关于Node2vec算法中Graph Embedding同质性和结构性的进一步探讨

原创 王喆的机器学习笔记 王喆的机器学习笔记 2019-05-07

收录于话题

#机器学习 7 #深度学习 10 #Graph Embedding 4 #图神经网络 3 #Embedding 4

关于Node2vec算法中

Graph Embedding

同质性和结构性的进一步探讨

这里是**「王喆的机器学习笔记」**的第十五篇文章，上篇文章 王喆：深度学习中不得不学的Graph Embedding方法 介绍了多种Graph Embedding的方法，其中node2vec方法的一个关键特性“**Graph Embedding的同质性和结构性”**非常有争议，也非常有意思，我觉得可以进一步跟大家探讨学习。

这里再回顾一下什么是网络的“**同质性**”和“**结构性**”。当然建议大家最好还是直接去看上一篇文章。

Node2vec是在DeepWalk的基础上更进一步，通过调整随机游走权重的方法使graph embedding的结果在网络的**同质性（homophily）**和**结构对等性（structural equivalence，下面简称结构性）**中进行权衡。

其中，网络的“**同质性**”指的是距离相近节点的embedding应该尽量近似，如图1中，节点u与其相连的节点s1、s2、s3、s4的embedding表达应该是接近的，这就是“**同质性**”的体现。

“结构性”指的是结构上相似的节点的embedding应该尽量接近，图1中节点u和节点s6都是各自

局域网络的中心节点，结构上相似，其embedding的表达也应该近似，这是“**结构性**”的体现。

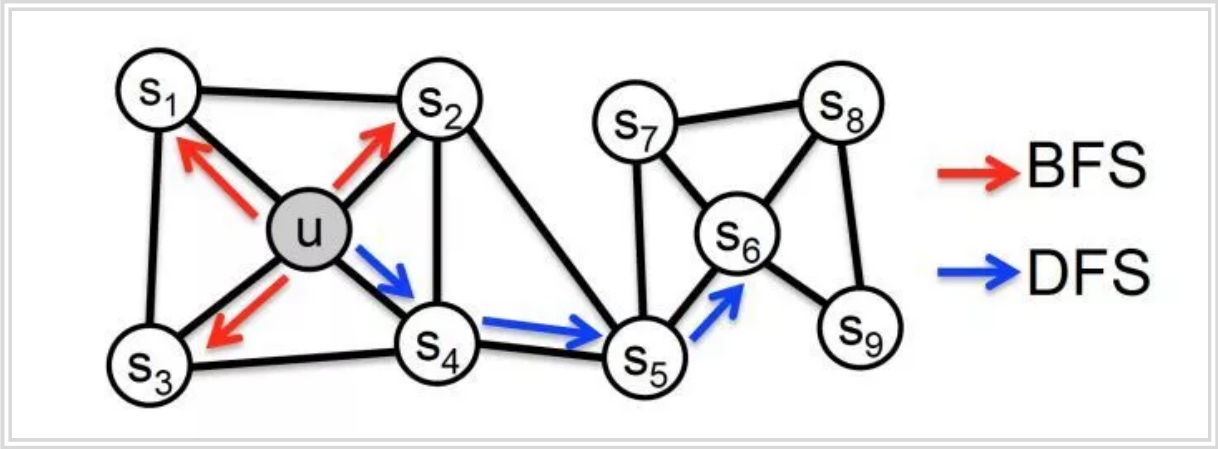

图1 宽度优先搜索（BFS）和 深度优先搜索（DFS）示意图

Node2vec在具体的实现上，是通过调整模型参数，使随机游走更倾向于宽度优先搜索（BFS），或者深度优先搜索（DFS），从而在embedding结果中更好的体现“结构性”或者“同质性”，那么，下面就到了比较容易让人困惑的问题了：

到底是宽度优先搜索（BFS）更能体现“结构性”，还是深度优先搜索（DFS）更能体现“结构性”呢?

当然，对于“同质性”，我们同样可以问出一个对偶问题。到底是宽度优先搜索（BFS）更能体现“同质性”，还是深度优先搜索（DFS）更能体现“同质性”呢?

我这里先不说论文原文的答案，还是希望大家能够自己思考起码五分钟。

好，五分钟过了，我说一下我最初的理解，大家看一下有没有道理。

为了使Graph Embedding的结果能够表达网络的**同质性**，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）**，因为BFS更喜欢游走到跟当前节点有直接连接的节点上，因此就会有更多同质性信息包含到生成的样本序列中，从而被embedding表达；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）**，因为DFS会更倾向于通过多次跳转，游走到远方的节点上，使得生成的样本序列包含更多网络的整体结构信息。

不知道大家跟我最初的想法是不是一致，遗憾的是，这种想法是与论文原文的解释相反的，恰恰是**BFS更多抓住了网络的结构性，而DFS更能体现网络的同质性。**为什么？

结合原文和自己的理解，我这里给出一个可能正确的解释，因为并没有通过实验去验证，最“正确”的做法当然是你自己实现之后评估一下哪种结果更能体现同质性。

对于宽度优先搜索（BFS）来说，其搜索往往是在当前节点（这里以节点u为例）的邻域进行的，特别是在node2vec中，由于存在所谓的“返回概率”，所以即使从u搜索到了s1，也有很大概率从s1再返回u，所以BFS产生的序列往往是在u附近的节点间进行来回的震荡，这就相当于对u周围的网络结构进行了一次**微观扫描（microscope view）**。

那么这个微观扫描当然更容易得到**微观结构**的观察，所以BFS就更容易使embedding结果更多反应网络的“结构性”。这里我需要纠正一下大家对“结构性”的理解，正如上面所说的一样，这里的“结构”更多的指的是微观结构，而不是大范围内，甚至整个网络范围内的宏观结构，而是一阶、二阶范围内的微观结构。

再举个例子理解一下，比如对于节点u和节点s9这两个节点来说，节点u是局部网络的中心节点，节点s9是一个十分边缘的节点。那么在对这个网络进行多次BFS随机游走的过程中，节点u肯定会被多次遍历到，而且会与s1-s4等更多节点发生联系，而边缘节点s9无论从遍历次数，还是从邻接点的丰富程度来说都远不及节点u，因此两者的embedding自然区别很大。

如果用DFS进行遍历，由于遍历存在更大的不确定性，因此s9有更大的可能被包含在更多的序列中，并跟更多的节点发生联系，这就减弱了局部结构性的信息。类似于平滑了结构性的信息，自然是不如BFS更能反应“微观结构”了。

另一方面，**为什么说DFS更能反应“同质性”呢**？

这里还要对“同质性”也进行一个纠正，这里的“同质性”不是指一阶、二阶这类非常局限的同质性，而是在相对较广范围内的，能够发现一个社区、一个群、一个聚集类别的“同质性”。要发现这类同质性，当然需要使用DFS进行更广范围内的探索。如果仅用BFS在微观范围内探索，如何发现一个社区的边界在哪里呢？

所以，**DFS相当于对网络结构进行了一次宏观扫描**（macroscope view），只有在宏观的视角，才能发现大的集团、社区的聚集性，和集团内部节点的“同质性”。

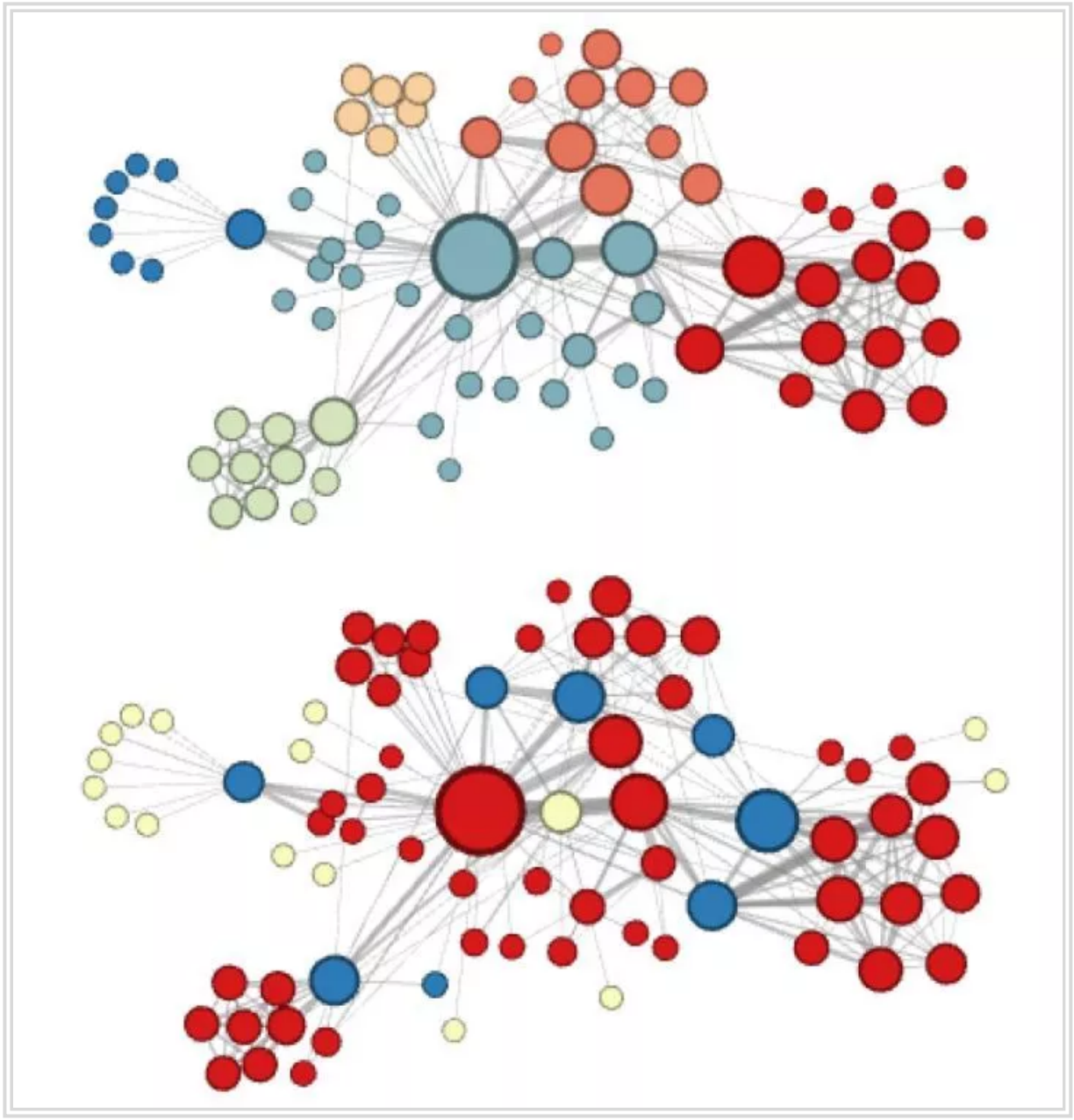

图2 上图是DFS结果，下图是BFS结果

论文中最后通过实验的方式验证了DFS和BFS对于“同质性”和“结构性”的挖掘结果。颜色接近的节点代表其embedding的相似性更强。

图2上图是node2vec更倾向于DFS的结果，可以看到各聚类的内部节点相似，这是网络“同质性”的体现；而图2下图中，结构类似节点的embedding更为相似，这是“结构性”的体现。

以上是我对node2vec中网络“同质性”和“结构性”的理解，**非常希望有实践经验的同学能够进一步分享node2vec或者其他graph embedding方法的理解和实践经验。**

最后是按照惯例的讨论题目：

如果是类似淘宝的商品推荐场景，那么什么样的商品之间是“同质性”较强的？什么样的商品之间是“结构性”相似度较强的？其实上一遍文章已经分享了一些我的看法，但还是希望大家能够分享自己的观点，谢谢参与。

最后欢迎大家关注我的 **微信公众号：王喆的机器学习笔记（wangzhenotes）**，跟踪计算广告、推荐系统等机器学习领域前沿。想进一步交流的同学也可以通过公众号加我的微信一同探讨技术问题，谢谢！

参考资料：

1. 王喆：深度学习中不得不学的Graph Embedding方法
2. [Node2vec] Node2vec - Scalable Feature Learning for Networks (Stanford 2016)

—END—

每周关注计算广告、推荐系统和其他机器学习前沿文章，欢迎关注**王喆的机器学习笔记**

收录于话题 #机器学习·7个

上一篇
毕加索的「公牛」和机器学习的「特征工程」

下一篇
前深度学习时代CTR预估模型的演化之路