网 络 表 征 学 习

(Network Representation Learning / Graph Embedding)

——数据挖掘中的机器学习

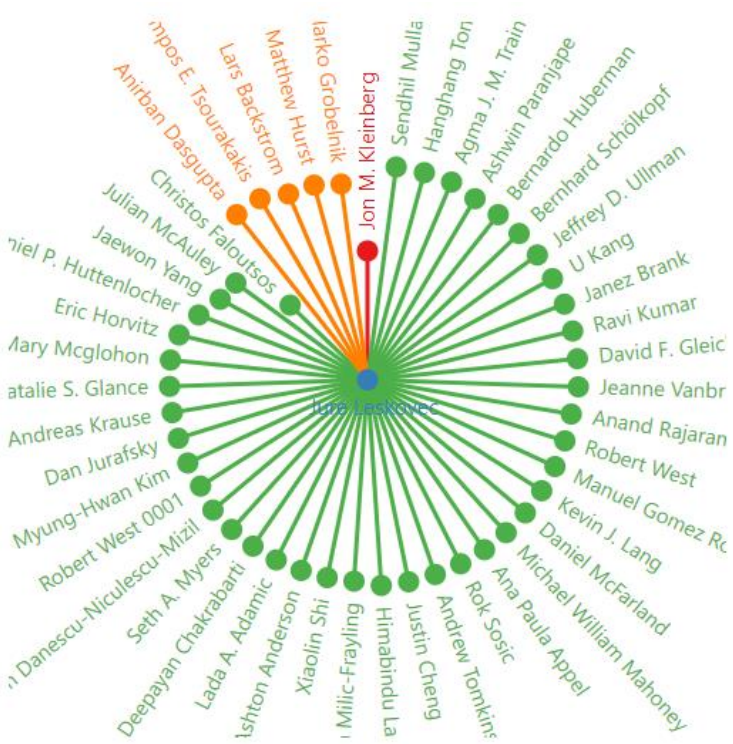
分享人：刘猛

通信网络和社会网络

计算机网络 (computer network) 和数据挖掘 (data mining) 两个领域都有网络的概念，它们有什么区别?

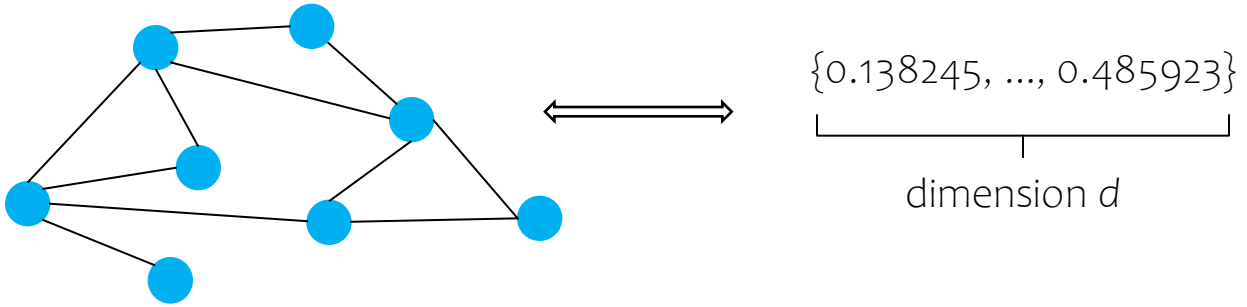
社会网络包含什么？

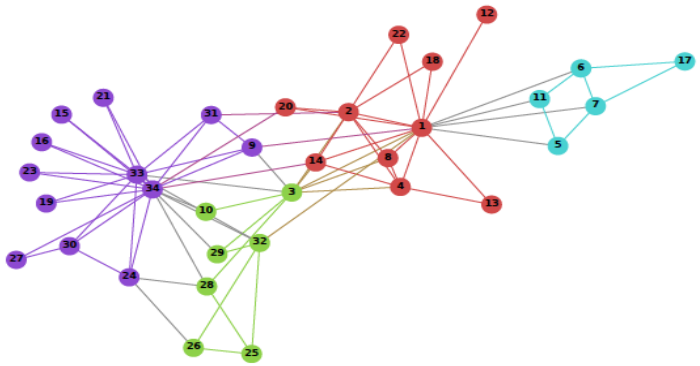
学术合作网络、购物消费网络、软件聊天网络等等，这些都属于社会网络，其根本标志在于以人为中心。

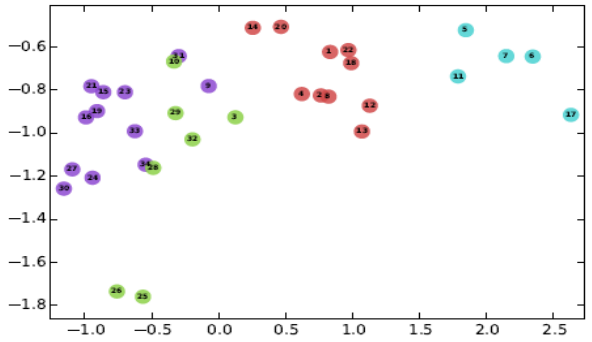


NRL是什么?

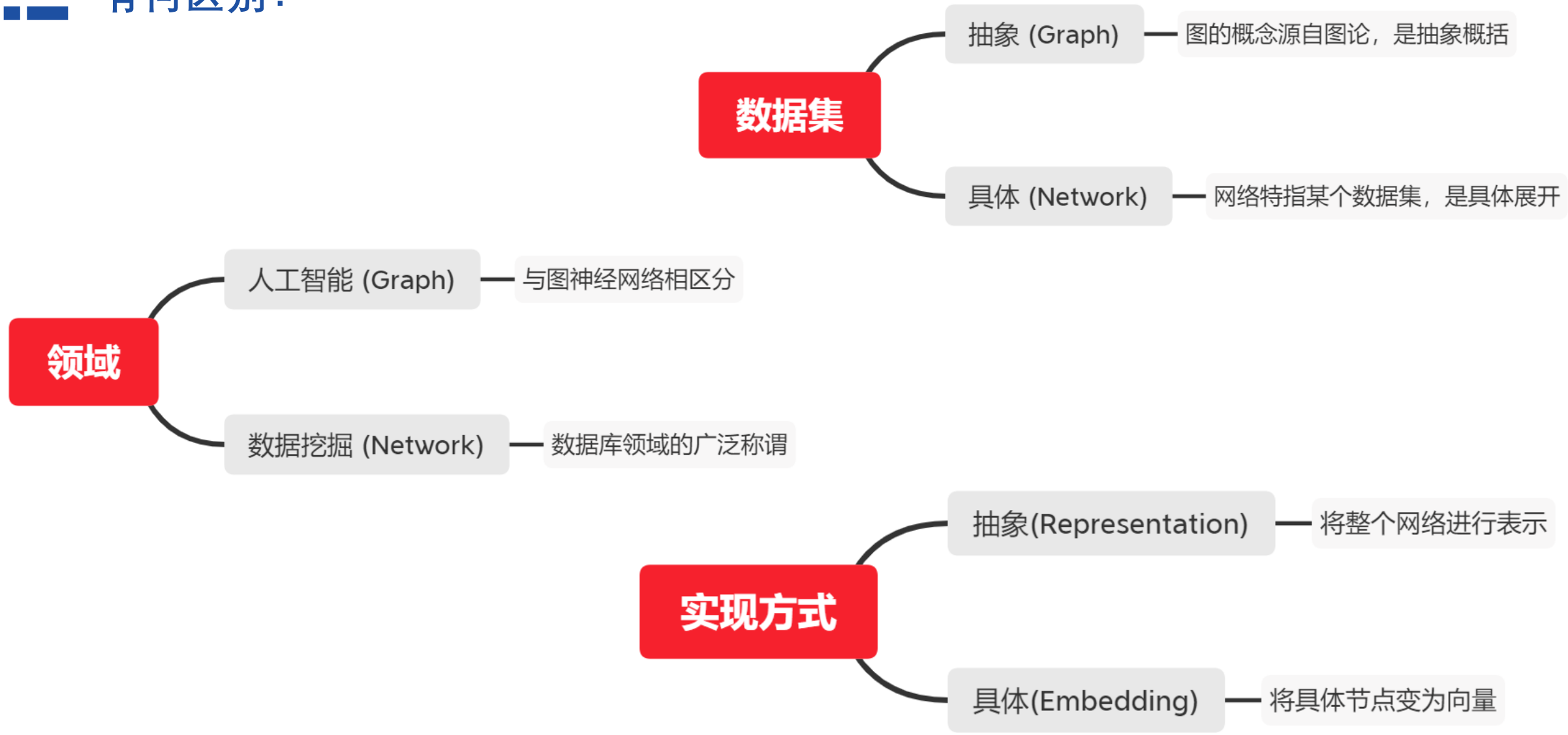
网络表征学习(Network Representation Learning)，又称图嵌入(Graph Embedding)，致力于将稀疏的高维网络结构数据用稠密的低维向量进行表示。





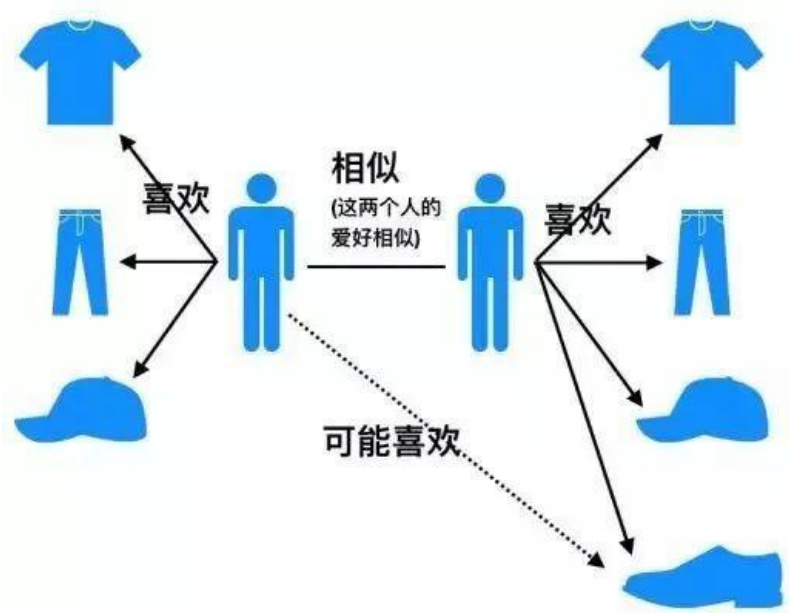


有何区别?

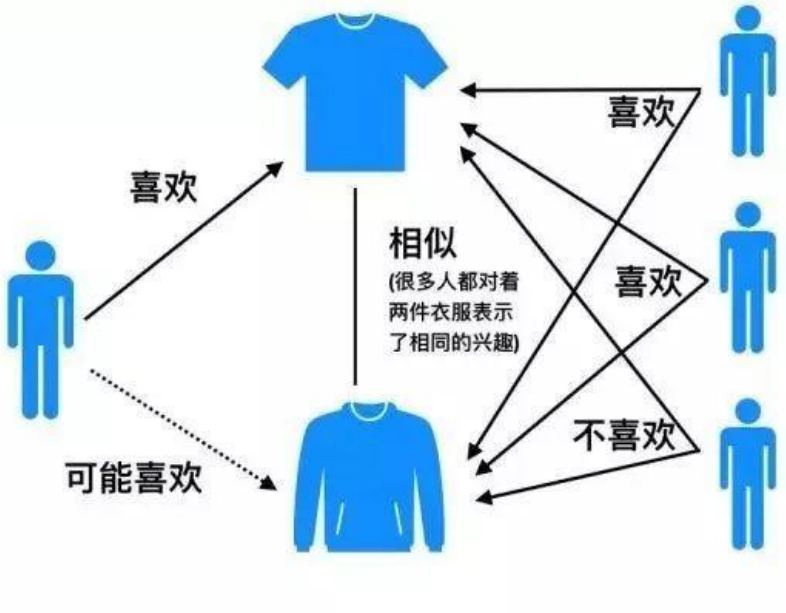

数据集
抽象 (Graph)
图的概念源自图论，是抽象概括
具体 (Network)
网络特指某个数据集，是具体展开
领域
人工智能 (Graph)
与图神经网络相区分
数据挖掘 (Network)
数据库领域的广泛称谓
实现方式
抽象(Representation)
将整个网络进行表示
具体(Embedding)
将具体节点变为向量

能做什么?

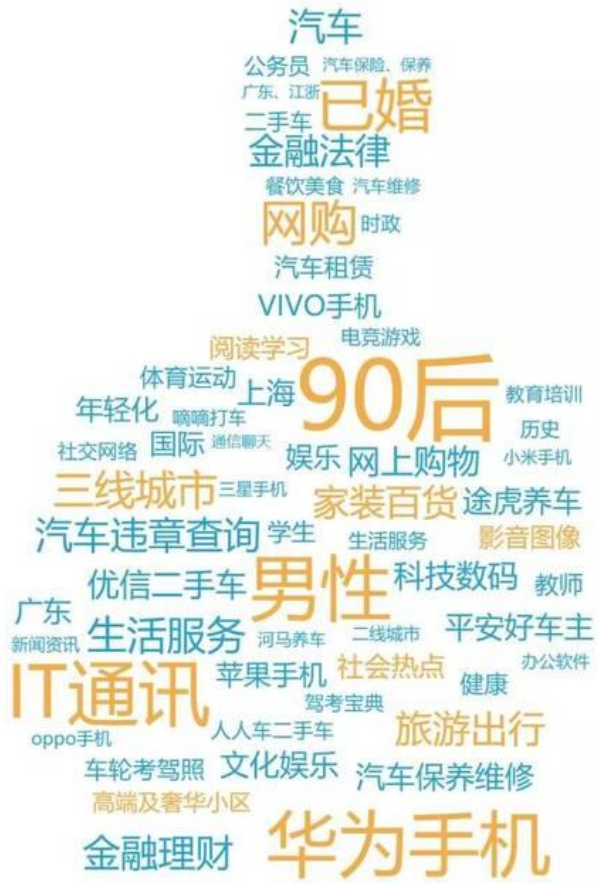
NRL能够捕捉网络的演变规律和个人行为模式，对现有的信息进行分类，对未来的交互进行预测，可以应用在兴趣推荐、用户画像等实际项目中。



“人以群分”的基于用户的协同过滤



“物以类聚”的基于物品的协同过滤



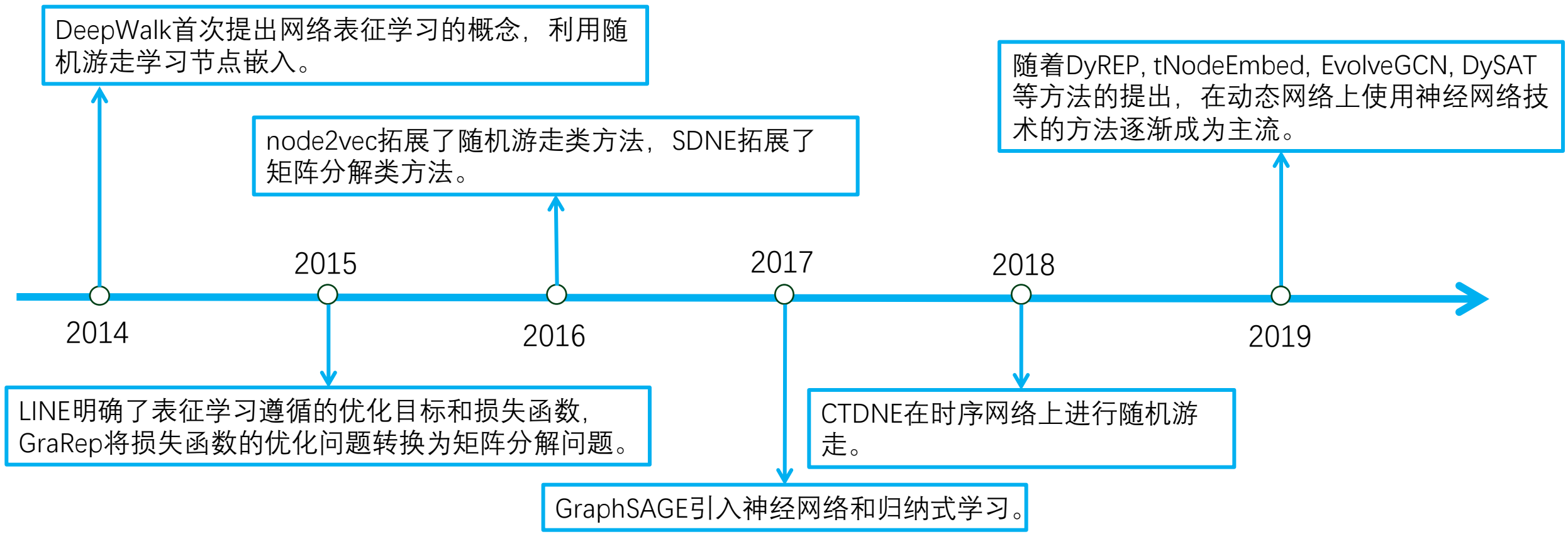
谁在做?

Jure Leskovec，数据挖掘领域影响力top-2的80后青年学者，斯坦福大学教授，Geogle引用量10万余次，H指数125，在NRL方向提出node2vec和GraphSAGE等影响深远的算法。

Steven Skiena，美国石溪大学杰出教授，Geogle引用量2万余次，H指数63，《算法设计手册》作者，提出DeepWalk算法开创了NRL方向。

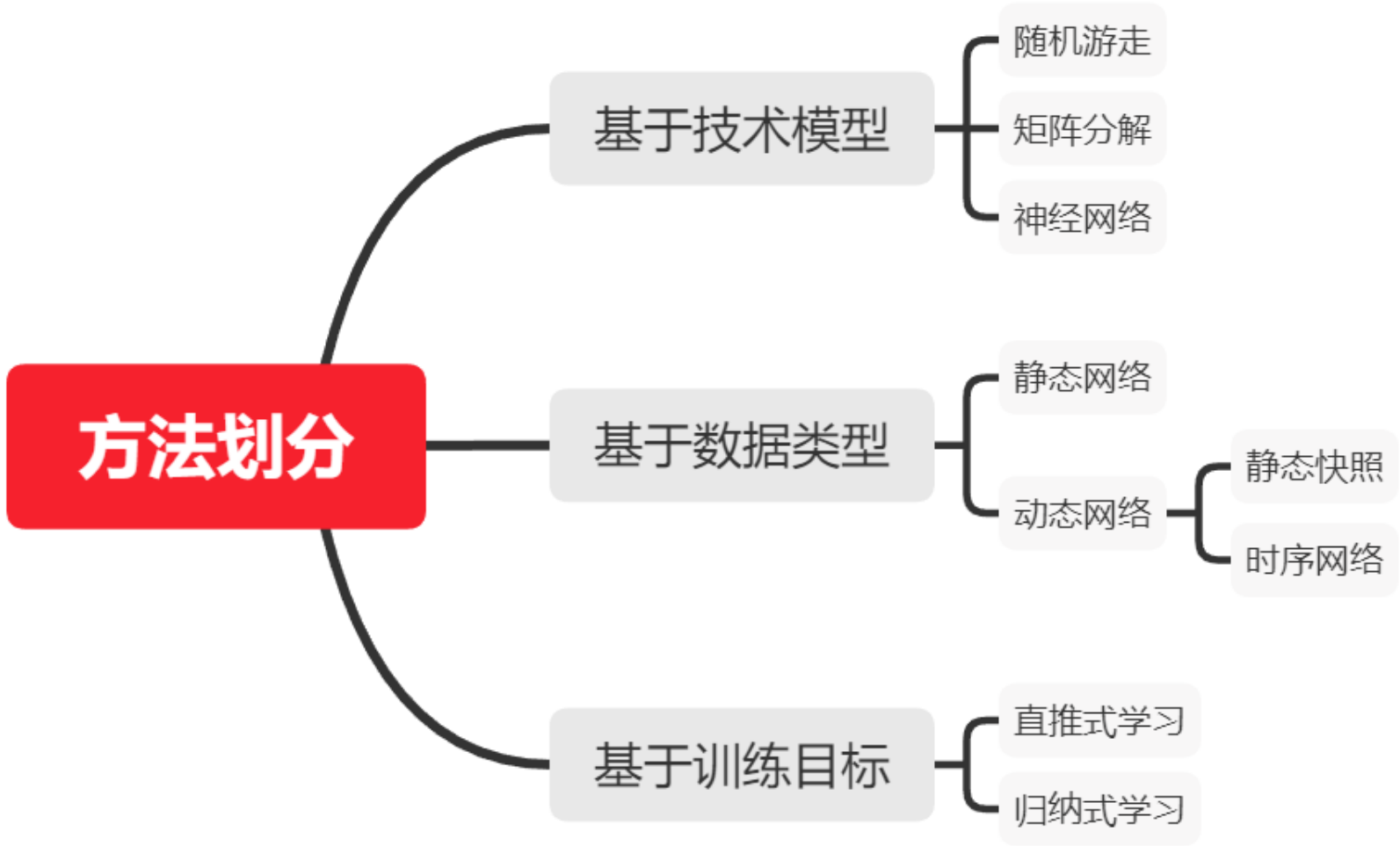
唐杰，清华大学计算机系副主任，国家杰青，Geogle引用量2万余次，H指数75，IEEE Transactions on Big Data期刊主编，开发了AMiner等数据挖掘平台。

发展历程

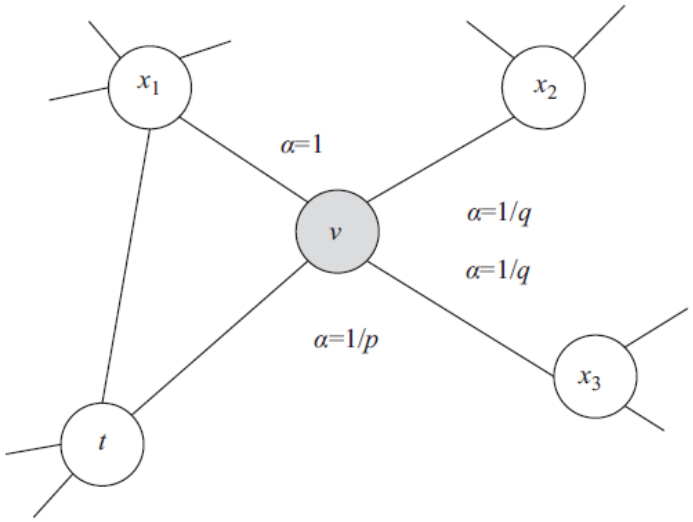

DeepWalk首次提出网络表征学习的概念，利用随机游走学习节点嵌入。
node2vec拓展了随机游走类方法，SDNE拓展了矩阵分解类方法。
随着DyREP, tNodeEmbed, EvolveGCN, DySAT等方法的提出，在动态网络上使用神经网络技术的方法逐渐成为主流。
2014
2015
2016
2017
2018
2019
LINE明确了表征学习遵循的优化目标和损失函数，GraRep将损失函数的优化问题转换为矩阵分解问题。
CTDNE在时序网络上进行随机游走。
GraphSAGE引入神经网络和归纳式学习。

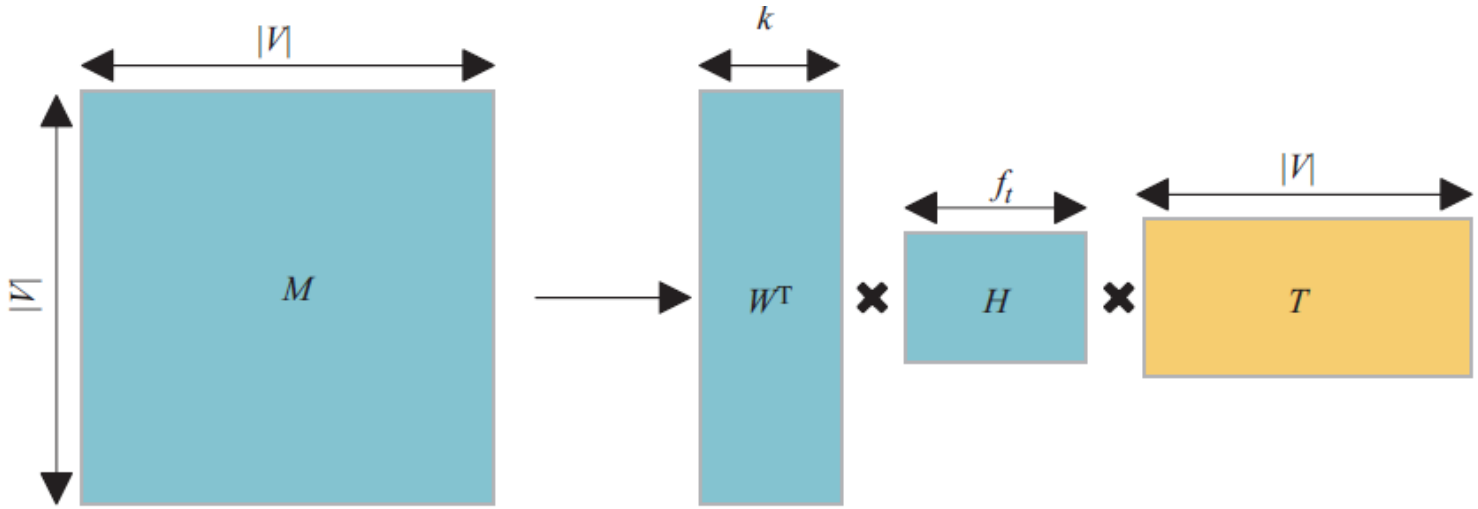
如何分类？

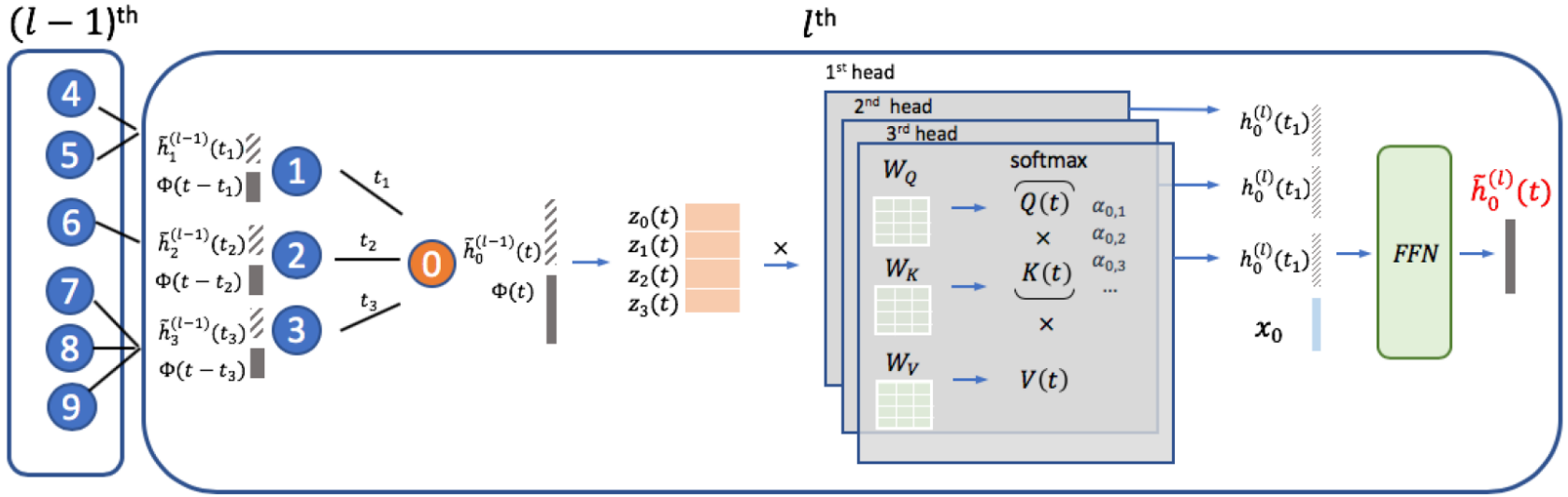
网络表征学习代表的是一类对网络结构数据进行挖掘的抽象思路，在具体实现上需要与现今数据挖掘和机器学习的实际方法相结合。



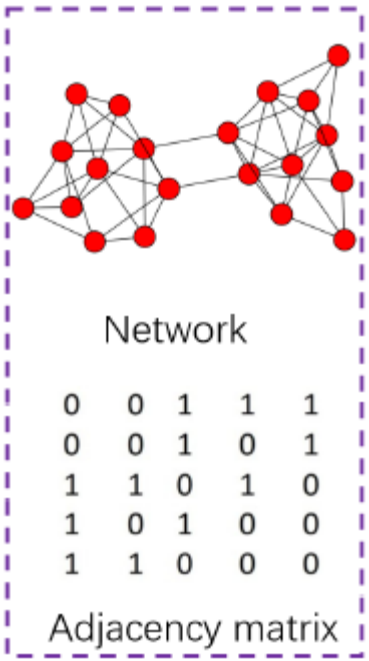
基于技术模型分类


x_1
x_2
x_3
v
t
α=1
α=1/q
α=1/q
α=1/p


$|V|$
k
$|V|$
M
W^T
f_t
H
$|V|$
T


$(l-1)^{\text{th}}$
l^{th}
4
5
6
7
8
9
1
2
3
0
t_1
t_2
t_3
$\tilde{h}_1^{(l-1)}(t_1)$
$\Phi(t-t_1)$
$\tilde{h}_2^{(l-1)}(t_2)$
$\Phi(t-t_2)$
$\tilde{h}_3^{(l-1)}(t_3)$
$\Phi(t-t_3)$
$\tilde{h}_0^{(l-1)}(t)$
$\Phi(t)$
$z_0(t)$
$z_1(t)$
$z_2(t)$
$z_3(t)$
×
1st head
2nd head
3rd head
W_Q
W_K
W_V
softmax
$Q(t)$
$K(t)$
$V(t)$
$\alpha_{0,1}$
$\alpha_{0,2}$
$\alpha_{0,3}$
$h_0^{(l)}(t_1)$
$h_0^{(l)}(t_1)$
$h_0^{(l)}(t_1)$
x_0
FFN
$\tilde{h}_0^{(l)}(t)$

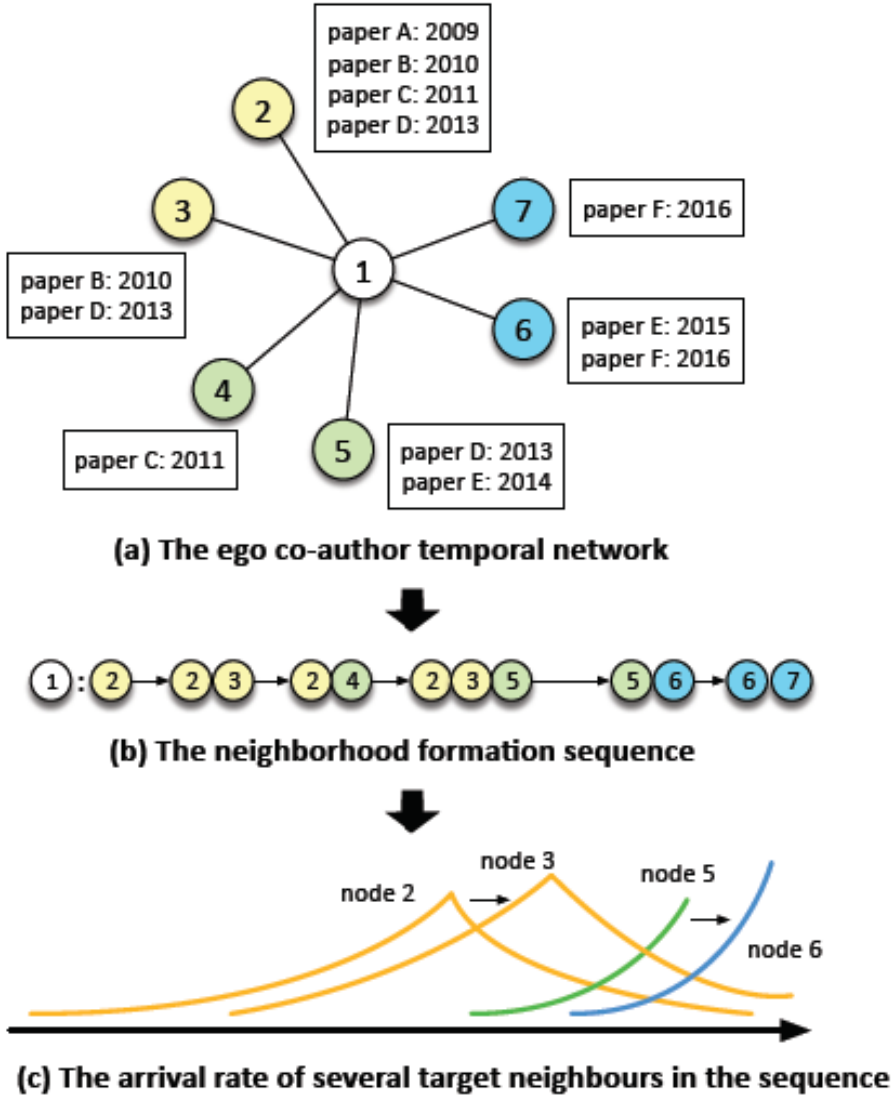
基于数据类型分类



静态图（邻接矩阵）

$$\mathbb{G} = \{\mathcal{G}^1, \ldots, \mathcal{G}^T\}$$

静态快照（多个静态图）



时序图（交互序列）

基于训练目标分类

Transductive（直推式）

结合现有数据反复训练，得到最终稳定的模型。

优点：性能通常较好，适合线下训练。

缺点：当数据发生增删等变化时，需要对整个模型重新训练。

Inductive（归纳式）

在训练模型的同时，捕捉数据演变的规律。

优点：数据发生变化时，只需修改少部分模型，适合线上训练。

缺点：相比直推式难以收敛，性能不够明朗。

正在关注什么？

截至2022年初，网络表征学习的发展方向已经非常广泛，其与图神经网络、推荐系统等领域也更为密不可分。具体的研究热点可归为如下几类：

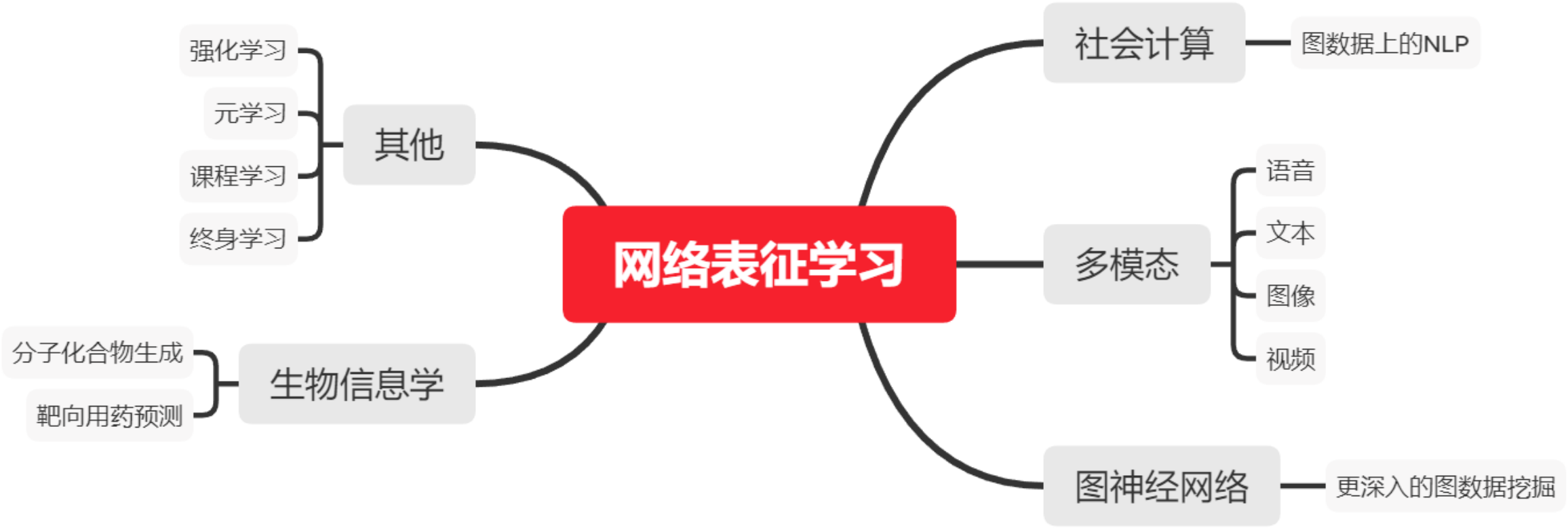
预训练：在大规模数据集上对模型进行预训练 (Pre-Training)，得到较好的参数定位和模型结构，后续用于其他数据集上仅需微调 (Fine-Tuning) 即可，这一思路来自NLP领域。

自监督：现实世界的数据集很难有充分清晰的标签，而使用人工标注的成本又过大。作为无监督学习中特殊的一类，自监督 (Self-Supervised) 从数据本身中提取和构造标签，使得模型对数据的理解更为深入。

统一架构：当前许多方法尝试提出泛化性较强的通用框架，可以搭建在任意模型上。此外，许多研究者还尝试用CNN或Transformer架构来统一NLP和CV的模型架构，这在图数据上也是一个新兴思路。

底层逻辑：一些研究者们开始尝试研究更为细节的内容，比如探究负采样的最优策略，讨论动态的嵌入维度大小对模型的影响，更进一步地，关于图数据模型上的可解释性问题仍饱受关注。

未来要走向何处?


网络表征学习
其他
强化学习
元学习
课程学习
终身学习
生物信息学
分子化合物生成
靶向用药预测
社会计算
图数据上的NLP
多模态
语音
文本
图像
视频
图神经网络
更深入的图数据挖掘

敬 请 指 正 !