# HIN2Vec：异质信息网络中的表示学习 | PaperDaily #18

原创 陆元福 PaperWeekly 2017-11-24


在碎片化阅读充斥眼球的时代，越来越少的人会去关注每篇论文背后的探索和思考。

在这个栏目里，你会快速 get 每篇精选论文的亮点和痛点，时刻紧跟 AI 前沿成果。

点击本文底部的「**阅读原文**」即刻加入社区，查看更多最新论文推荐。

这是 PaperDaily 的第 **18** 篇文章

本期推荐的论文笔记来自 PaperWeekly 社区用户 **@YFLu**。这篇论文发表在刚刚结束的 2017CIKM 会议上，论文提出了一种**针对异质信息网络的表示学习框架 HIN2Vec**。

不同于之前很多基于 Skip-gram 语言模型的工作，**HIN2Vec 的核心是一个神经网络模型，不仅能够学习网络中节点的表示，同时还学到了关系（元路径）的表示。**

如果你对本文工作感兴趣，点击底部的**阅读原文**即可查看原论文。

**关于作者：陆元福，北京邮电大学计算机系硕士生，研究方向为异质信息网络的表示学习。**

■ 论文 | HIN2Vec: Explore Meta-paths in Heterogeneous Information Networks for Representation Learning
■ 链接 | https://www.paperweekly.site/papers/1182
■ 作者 | YFLu

HIN2Vec 是一篇关于异质信息网络中的表示学习的论文，发表在刚刚结束的 2017CIKM 会议上。这篇论文和我最近的工作有一些相似之处，一些想法甚至有些相同，同样有很多地方值得借鉴。

论文提出了一种**针对异质信息网络的表示学习框架 HIN2Vec**，不同于之前很多基于 Skip-gram 语言模型的工作，**HIN2Vec 的核心是一个神经网络模型，不仅能够学习网络中节点的表示，同时还学到了关系（元路径）的表示。**

同时论文还对异质信息网络中表示学习的一些问题做了研究实验，例如：**元路径向量的正则化、负采样**过程中节点的选择以及随机游走中的**循环序列**问题。

## Introduction

论文首先指出了现有模型存在的一些问题，之前的很多工作仅仅局限于同质信息网络，而且往往只考虑节点之间的整合的信息或者限制类型的关系。虽然 ESim 模型考虑了节点间的不同关系，但是该模型过于依赖人为定义的元路径以及每条元路径人为设置的权重。

基于现有模型存在的问题，论文提出了 HIN2Vec 模型，通过研究节点之间不同类型的关系和网络结构，学习异质信息网络中丰富的信息。由于**不同的元路径可能有不同的语义信息**，所以作者认为**对嵌入在元路径和整个网络结构中的丰富信息进行编码，有助于学习更有意义的表示。**

和之前的一些模型相比，HIN2Vec 模型**保留了更多的上下文信息**，不仅假设**存在关系的两个节点是相关的**，而且还**区分节点之间的不同关系**，并通过共同**学习关系向量**区别对待。

论文的主要贡献：

- 证明了节点间不同类型的关系能够更好的捕获更多嵌入在网络结构中的细节信息，因此通过捕获节点间各种不同类型的关系，有助于网络的表示学习。

- 提出了 HIN2Vec 模型，包括两部分：首先，**基于随机游走和负采样生成训练数据**，然后，设计**逻辑二元分类器用于预测两个给定的节点是否存在特定的关系。**同时，考虑了循环序列、负采样和正则化问题。

- 实验很充分，包括多标签分类和链路预测，同时实验研究了循环序列、负采样以及正则化对实验分类结果的影响。

# HIN2Vec

## Framework

HIN2Vec 模型分为两部分：基于随机游走的数据生成部分和表示学习部分。数据生成部分，基于随机游走和负采样生成符合目标关系的数据，以用于表示学习。表示学习部分是一个神经网络模型，通过最大化预测节点之间关系的可能性，**同时学习节点和关系的表示向量**，模型的整体框架可以见下图。

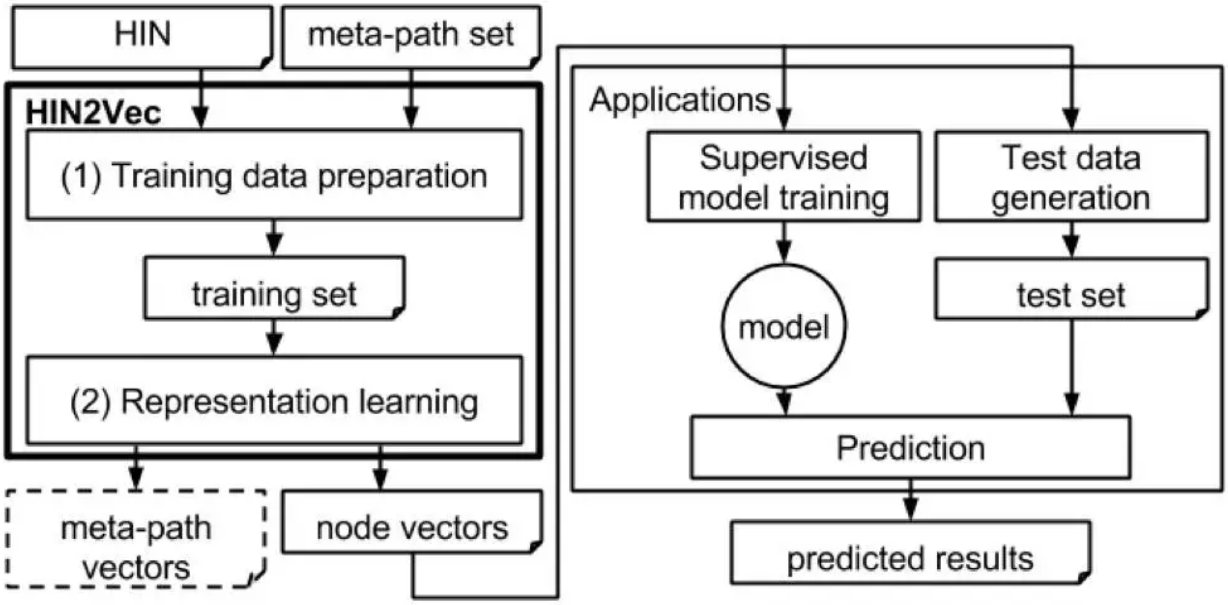


**Figure 1: Overview of the HIN2Vec framework**

值得注意的是，HIN2Vec 模型同时学习了节点和关系（元路径）的表示向量，这种多任务学习（multi-task learning）方法能够把不同关系的丰富信息和整体网络结构联合嵌入到节点向量中。

## Representation Learning

HIN2Vec 模型的基本想法是**对于多个预测任务，每个任务对应于一条元路径，联合学习一个模型，学到每个节点的向量表示**，所以一个简单的想法就是构建一个神经网络模型，**预测任意给定节点对之间的一组目标关系。**

论文最初的想法是一个多分类模型，即给定两个节点和一个目标关系集合，通过下图 2 的神经网络模型训练预测概率值 P(ri|x,y),(i=1...|R|)，但是对于这样的一个模型的训练数据的获取是非常复

杂的，**因为对于一个复杂网络而言，获取网络中两个节点的所有关系是很困难的。**

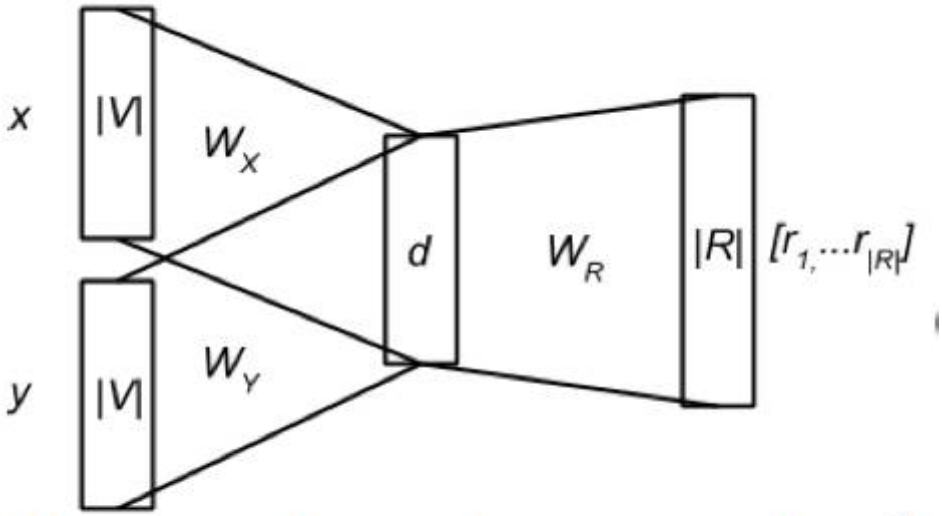


**Figure 2: A conceptual model for HIN2Vec**

所以论文退而求其次，将问题简化成**二分类问题**，即**给定两个节点 x,y，预测节点间是否存在确定的关系 r**，这样就避免了遍历网络中的所有关系，图 3 所示就是 HIN2Vec 的神经网络模型。

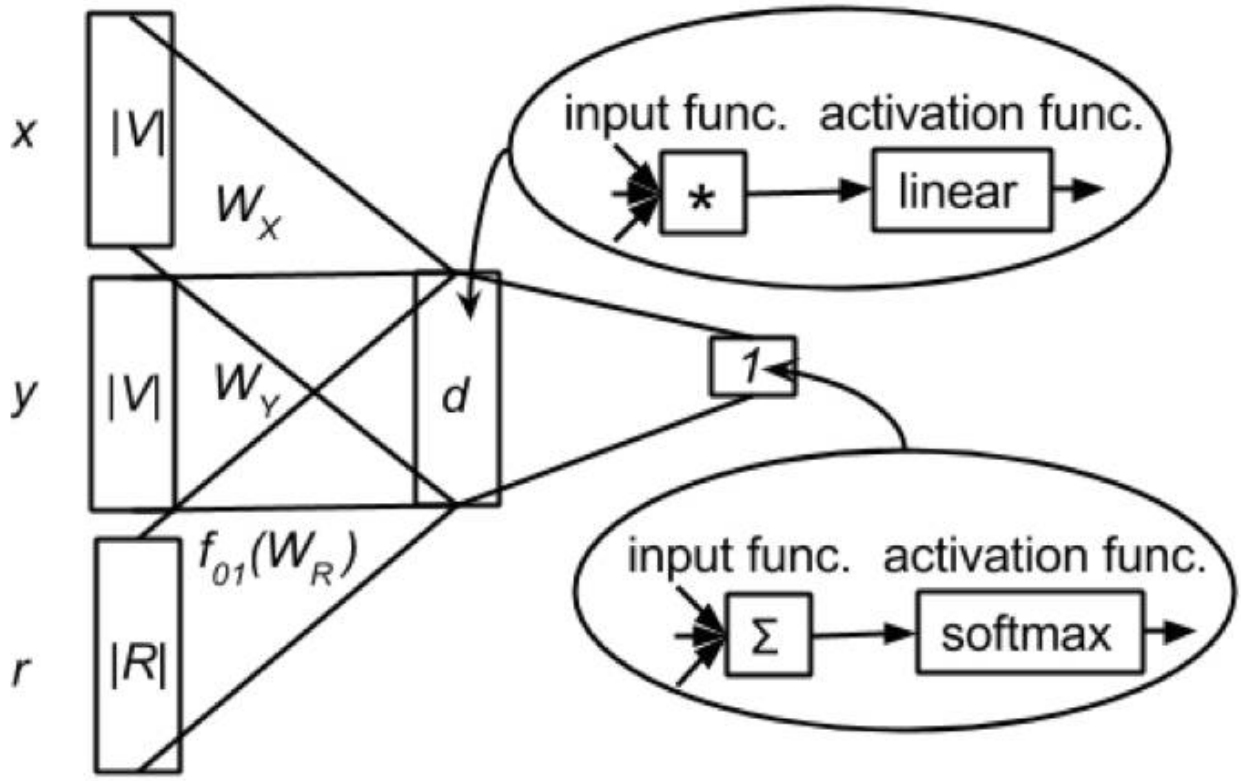


**Figure 4: The HIN2Vec NN model**

模型的输入层是三个 one-hot 编码的向量 $\vec{x}, \vec{y}, \vec{r}$，经过隐层层转换到隐含向量 $W'_X\vec{x}, W'_Y\vec{y}, f_{01}(W'_R\vec{r})$，值得注意的是，因为关系和节点的语义含义是不同的，所以论文**对关系向量 r 做了正则化处理**，这种处理方式限制了关系向量的值在 0 到 1 之间。

然后，模型对三个隐含向量运用一个 Hadamard 函数（例如，对应元素相乘），对乘积再运用一个线性激活函数，最后输出层对成绩求和后的值进行一个 sigmoid 非线性转换。

模型的训练数据集是以四元组的形式给出的，形如 $，其中，其中 L(x,y,r) 指示指示 x,y 之间是否存在关系之间是否存在关系 r。具体的：

$$O_{x,y,r}(x,y,r) = \begin{cases} P(r|x,y) & \text{if L(x,y,r)} = 1 \\ 1 - P(r|x,y) & \text{if L(x,y,r)} = 0 \end{cases} \$$

$$\log O_{x,y,r}(x,y,r) = L(x,y,r)\log P(r|x,y) + [1 - L(x,y,r)]\log[1 - P(r|x,y)]$$

$$P(r|x,y) = sigmoid(\sum W'^{\vec{r}}_{X}\vec{x} \odot W'^{\vec{r}}_{Y}\vec{y} \odot f_{01}(W'^{\vec{r}}_{R}\vec{r}))$$

## Traning Data Preparation

论文采用随机游走的方式生成节点序列，但是需要注意的是，**不同于 metapath2vec[1] 按照给定元路径模式游走的方式，HIN2Vec 模型完全随机选择游走节点，只要节点有连接均可游走。**

例如，随机游走得到序列 P1,P2,A1,P3,A1，那么对于节点 P1，可以产生训练数据 $和$。

在论文中，作者讨论了随机游走过程中可能出现的**循环**的情况，提出通过检查重复节点的方式消除循环，并在实验部分分析了是否消环对实验结果的影响，但是个人认为这个地方的原理性介绍比较欠缺，对于消除循环的具体做法没有给出很详细的说明解释，循环的检测是根据前面已生成的所有节点还是部分节点，也没有给出说明。

论文还讨论了训练数据集中负样本的选择，论文也是采用 word2vec 中的负采样的方法产生负样本。对于一个正样本 $，通过随机替换，通过随机替换 x,y,r 中的任意一个，生成负样本中的任意一个，生成负样本，其中，其中 x'' 和 y'' 之间不一定有确定的关系之间不一定有确定的关系 r''$。

但是，由于网络中的关系数量是很少的，节点的数量远远大于关系的数量，这样就很容易产生**错误的负样本**（其实是可能正样本），所以论文采用**只随机替换 x 或 y 中的一个，而保持 r 是不变的，同时保持 x 或 y 的类型不变。**

## Summary

总体来说，论文的想法还是很新颖的，把节点和节点间的关系作为一种二分类问题考虑，给定两个节点 x,y，通过预测节点之间是否存在确定的关系 r，同时学习到了节点和关系的向量表示。

此外，论文考虑到了节点和关系的语义是不同的，因此它们的表示空间也应该不通，所以论文对关系向量运用了一个正则函数。对于随机游走过程中可能会出现循环节点的问题，论文也给出了实验分析，同时阐述了负采样时候节点及节点类型的选择。

个人认为，论文的不足之处在于随机游走过程中如何消除循环，没有给出较为详细的说明。此外，对于学习到的关系的表示如何应用到实际的数据挖掘任务中，论文也没有给出实验分析。

## Reference

[1] Dong Y, Chawla N V, Swami A. metapath2vec: Scalable representation learning for heterogeneous networks[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017: 135-144.