

# DeepWalk随机游走实现图向量嵌入

在NLP任务中，word2vec是一种常用的word embedding方法，word2vec通过语料库中的句子序列来描述词与词的共现关系，进而学习到词语的向量表示。

DeepWalk的思想类似word2vec，使用**图中节点与节点的共现关系**来学习节点的向量表示。那么关键的问题就是如何来描述节点与节点的共现关系，DeepWalk给出的方法是使用随机游走(RandomWalk)的方式在图中进行节点采样。

RandomWalk是一种**可重复访问已访问节点的深度优先遍历**算法。给定当前访问起始节点，从其邻居中随机采样节点作为下一个访问节点，重复此过程，直到访问序列长度满足预设条件。

获取足够数量的节点访问序列后，使用skip-gram model 进行向量学习。

对比维度	Word2Vec	DeepWalk
所属领域	自然语言处理 (NLP)	图表示学习 (Graph Representation Learning)
核心目标	学习词的低维向量表示，捕捉语义/语法相似性	学习图中节点的低维向量表示，捕捉结构/角色相似性
输入数据	大量文本语料 (句子序列)	一个图 (Graph)，，由节点和边组成
基本单元	词(word) 自然语言中的真实句子	节点 (node)
“句子”来源 上下文定义	固定窗口内的邻近词 (如前后5个词)	通过随机游走 (Random Walk) 在图上生成的节点序列 随机游走序列中窗口范围内的邻近节点
模型架构	Skip-gram或CBOW	仅使用Skip-gram (通常)
是否需要标签	否 (无监督)	否 (无监督)
输出	每个词的embedding向量 (如100-300维)	每个节点的embedding向量(如64-128维)
相似性含义	语义相似 (如“king”=“queen”)	结构相似 (如两个社区中心节点)
关键技术	负采样、层次Softmax、向量运算	随机游走+Word2Vec (Skip-gram)
可处理数据类型	序列数据 (文本)	非欧几里得结构数据 (图)
典型应用	机器翻译、文本分类、情感分析、问答系统	社交网络分析、推荐系统、异常检测、链接预测、社区发现
是否依赖图结构	×否	是
是否利用节点属性	不适用 (只有词)	×否 (仅用拓扑结构，忽略节点特征)
扩展性	高 (可处理十亿级词)	中高 (依赖游走生成效率，适合中等规模图)
后续发展	FastText、GloVe等	Node2Vec、LINE、GraphSAGE、GCN等

- DeepWalk 不是 Word2Vec 的替代品，而是其思想在图数据上的成功迁移。
- DeepWalk = 随机游走（生成“句子”） + Word2Vec（Skip-gram，学习“词向量”）
- 二者都基于 分布假设（Distributional Hypothesis）：
  - NLP：“出现在相似上下文中的词，语义相似”
  - 图嵌入：“出现在相似游走上下文中的节点，结构相似”