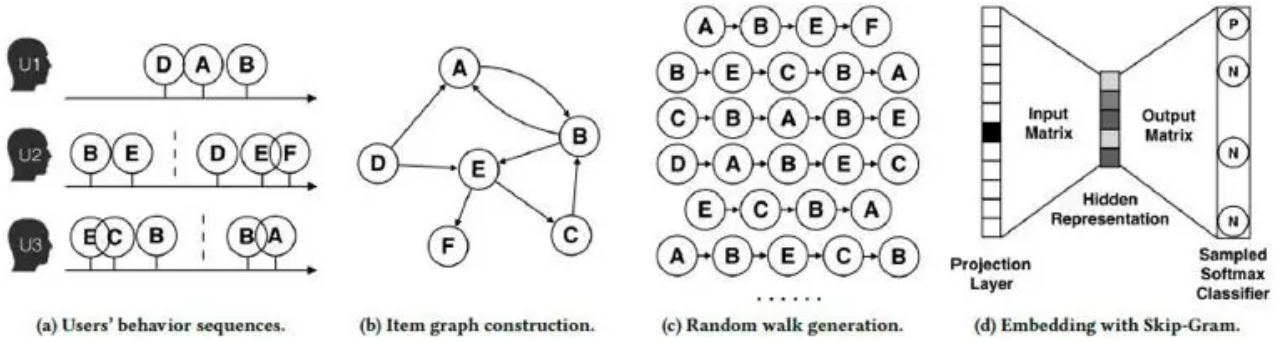
Graph Embedding方案之DeepWalk

原创 傲海 凡人机器学习 2019-11-13

对于算法不太了解的同学，这是一个看上去很没吸引力的标题，预计点击量超不过200。最近非常迷恋一句话“万物皆可Embedding”，讲的是世间所有的事物都能通过某种方法被向量表示，一旦事物被向量表示了就可以通过乘法去做进一步逻辑处理。比如商品A被表示为向量m，商品B被表示为向量n，则m*n的结果就是A和B的相似关系。把事物Embedding的方法有很多，今天就来介绍DeepWalk，一种把图关系向量化的方法。

DeepWalk有什么用呢？在推荐、文本分类等领域都有很多场景。比如用户观看视频这件事，所有用户的视频观看顺序组合到一起会构成一种图关系，每个节点是视频，边是观看次数。当有一个客户先后看了A->B->C三个视频，那么下个视频会看什么呢，就可以用DeepWalk将所有视频向量化，然后所有视频的向量分别与C视频的向量相乘，分数最高的就可以作为下一个推荐视频。

下面这张图直观展示了DeepWalk处理数据的过程：

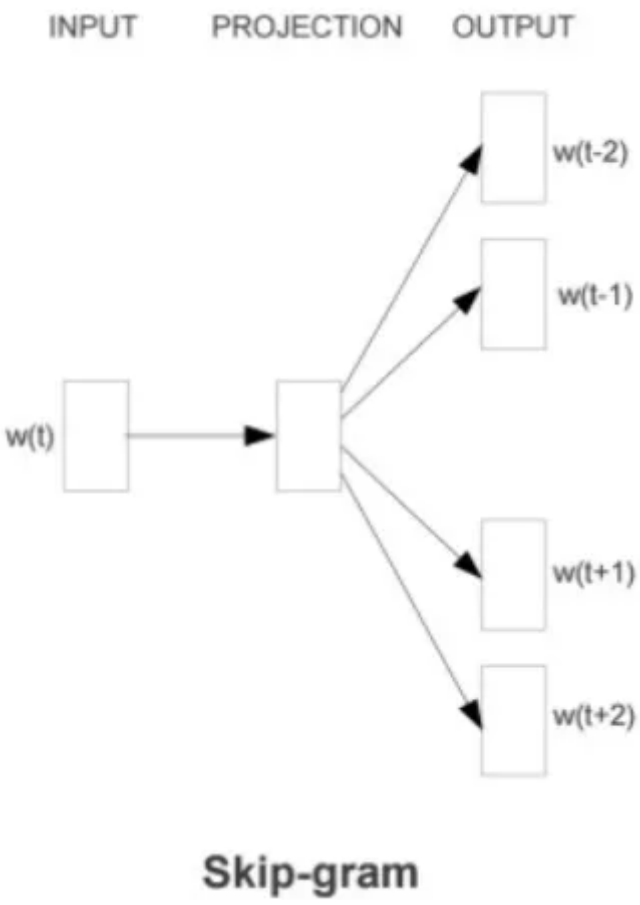


(a) Users' behavior sequences. (b) Item graph construction. (c) Random walk generation. (d) Embedding with Skip-Gram.

具体DeepWalk做事物Embedding可以分两步，第一步是Random Walk，第二步是Word2Vector。为了大家更好的理解，我们先从Word2Vector开始讲。

Word2Vector

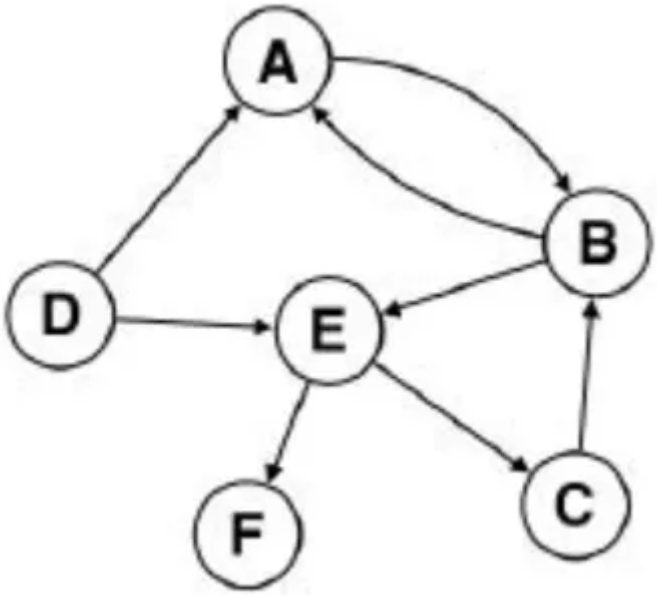
有的同学会奇怪，做Graph Embedding跟Word2Vector有什么关系，Word2Vector不是一个文本向量化的方法么？图像量化其实可以巧妙的转化成文本向量化。Word2Vector有一种模式叫做Skip-Gram，就是给定一个文章中的某个词，预测这个词的上下文，如下图展示：



Skip-gram

比如我们输入一个文本“傲海是北京朝阳地区第一帅哥”，“朝阳”的上一个词是“北京”，下一个词是“地区”。当这种训练样本很多的时候，比如我们找到大量含“朝阳”的词的语句进行训练，就可以通过Skip-Gram得出“朝阳”这个词的向量形式，这就是Skip-Gram这个算法的功能，这里需要指定一个窗口的概念，就是训练的时候取每个词上下几个词作为输入。那么这种上下文关系如何映射到图向量化工作呢？我们接着看。

Random Walk



先来看看这个有向图，我们可以从图关系中找到几个序列，比如A->B->E->C，这是一个有向关系链。这个关系跟上文提到的“傲海是北京朝阳地区第一帅哥”是有一定相似性的，我们把“北京”看成A，“朝阳”看成B，“地区”看成E，这种图的先后关系可以映射成文本的先后出现关系。Random Walk做的事就是设置一个窗口，然后顺着图关系去随机找到类似于A->B->E这样的先后关系。

比如我们设窗口为2，那么以B为定点，可以找到如B->E、B->A这样的前后关联关系，这种关系放到Word2Vector里就能生成每个定点的向量表示。

Ok~讲的比较浅，希望大家可以理解哈，希望有帮助，谢谢