关于论文《Network Representation Learning: A Survey》的学习报告

这篇论文是针对信息网络的网络表示学习（NRL）方法进行研究的。NRL 的任务是把一个高维复杂的网络结构嵌入到一个潜在的低维空间中，而这个低维的空间应该尽可能多地保存原来网络的架构连通性以及节点属性，NRL 的结果是把原来的网络中的每个节点都表示成一个特征向量的形式，而这些节点对应的低维向量能应用到目前存在的简单但高效的机器学习算法中，而不必针对原来的复杂网络架构研发出较为复杂的机器学习算法。论文中也提出当前的 NRL 研究领域面临着的一些挑战：包括如何同时保存网络架构的局部与全局信息；如何保存网络中成员的相关信息；数据稀疏性以及算法的可扩展性等。

论文中作者首先提出了 NRL 问题的定义以及涉及这个问题的一些术语。然后给出了当前研究中 NRL 算法的分类以及详细介绍了一些当前应用较为成功与普遍的 NRL 算法。紧接着作者们介绍了 NRL 的一些成功应用。然后作者介绍了一些数据集以及算法的评估标准以便后续的研究者用来检验自己的算法的性能。论文的最后提出了在 NRL 领域的一些有潜力的研究方向。

1. **NRL 问题的定义以及一些相关术语**
   1) 信息网络：信息网络由点集、边集、节点属性与节点标签组成。信息网络中存在的公共网络性质是网络中很有用的信息，这些信息在一定程度上可以准确地表示该网络的潜在组成机制，因此对网络表示学习的过程来说是非常重要的信息。
   2) 一阶连通：一阶连通体现了网络包含的直接信息，具体而言就是网络中各个节点之间的直接连接关系。
   3) 二阶连通和高阶连通：与一阶连通不同的是，二阶连通与高阶连通反映的是节点之间的相似度，从而体现出这些节点在网络中是否扮演着相同的角色，这一点并不能直接从网络结构中看出。
   4) 社区内的连通性：网络结构中通常存在社区架构，即网络可以划分为许多个大大小小的社区，同一社区中成员的相似性大于不同社区间成员的相似性。社区内的连通性保存了同一社区中成员共同拥有的特性。这里要提出的一点是，除了网络的架构以外，节点属性也能用以评价两个节点之间的相似程度，因此在计算社区内的连通性时，应就可能考虑网络架构与节点属性，充分应用信息网络中的信息。
   5) 网络表示学习：网络表示学习是结合信息网络中的信息学习得到一个映射函数，将原来的网络通过该映射函数可以把原网络映射到一个新的网络空间中，使得两个在原来网络中相似的节点在新的映射空间中也能具有很大的相似性。NRL 得到的节点表示需要满足以下三个条件：低维、具备就可能多原来网络的信信息、以及连续（即节点需要表示成连续实数以便网络分析任务）。

2. **NRL 算法：**

   目前使用较多的 NRL 算法主要可以分为四种类型：基于矩阵分解的方法；基于 random walk 的方法；基于边模型的方法；基于深度学习的方法。

   优化方法：特征分解；替代优化；梯度下降法；随机梯度下降法。

这一部分没有做深入的学习与理解，只知道针对不同的分类问题存在哪些可以应用的算法，打算以后具体应用的时候才去详细学习。

3. NRL 的成功应用方面：

1) 节点分类：NRL 可以将网络中的节点表示成一个向量的形式，而相似的节点的向量应该也比较相近，利用这一性质以及某些具有标签的节点的向量可以很好地对网络中的节点进行分类。
2) 链接预测：链接预测的主要任务是从当前的网络信息中推导出潜在的网络成员间的关系，可以用来完善网络结构图或发现伪的成员关系，或进行社交网络中的好友推荐。而 network representation 的作用是利用节点表示挖掘出不同节点之间的相似度，从而可以更好地进行链接预测。
3) 可视化：传统的可视化方法对于小规模的数据量来说是非常有效的，然而当数据规模变大的时候，这些方法的效果就变得不好，因此，如何把大规模数据降维继而实现数据可视化就值得探讨。而这个问题就可以通过 network embedding 解决，network embedding 的任务就是把原来的高维数据映射到低维空间中。
4) 推荐系统：POI 推荐是根据用户的网上记录来进行推荐的，然而，应用到推荐系统中时使用的是用户-商品矩阵，从而导致该矩阵可能非常稀疏，当对稀疏矩阵应用数据挖掘算法时得到的效果一般不好。因此，应用 network embedding 把每个用户的所有商品记录变成一个低维向量，可以有效解决矩阵稀疏性问题，从而得到更好的推荐效果。
5) 知识图：在知识图中存在数百万的信息实体，而这些实体间也存在很多关系，同时，知识图有可能是一个异构网络，从而导致网络中存在不同类型的成员以及不同类型的连接，这与我们一般遇到同构网络有很大差别。由于知识图是一个非常巨大的网络，加上异构性导致了网络的表示变得非常复杂，从中进行数据挖掘任务就变得异常困难。然而，利用 network embedding 将原本的知识图中的每个实体都表示成一个低维的向量，通过向量之间的比较找到相似的实体可以有效地进行数据库查询任务。

4. 评估模型：

1) 网络重构：利用 NRL 得到的节点的表示向量，从这些向量中预测节点间的连接关系以及节点相似度，继而重构原来的网络。利用重构网络与原始网络的差别评估得到的节点表示向量的正确性。
2) 节点分类：通过 NRL 以后，将网络中的节点表示成向量的形式，对一部分节点使用某个分类算法训练出一个分类器以后，用这个分类器对剩余节点进行分类，最终用 F1-score 评测分类的效果。
3) 链接预测：链接预测可以用来评估得到的节点表示是否具有足够的信息表示网络的演变机制。利用链接预测来评估 NRL 的性能的具体方法是从原来的网络去掉一定数量的边，然后从剩余的网络结构中学习到节点表示，对节点表示进行链接预测观察是否能预测到去掉的边。
4) 聚类：通过 NRL 以后，将网络中的节点表示成向量的形式，对向量形式的网络节点运用 k-Means 算法对网络中的所有节点进行聚类，以网络中原来的存在的社区结构作为评测标准，利用聚类结果与评测标准计算 NMI 评测聚类

的效果。

5) 可视化：对 NRL 学习到的节点表示向量，利用 t-SNE 算法把多维向量映射到 2 维空间中，观察网络节点在 2 维空间中的分布情况，如同一类节点是否密集分布在一起，从而评测 NRL 得到的节点表示的效果。

5. 潜在研究方向:

1) 非线性：NRL 的目的是就可能多地保存原来网络的信息，当前存在的许多 NRL 方法是 matrix factorization 和 random walk，而深度学习的方法能够得到一个非线性模型从而能保存更多更复杂的网络信息。因此在 NRL 方法中嵌套深度学习的方法是一个值得研究的课题。

2) 基于任务的 NRL 算法：NRL 问题往往与许多数据挖掘任务联系起来，不同的任务需要保存的信息往往不一致，而一般的 NRL 算法都是在这些信息中找一个权衡，因此对于一个特定任务，应该针对这个目标任务来保存与该人物相关的信息，而其他信息可以适当忽略。

3) 动态网络：由于网络可能会一直演变，随着时间的推移，某些节点可能会加入网络中或从网络中删除，从而导致了对静态网络的网络表示没有很大的作用。因此如何把静态网络的 NRL 方法拓展到动态网络中值得研究人员的更深探索。

4) 可扩展性：NRL 问题的可扩展性仍然是一个很大的挑战，当前存在的 NRL 方法诸如矩阵分解需要很大的时间复杂度，这对于大规模的网络来说是无法应用的。因此研究出一种时间复杂度低的 NRL 算法显得非常重要。

5) 异构性：随着大数据时代的到来，对异构网络的研究也越来越热门。在异构网络中，网络成员（即网络节点）的类型并不总是一样的，因而也导致了成员之间的连接类型是多种多样的。这就导致了很难测量不同节点之间的连通性，从而导致很难得到对异构空间中各个空间都通用的低维空间来进行网络表示学习。因此如何更好地测量 cross-modal 数据的连通性值得研究人员去探索。