Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

网络嵌入研究方法综述

李新帅　硕士研究生

2020年3月22日

内容提要

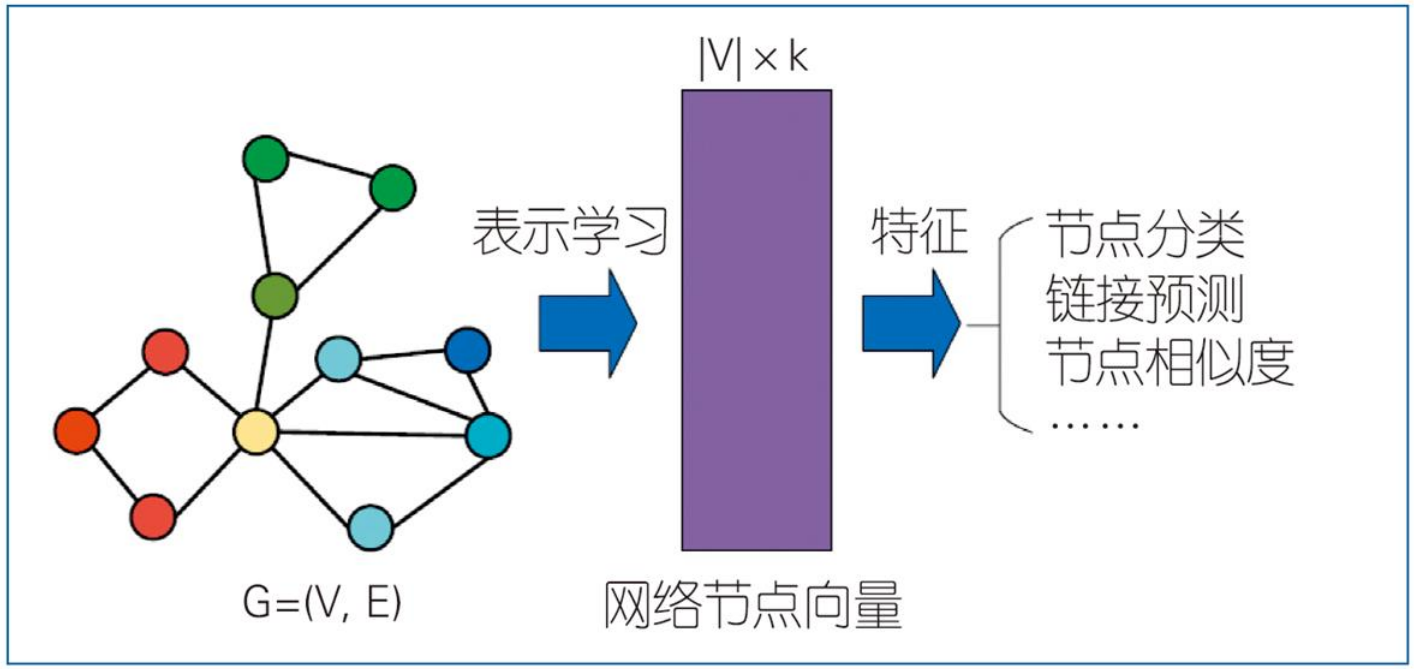
- 背景简介
- 基本概念
- 算法原理
- 应用总结
- 参考文献

预期收获

- 预期收获
  - 1. 了解网络嵌入基本思想
  - 2. 了解网络嵌入的学习方法
  - 3. 了解网络嵌入的应用

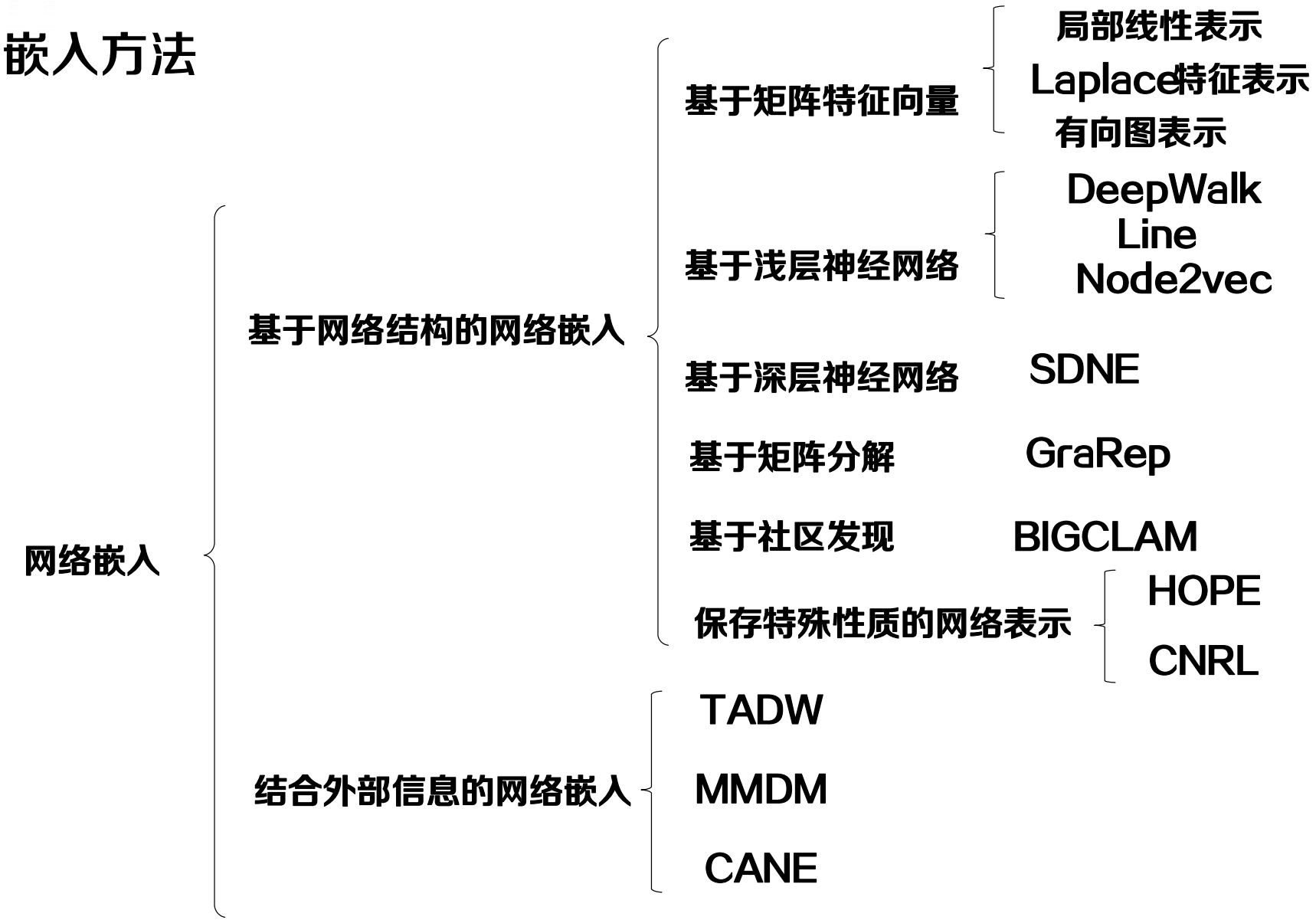
背景简介

- 网络嵌入（Network Embedding）
  - 网络表示学习（Network Representation Learning）
  - 图嵌入（Graph Embedding）



背景简介

- 网络嵌入方法


网络嵌入
基于网络结构的网络嵌入
基于矩阵特征向量
局部线性表示
Laplace特征表示
有向图表示
基于浅层神经网络
DeepWalk
Line
Node2vec
基于深层神经网络
SDNE
基于矩阵分解
GraRep
基于社区发现
BIGCLAM
保存特殊性质的网络表示
HOPE
CNRL
结合外部信息的网络嵌入
TADW
MMDM
CANE

基本概念

基本概念

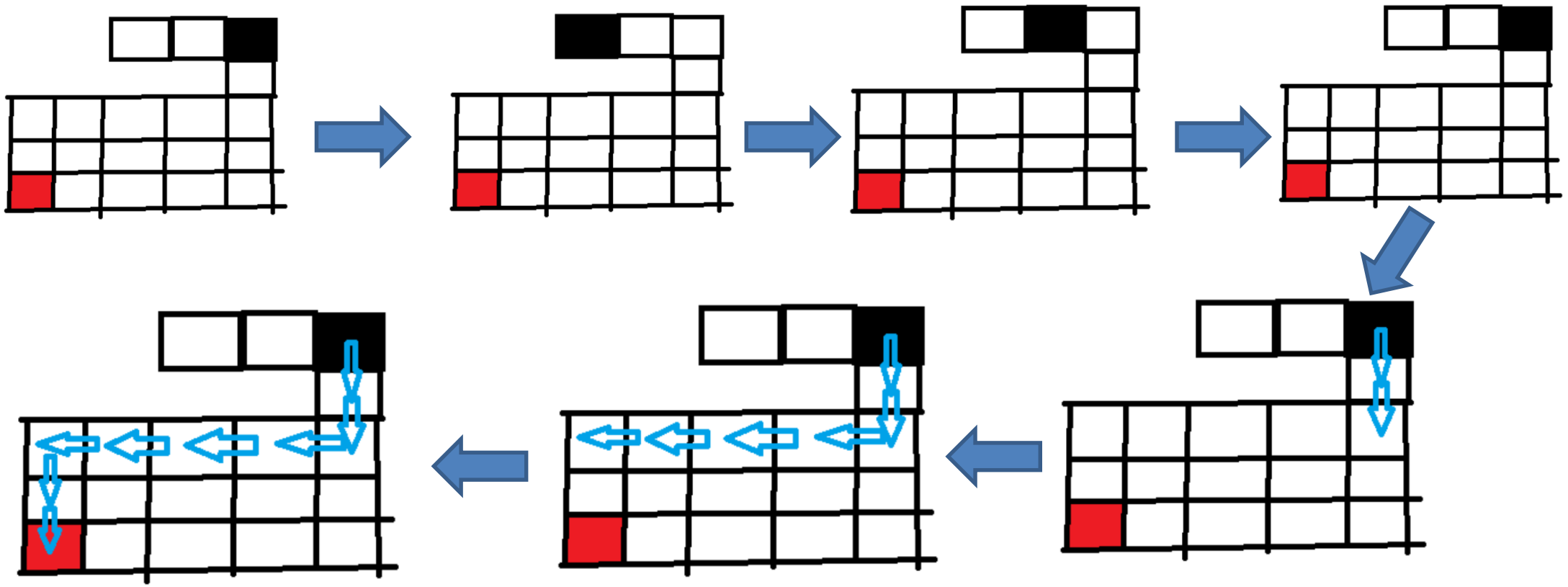
- 图
  - 将网络记为图$G(V,E)$，其中$V=\{v_1,v_2,\cdots,v_n\}$是节点集合，$E=\{e_{ij}\}_{i,j=1}^{n}$是边的集合，e_{ij}表示节点v_i和v_j之间的边。
- 邻接矩阵
  - 图$G(V,E)$的邻接矩阵定义为$\mathrm{A}\in\mathbb{R}^{n\times n}$，包含与每个边相关的非负权重：$a_{i,j}\geq 0$。如果$v_i$和$v_j$没有相互连接，则$a_{i,j}=0$。

基本概念

- 搜索方式
  - DFS（Deep First Search）深度优先搜索
  - BFS（Breath First Search）广度优先搜索

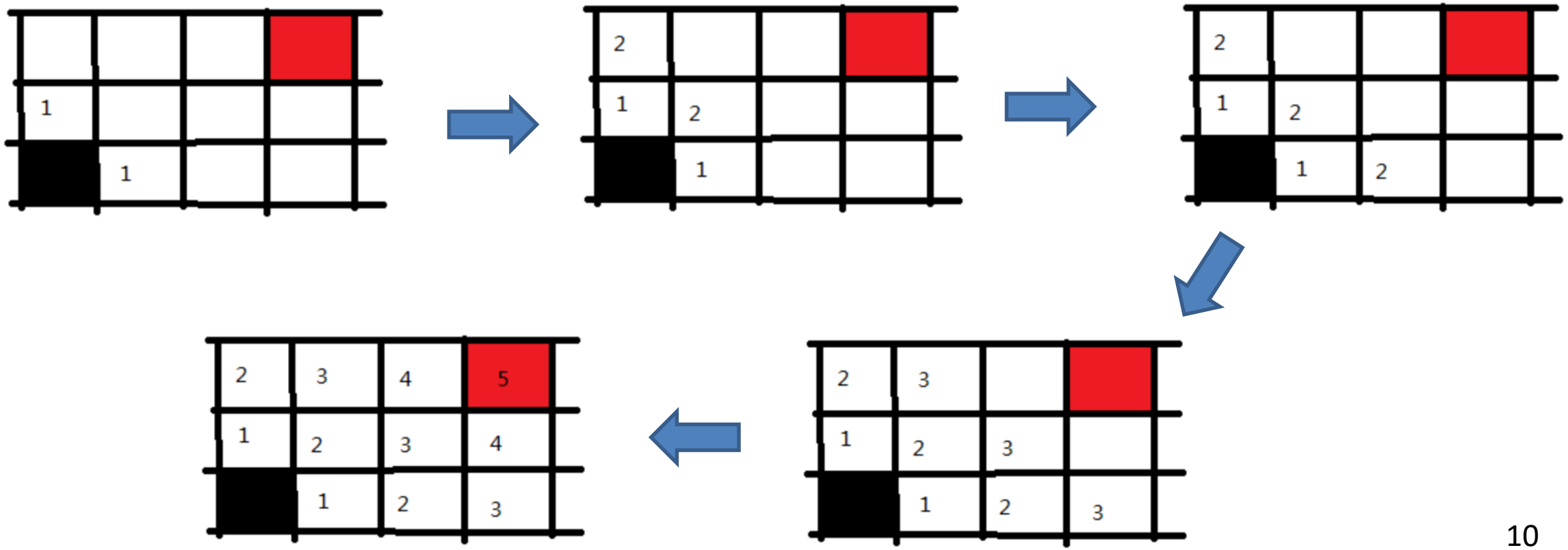
基本概念

- DFS（Deep First Search）深度优先搜索
  - 搜索步骤：递归下去，回溯上来
  - 深度优先



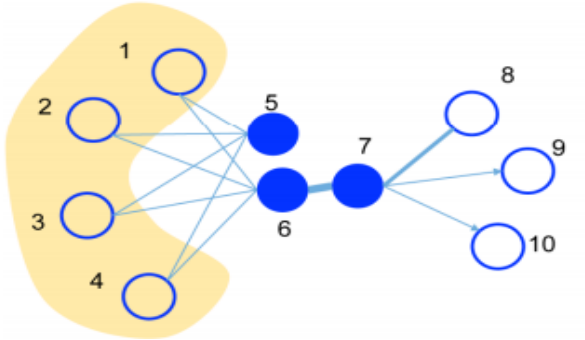
基本概念

- BFS（Breath First Search）广度优先搜索
  - 搜索步骤：标记所有路径，最后选择
  - 广度优先



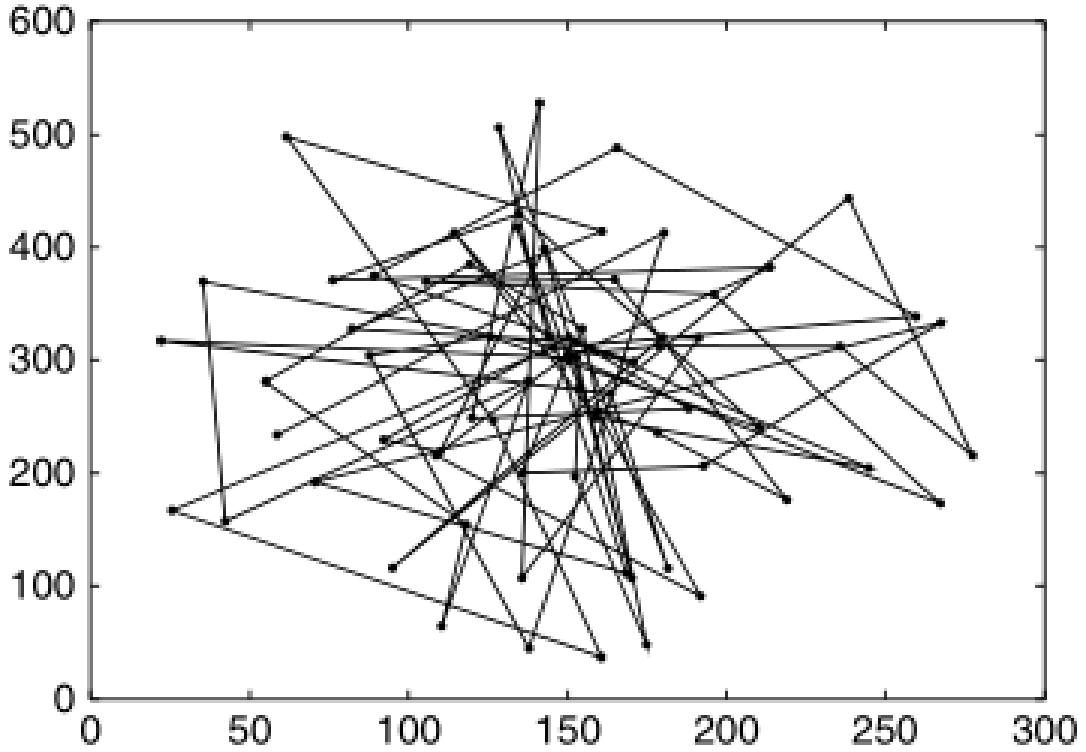
基本概念

- 一阶相似度
  - 邻接矩阵A中的一行 $A_i = \{a_{i,1}, a_{i,2}, \cdots, a_{i,|V|}\}$ 表示v_i与其他顶点之间的一阶相似度。如果权重$a_{i,j} > 0$，则v_i和v_j之间存在正的一阶相似度，权重越高，两个节点越相似。如果节点之间没有连接，一阶相似度为0，即权重$a_{i,j} = 0$ 。
- 二阶相似度
  - 二阶相似度描述了一对节点的邻域结构的接近程度。
  - 节点 v_i和v_j之间的二阶相似度定义为：A_i和A_j之间的相似性。
  - 两个节点相同的用户越多，这两个节点越相似.



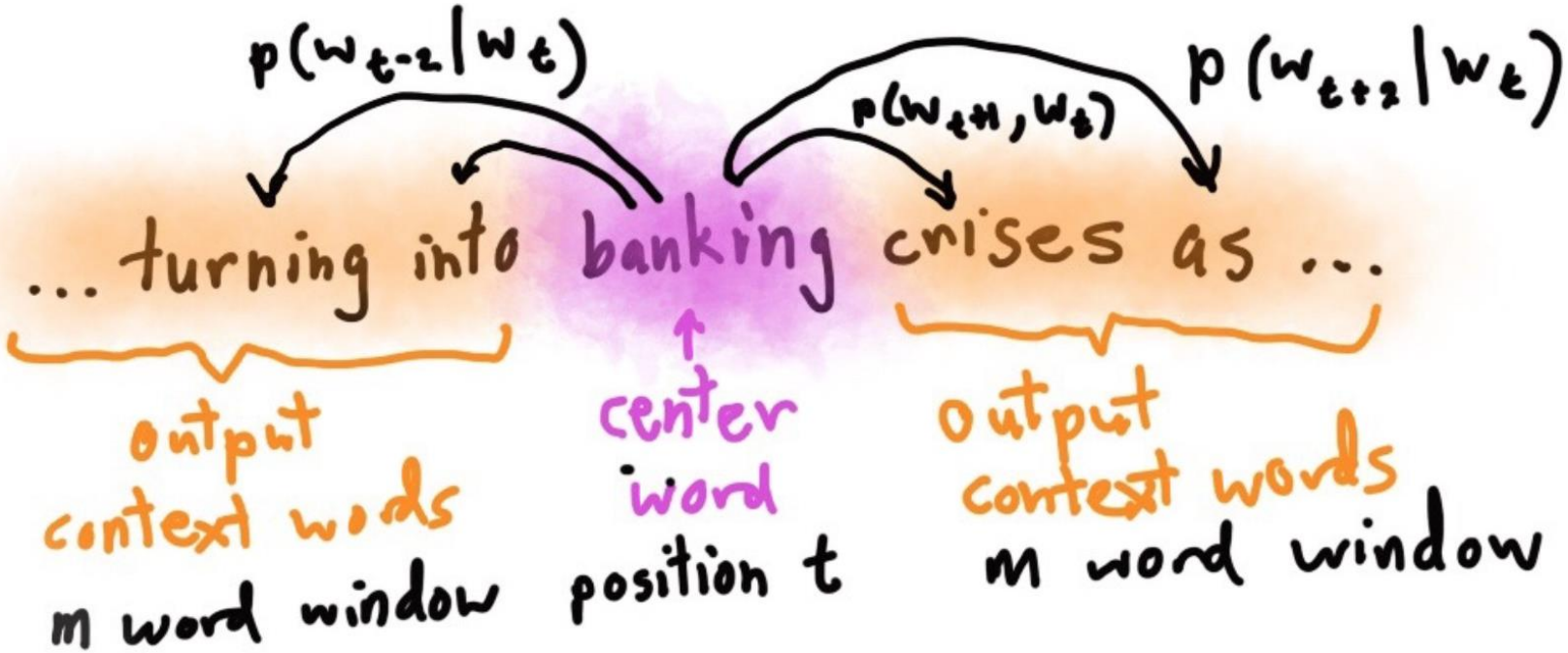
基本概念

- 随机游走
  - 基于过去的表现，无法预测将来的发展步骤和方向。



基本概念

- SkipGram模型
  - 通过中心词预测上下文，该模型给出了给定中心词后，上下文中某个词出现的概率，即图中的$P(w_{t-2}|w_t)$, $P(w_{t-1}|w_t)$, $P(w_{t+1}|w_t)$, $P(w_{t+2}|w_t)$, SkipGram要做的事情就是最大化这些概率。



算法原理

算法原理

T	将网络中的节点表示成低维、实值、稠密的向量形式
I	网络拓扑结构
P	网络嵌入方法
O	节点向量表示

P	尽可能利用图的拓扑结构信息
C	具备网络拓扑结构
D	1.高阶网络嵌入 2.网络外部信息的引入应用
L	KDD，WWW

算法原理

- 谱聚类方法
  - 局部线性表示 (locally linear embedding)
    - 局部线性表示假设一个节点和它邻居的表示都位于该流形的一个局部线性的区域。一个节点的表示可以通过它的邻居节点的表示的线性组合来近似得到。
  - Laplace 特征表 (Laplace eigenmap)
    - 通过平滑项的方式，使得原始空间中两个相似的节点，在低维的向量空间中有近似的表示。
  - 有向图表示 (directed graph embedding)
    - 进一步扩展了 Laplace 特征表方法, 给不同点的损失函数以不同的权重. 其中点的权重是由基于随机游走的排序方法来决定。

算法原理

- DeepWalk

模型	目标	输入	输出
Word2vec	单词	句子	词向量
DeepWalk	节点	节点序列	节点向量

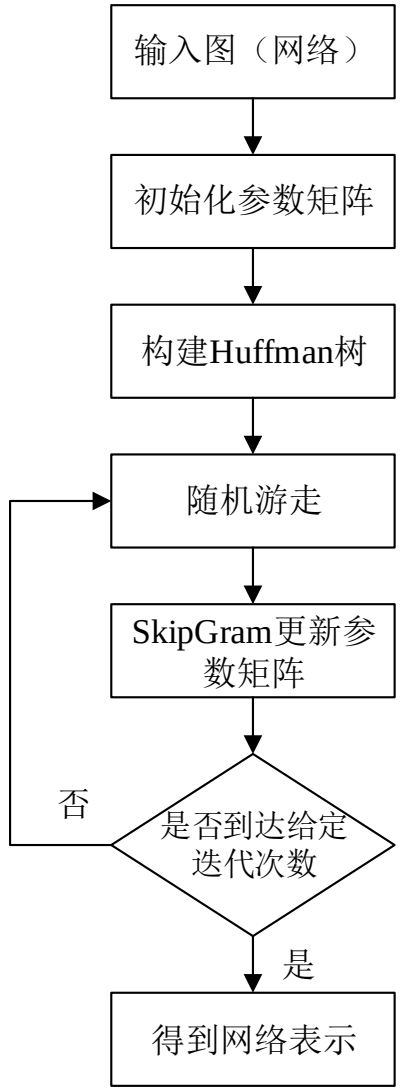
算法原理

- DeepWalk

Algorithm 1 DeepWalk(G, w, d, γ, t)

Input: graph $G(V, E)$
window size w
embedding size d
walks per vertex γ
walk length t
Output: matrix of vertex representations $\Phi \in \mathbb{R}^{|V| \times d}$

1: Initialization: Sample Φ from $\mathcal{U}^{|V| \times d}$
2: Build a binary Tree T from V
3: **for** $i = 0$ to γ **do**
4: $\quad \mathcal{O} = \text{Shuffle}(V)$
5: $\quad$ **for each** $v_i \in \mathcal{O}$ **do**
6: $\quad\quad \mathcal{W}_{v_i} = RandomWalk(G, v_i, \text{t})$
7: $\quad\quad \text{SkipGram}(\Phi, \mathcal{W}_{v_i}, w)$
8: $\quad$ **end for**
9: **end for**

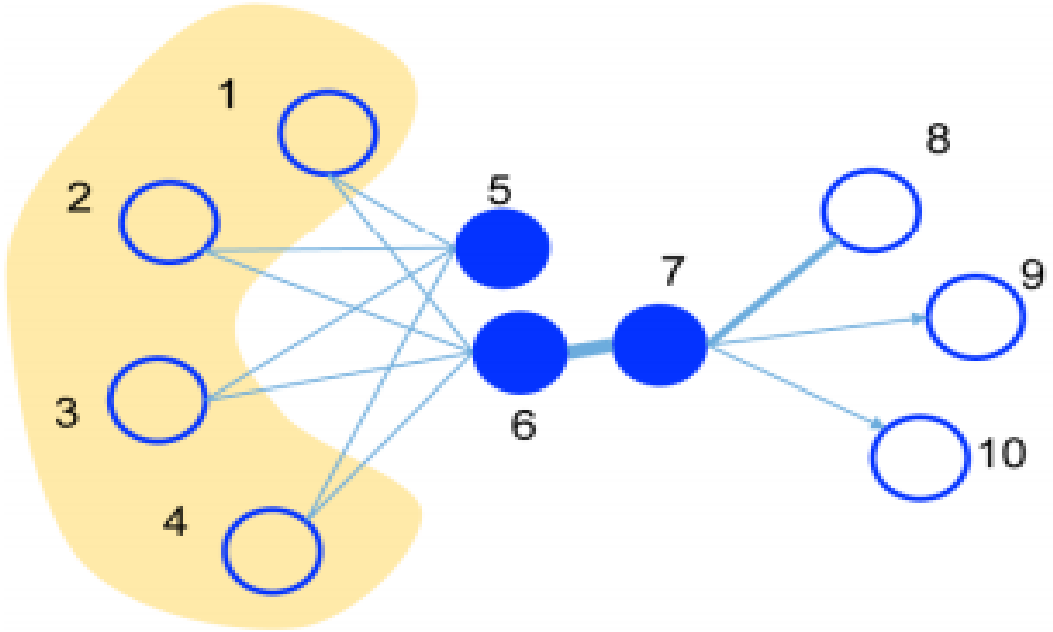


算法原理

- DeepWalk
  - 在线学习：DeepWalk是可扩展的
  - 容易实现并行性。几个随机游走者（不同的线程，进程或机器）可以同时探索同一网络的不同部分。
  - 适应性。当图变化后，不需要全局重新计算，可以迭代地更新学习模型

算法原理

- LINE
  - 可适应任意类型的网络：有向、无向、有权、无权。
  - 采用一阶相似度和二阶相似度结合

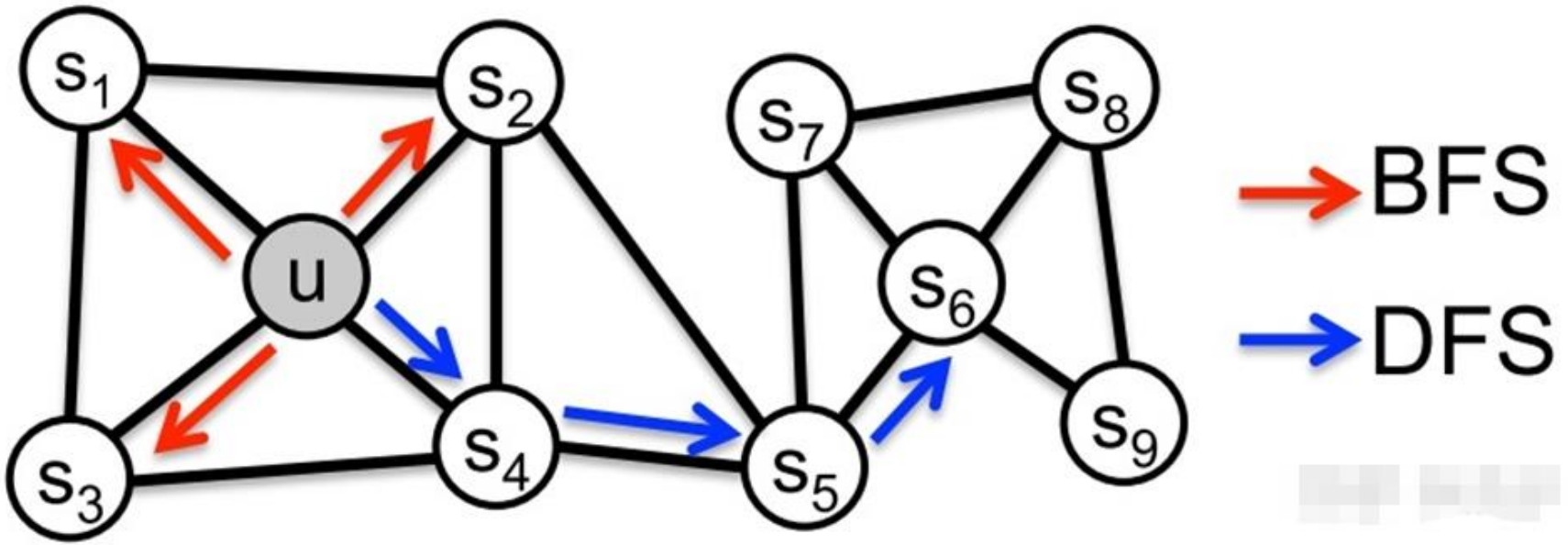


算法原理

- 一阶相似度
  - 两个相邻节点之间的关系
- 二阶相似度
  - 一个节点维护两个嵌入向量
  - 一个本身的表示向量
  - 一个作为其他顶点的上下文节点的表示向量
- 负采样
  - 二阶相似度计算，涉及Softmax函数，计算量大
- 边采样
  - 边采样优化方法解决了SGD的局限性（边的权值变换很大时，学习率难以选择，并且权值和乘以梯度导致梯度爆炸）在信息较少的稀疏网络表现优越。

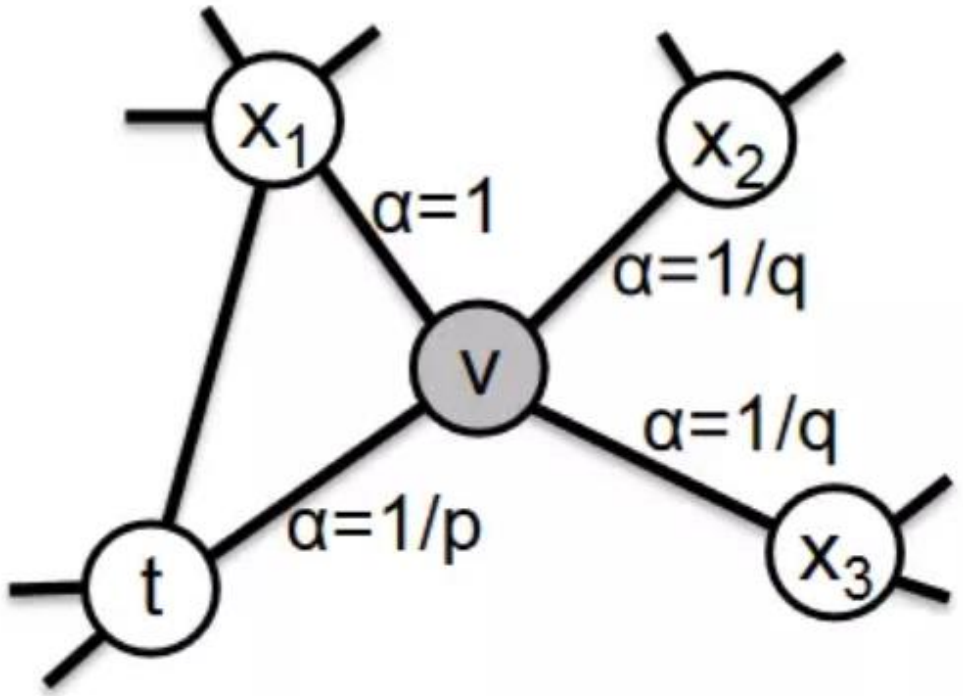
算法原理

- Node2vec
  - 同质性（homophily）
  - 结构一致性（structual equivalence）



算法原理

- Biased random walk



$$\pi_{vx} = \alpha_{\mathrm{pq}}(t,x) \cdot \omega_{vx}$$

$$\alpha_{pq}(t,x) = \begin{cases} \frac{1}{p} & \text{If } d_{tx} = 0 \\ 1 & \text{If } d_{tx} = 1 \\ \frac{1}{q} & \text{If } d_{tx} = 2 \end{cases}$$

算法原理

- Node2vec

Algorithm 1 The *node2vec* algorithm.

LearnFeatures (Graph $G = (V, E, W)$, Dimensions d, Walks per node r, Walk length l, Context size k, Return p, In-out q)
π = PreprocessModifiedWeights(G, p, q)
$G' = (V, E, \pi)$
Initialize $walks$ to Empty
for $iter = 1$ **to** r **do**
  for all nodes $u \in V$ **do**
    $walk$ = node2vecWalk(G', u, l)
    Append $walk$ to $walks$
f = StochasticGradientDescent(k, d, $walks$)
return f

算法原理

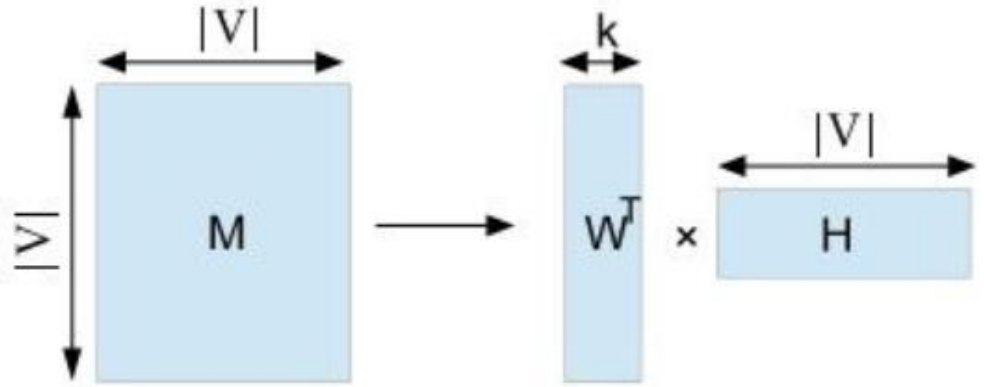
- 算法对比
  - DeepWalk：只可用于无权图，采取完全随机的随机游走策略，只考虑了二阶相似度
  - LINE：可以用于有向、无向、有权、无权图，同时考虑一阶相似度和二阶相似度
  - Node2vec：结合BFS和DFS的随机游走策略

算法原理

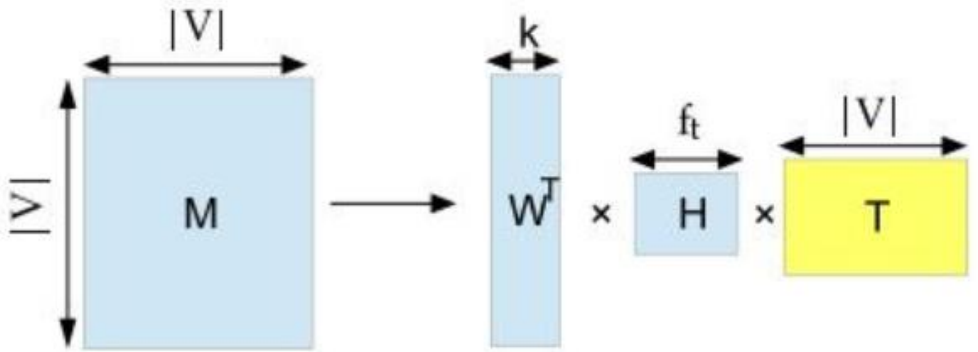
- 结合外部信息的网络嵌入
  - TADW：结合节点的文本信息
  - MMDW：结合节点的标签类别信息
  - TransNet：结合节点和节点间的标签信息

算法原理

- TADW



DeepWalk模型



TADW模型

应用总结

应用总结

- 算法的应用领域
  - 节点分类
  - 链接预测
  - 社区发现
- 未来发展
  - 动态网络嵌入

参考文献

[1] 涂存超, 杨成, 刘知远,等. 网络表示学习综述[J]. 中国科学:信息科学, 2017(8).
[2] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. KDD2014.
[3] JianTang, Meng Qu , Mingzhe Wang, Ming Zhang, Jun Yan, Qiaozhu Mei. LINE：Large-scaleInformation Network Embedding . WWW2015.
[4] A Grover, J Leskovec . node2vec: Scalable Feature Learning for Networks. KDD2016.

道德经

大成若缺，其用不弊。
大盈若冲，其用不穷。
大直若屈。大巧若拙。
大辩若讷。静胜躁，寒
胜热。清静为天下正。

谢谢！