大咖聊技术|一切皆可Embedding：聊一聊node2vec算法

原创 Huang Mingxia 百威亚太数据科学 2020-07-20

Embedding思想的发展要归功于词嵌入（word embedding）模型word2vec。随着自然语言处理（NLP）技术的发展和推广，这一思想也逐渐被应用到推荐、搜索、广告等其它领域。简单来说，embedding就是用一个低维向量表示一个物体，可以是一个词，或一个商品，或是一个电影等等。这个向量的性质是能使距离相近的向量对应的物体有某种相似的特点，比如是一组近义词，或经常被一起购买的商品，或是同类型的电影。更神奇的是，这些向量之间还具有数学运算的关系，例如word2vec中经典的

$$\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} = \overrightarrow{queen}$$

今天给大家带来的是一个基于图数据结构的embedding算法node2vec，从名字上就可以看出它和word2vec一定有血缘关系，相信熟悉word2vec的同学们很快就能理解这个算法的思想。

图表示学习

Word2vec是基于句子序列中词与词的上下文共现关系来学习词的向量表征。很快大家发现，在非NLP的领域里，只要我们能用item构建出合理的序列，同样可以基于item的共现关系得到item的向量表示。但这必须是基于“序列”样本，在很多尤其是互联网场景下，数据对象之间的关系更多呈现的是图结构。而图嵌入（graph embedding）的大部分工作，就是如何构造合理的节点序列。

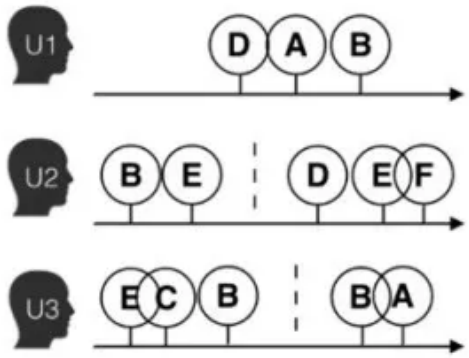
从word2vec到Graph Embedding

经典的DeepWalk方法

早期影响力较大的graph embedding方法是2014年提出的DeepWalk，它的主要思想是在网络上进行随机游走，从一个节点出发，随机采样它的一个邻居作为下一个节点，重复得到大量的节点序列，然后将它们作为训练样本输入word2vec得到节点的embedding。下图展示一个电商场景下Deepwalk的过程：

a

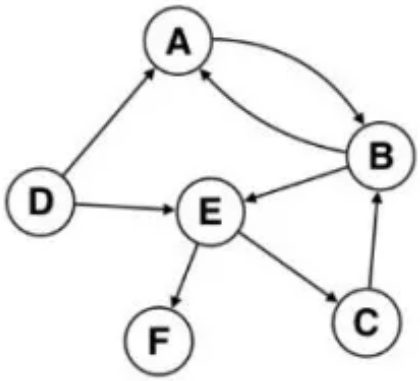
原始的用户行为序列：例如电商场景中用户U1先后购买了物品D、物品A和物品B。



(a) Users' behavior sequences.

b

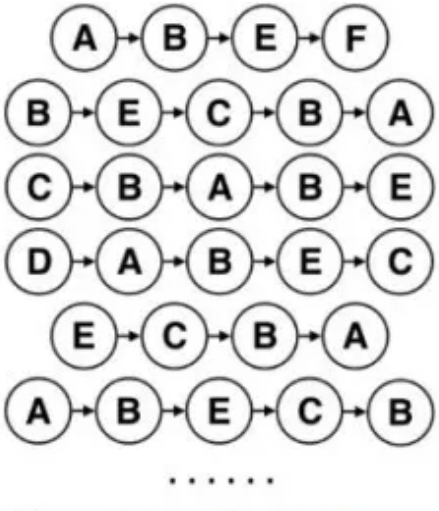
基于用户行为序列构建了物品相关图：例如当用户先后购买了物品A和物品B，就建立一条A到B的有向边。如果产生了多条相同的有向边，则它的权重就会被加强。



(b) Item graph construction.

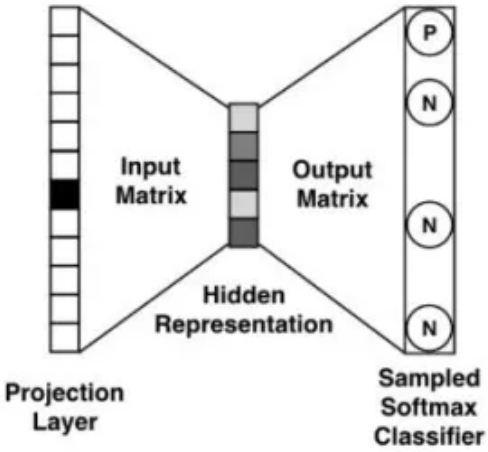
c

采用随机游走的方式随机选择初始点，重新生成物品序列。



(c) Random walk generation.

d

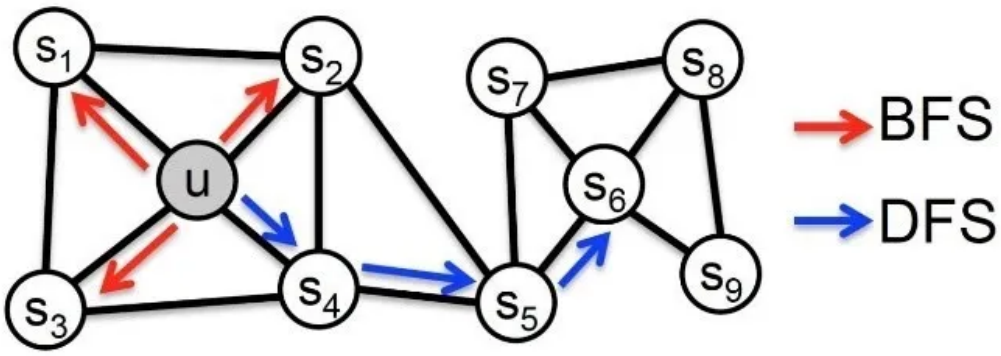


(d) Embedding with Skip-Gram.

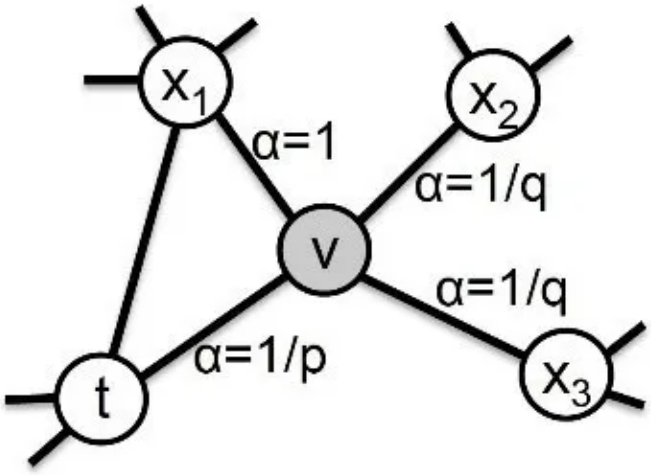
使用word2vec中的skipgram模型来训练得到节点的向量。

Node2vec
基于DeepWalk的改进

Node2vec在2016年被提出，考虑了网络的同质性（homophily）和结构性（structural equivalence）的权衡。具体来讲，网络的"同质性"指的使近距离节点的embedding应该尽量相似，如下图中u和s_1、s_2和s_3都应该有相近的向量表示。而"结构性"指的是结构上相似的节点的embedding也应该尽量接近，如下图的u和s_6应该相似。



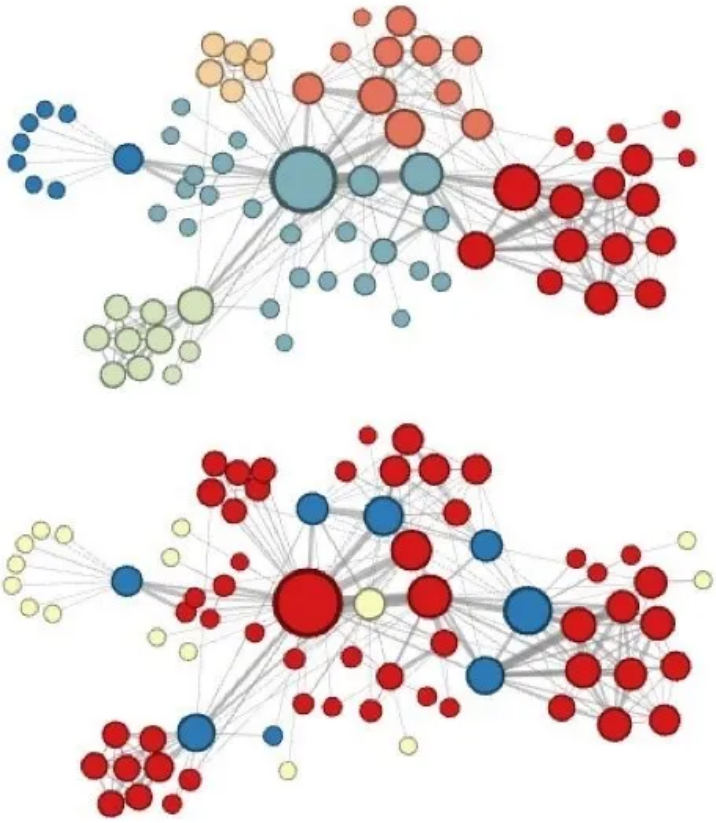
为了使Graph Embedding的结果具有结构性，在随机游走的过程中，需要让游走的过程更倾向于宽度优先搜索（BFS）；另一方面，为了抓住网络的结构性，就需要随机游走更倾向于深度优先搜索（DFS）。在node2vec算法中，通过两个超参数p和q来控制节点间的跳转概率来控制随机游走的倾向：



如图所示，当从节点*t*跳转到当前节点*v*后，下一步的跳转到节点*x*的概率为

$$P_{vx} = \alpha w_{vx}$$

其中α如图所示，参数*p*为返回参数（return parameter），*p*越小则随机游走返回节点*t*的可能性越大，算法更倾向于BFS。参数*q*为进出参数（in-out parameter），*q*越小则随机游走到远方的节点可能性越大，算法更倾向于DFS。



这种灵活的游走方式使得node2vec可以更好的挖掘出图关系的特征表达，以推荐系统为例：

- 同质性相同的物品可能是同类目、同属性，或者经常被加到同一个购物车的商品；
- 结构性相同的物品可能是爆款、满减凑单，套餐折扣等具有相似结构特征的商品。

这两种信息都很重要，都可以是推荐的理由。因此，甚至可以通过不同参数的node2vec来召回不同的商品列表。

如果前面提及DeepWalk的时候你还有为什么明明有了序列数据还要用Graph Embedding这样的疑问，相信到这里也找到一些答案了。首先，在图上进行随机游走可以生成出原始数据里没有出现过的序列，丰富了embedding的训练数据。另外，像Node2vec提供的游走方式可以挖掘出结构相似等序列本身没有的信息。

Node2vec只是Graph Embedding的一个开始，主要针对同构的图。但现实中的图数据往往是异构的，不仅包含异构的节点（用户和商品），而且包含异构的边（点击、购买等多种行为），将这些异构的信息利用进来就能有更好的结果，一些基于异构网络的嵌入学习也被广泛研究（感兴趣的同学可以看看metapath2vec和GATNE）。对于我们的场景，也有需要挖掘客户（消费者或经销商）之间共性的。那么除了聚类、画像这些手段，是不是也可以通过node2vec或者其它的graph embedding技巧去得到用户甚至是SKU的embedding呢?

- END -

本期员工大咖

Dr. Huang Mingxia

PhD, Applied Mathematics, Tongji University

Bachelor, Mathematics and Applied Mathematics, Tongji University