率，如下图 所示：

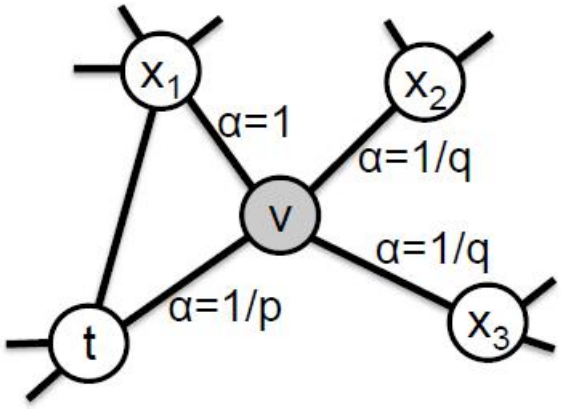


图 5 Node2vec 节点跳转概率

当下一个节点 x 与前一个节点 t 和当前节点 v 等距时，α =1；当下一个节点 x 是上一个节点时，α =$\frac{1}{p}$；在其他情况下，α =$\frac{1}{q}$。假设上一步游走的边为（t,v），那么对于节点 v 的不同邻居，node2vec 根据 p 和 q 定义了不同的邻居的跳转概率，p 控制跳向上一个节点的邻居的概率，q 控制跳向上一个节点的非邻居的概率，具体的未归一的跳转概率值 $\pi_{vx} = \alpha_{pq}(t,x)$，如下所示：

$$\alpha_{pq}(t,x) = \begin{cases} \frac{1}{p} & \text{if } d_{tx} = 0 \\ 1 & \text{if } d_{tx} = 1 \\ \frac{1}{q} & \text{if } d_{tx} = 2 \end{cases}$$

其中，d_{tx} 表示节点 t 和 x 之间的最短距离。为了获得最优的超参数 p 和 q 的取值，node2vec 通过半监督形式，利用网格搜索最合适的参数学习节点表示。

4、SDNE

SDNE 使用一个自动编码器结构来同时优化一阶和二阶相似度，而 LINE 是分别优化的，学习得到的向量表示能够保留局部和全局结构，并且对稀疏网络具有鲁棒性。

为了捕捉高度非线性的网络结构，提出了一种深度架构，它由多个非线性映射函数组成，通过将输入数据映射到高度非线性的潜在空间以捕获网络结构。通过重构每个节点的邻域结构来设计无监督学习部分来保持二阶相似性。设计监督学习部分来利用一阶相似性作为监督信息来改进潜在空间中的表示。

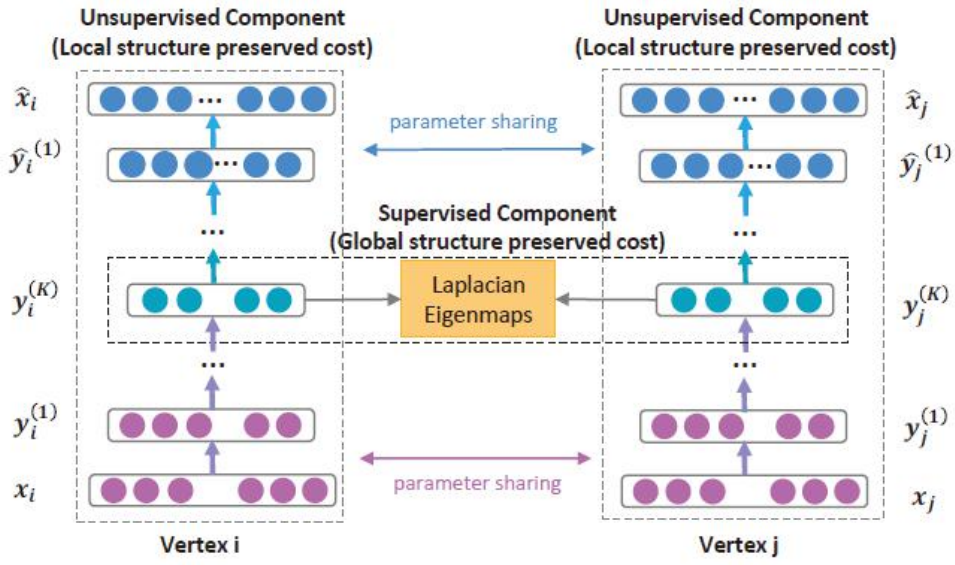


图 6 SDNE 模型框架

总结

现有的网络表示学习算法在节点分类、链接预测和社群发现等任务上都有不错的表现，对大规模的网络结构具有重要的意义。现有的网络表示学习方法主要依赖于静态的网络拓扑结构信息，而忽略了网络结构的动态性、网络中节点的异质性、节点拥有的大量外部信息等。

因此，基于网络的表示学习旨在探索能够更好地研究分析复杂信息网络中的节点间的联系，寻找解决网络背景下的各种实际问题的通用方法，有效融合网络结构与节点外部信息，形成更 具有区分性的网络表示。近年来，网络表示学习问题也吸引了大量研究者的目光，相关的论文工作也层出不穷。