从Item2vec到GNN——Embedding在推荐系统中的发展和应用

- 姓名：王喆
- 公司：Roku
- Title：资深机器学习工程师，推荐系统架构负责人

2020 ProCon 万人开发者大会

7月3日-4日

CSDN

关于王喆

- 现任美国流媒体公司Roku资深机器学习工程师，推荐系统架构负责人。
- 曾任Hulu高级研究员，品友互动广告效果算法组负责人。
- 毕业于清华大学计算机系，清华大学KEG实验室学术搜索引擎AMiner早期贡献者。
- 主要研究方向为推荐系统、计算广告，发表相关领域学术论文和专利10余项，曾担任DLP-KDD联合主席，KDD、CIKM等国际会议审稿人。
- CTRmodel，Ad-papers 等开源项目发起人和主要贡献者，4.5k stars+。
- 著有《百面机器学习》，《深度学习推荐系统》等技术书籍，读者5万+。
- 知乎专栏/微信公众号：王喆的机器学习笔记

从Item2vec到GNN—Embedding在推荐系统中的发展和应用

什么是Embedding？

基于序列数据的Embedding方法

- Word2vec
- Item2vec

基于Random Walk的Graph Embedding方法

- DeepWalk
- Node2vec

图神经网络

- GraphSAGE
- PinSAGE

Embedding技术在推荐系统中的落地

从Item2vec到GNN—Embedding在推荐系统中的发展和应用

什么是Embedding？

基于序列数据的Embedding方法

- Word2vec
- Item2vec

基于Random Walk的Graph Embedding方法

- DeepWalk
- Node2vec

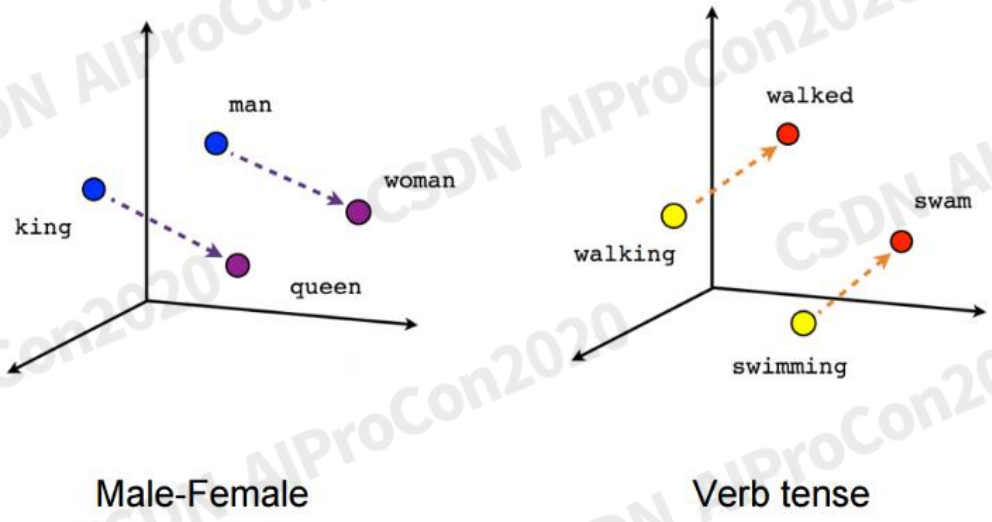
图神经网络

- GraphSAGE
- PinSAGE

Embedding技术在推荐系统中的落地

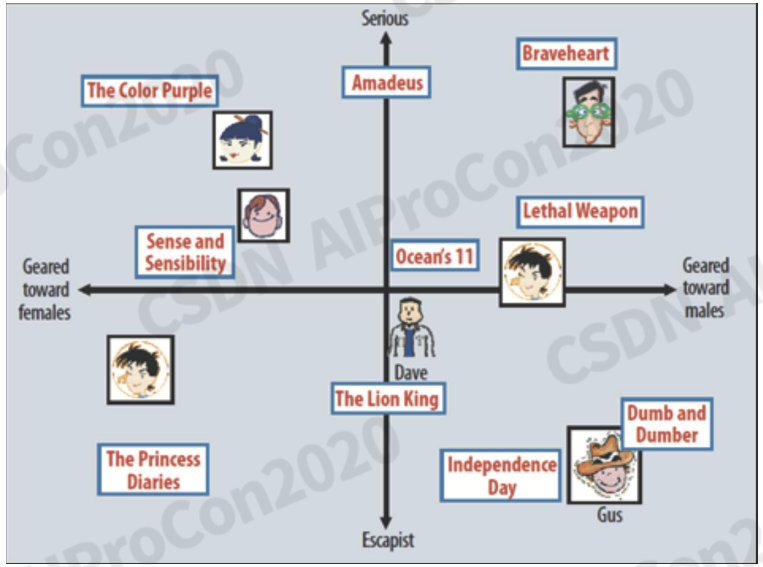
什么是Embedding?

Embedding是一个客观事物到某向量空间的映射，不同Embedding向量间的距离保存了客观事物间的关系信息。



Male-Female　　Verb tense

词向量



视频-用户向量

从Item2vec到GNN—Embedding在推荐系统中的发展和应用

什么是Embedding？

基于序列数据的Embedding方法

- Word2vec
- Item2vec

基于Random Walk的Graph Embedding方法

- DeepWalk
- Node2vec

图神经网络

- GraphSAGE
- PinSAGE

Embedding技术在推荐系统中的落地

基于序列数据的Embedding方法——Word2vec

从 | Item2vec | 到 | GNN | —— | Embedding | 在 | 推荐系统 | 中的 | 发展 | 和 | 应用 |

从 | Item2vec | 到 | GNN | —— | Embedding | 在 | 推荐系统 | 中的 | 发展 | 和 | 应用 |

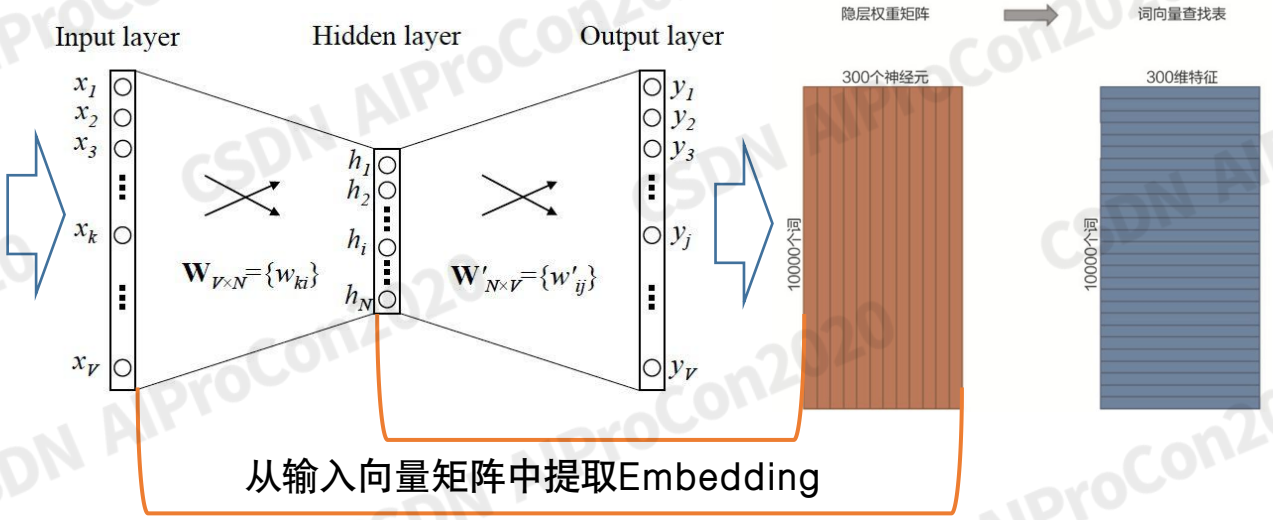
基于Skip-gram架构的样本

Embedding -> GNN，推荐系统

推荐系统 -> Embedding，发展

发展 -> 推荐系统，应用

……

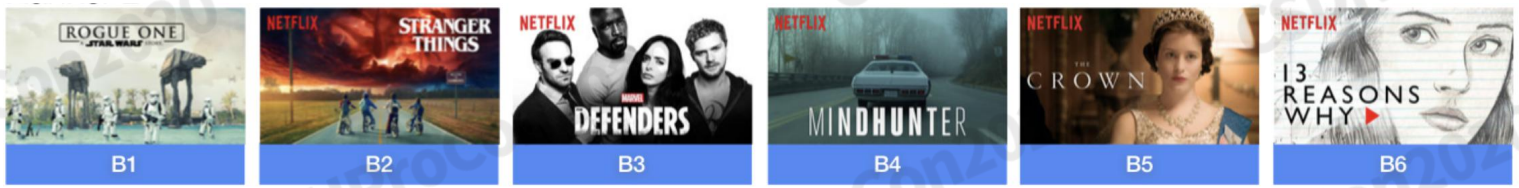


基于序列数据的Embedding方法——Item2vec

从文本序列到一般序列数据的推广

文本序列：从 | Item2vec | 到 | GNN | —— | Embedding | 在 | 推荐系统 | 中的 | 发展 | 和 | 应用 |

观影序列：

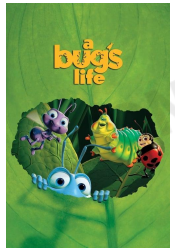


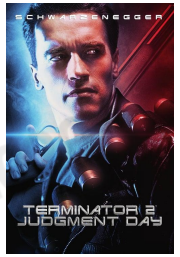
购物序列：

从Word2vec到Item2vec的推广

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leqslant j\leqslant c, j\neq 0}\log p\left(w_{t+j}|w_t\right) \Rightarrow \frac{1}{K}\sum_{i=1}^{K}\sum_{j\neq i}^{K}\log p\left(w_j|w_i\right)$$

基于序列数据的Embedding方法——Movie2vec的例子





从Item2vec到GNN—Embedding在推荐系统中的发展和应用

什么是Embedding？

基于序列数据的Embedding方法

- Word2vec
- Item2vec

基于Random Walk的Graph Embedding方法

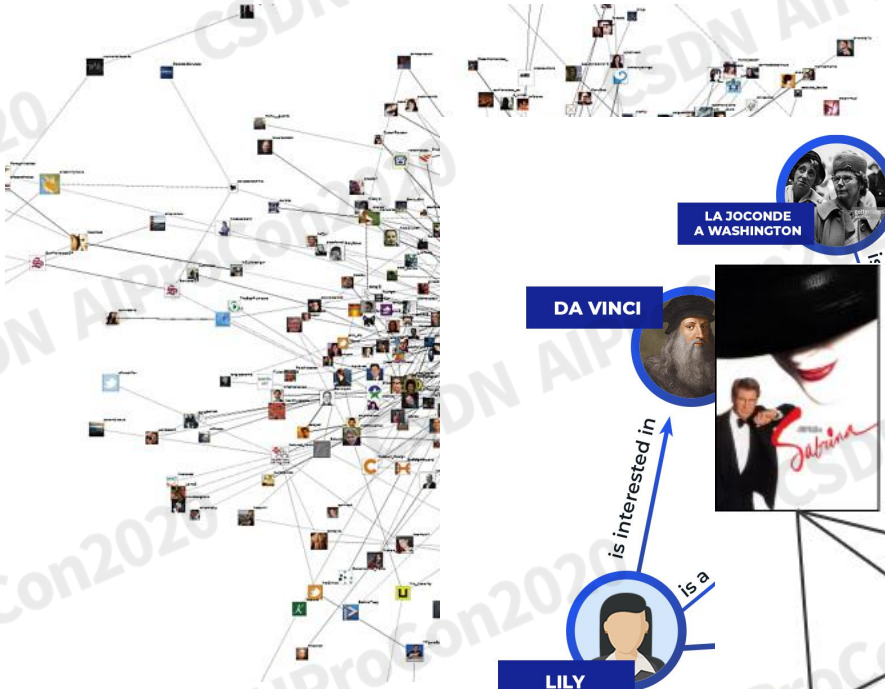
- DeepWalk
- Node2vec

图神经网络

- GraphSAGE
- PinSAGE

Embedding技术在推荐系统中的落地

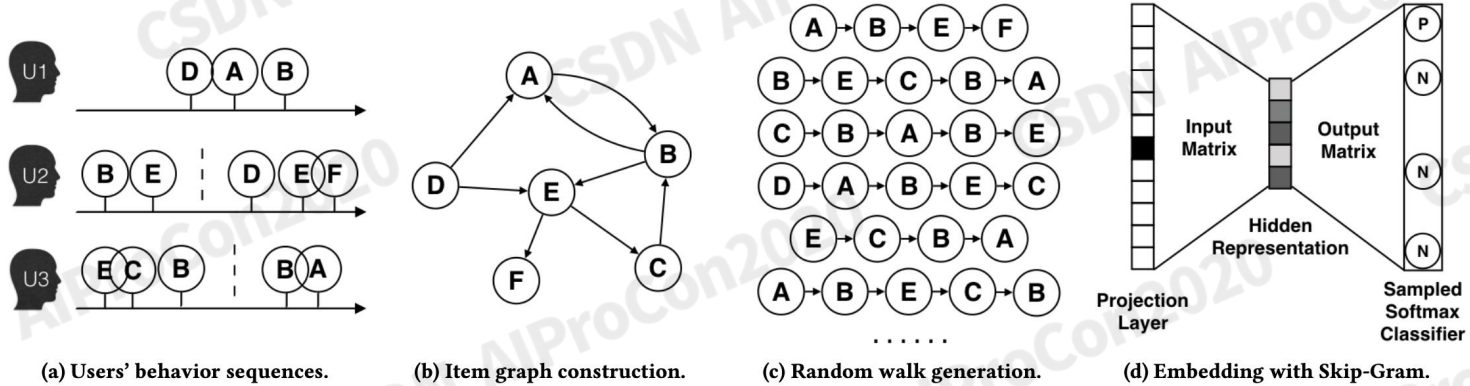
互联网数据更多是以图的形式存在



社交网络

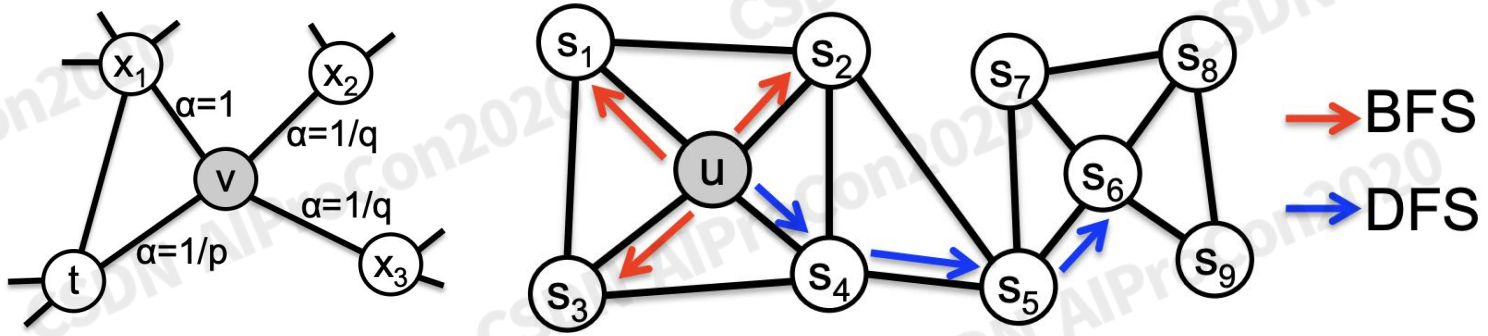
行为关系图

基于Random Walk的Graph Embedding方法



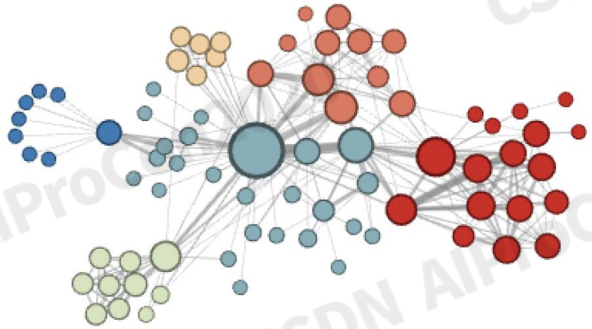
(a) Users' behavior sequences. (b) Item graph construction. (c) Random walk generation. (d) Embedding with Skip-Gram.

DeepWalk



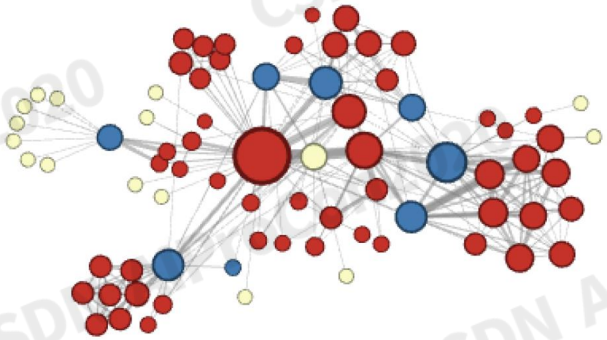
Node2vec

基于Random Walk的Graph Embedding方法



DFS

节点的同质性

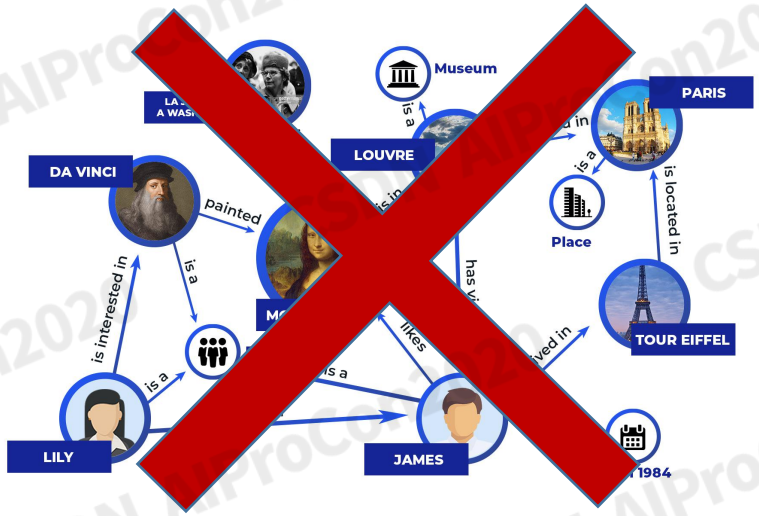


BFS

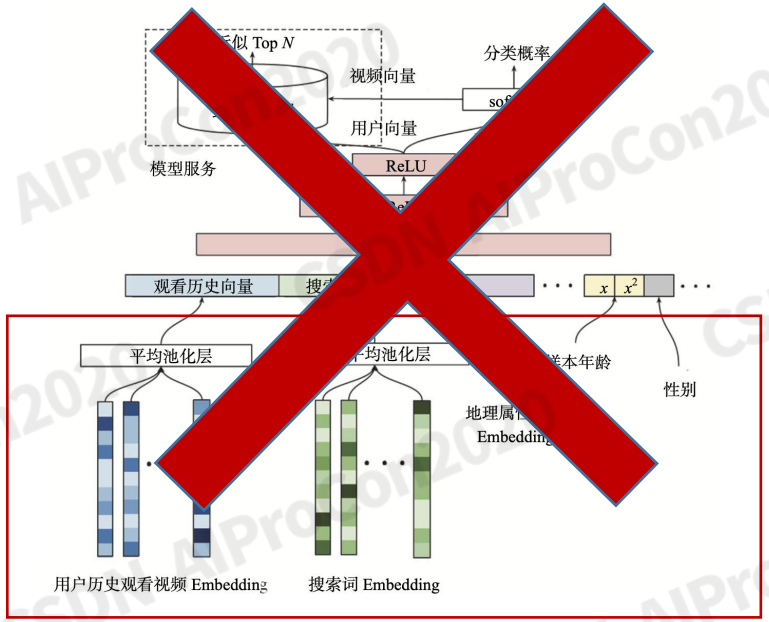
网络的结构性

基于Random Walk的Graph Embedding方法的局限性

- 没有真正处理图结构



- 无法直接引入其他特征



从Item2vec到GNN—Embedding在推荐系统中的发展和应用

什么是Embedding？

基于序列数据的Embedding方法

- Word2vec
- Item2vec

基于Random Walk的Graph Embedding方法

- DeepWalk
- Node2vec

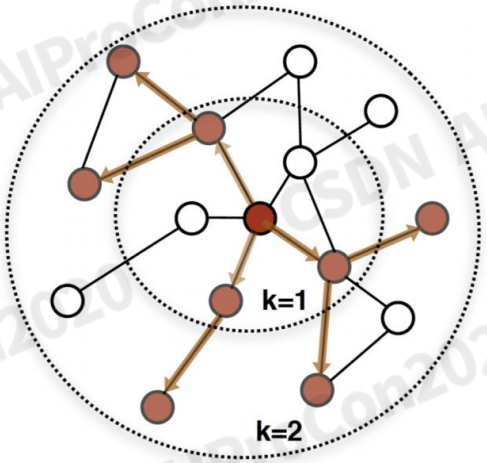
图神经网络

- GraphSAGE
- PinSAGE

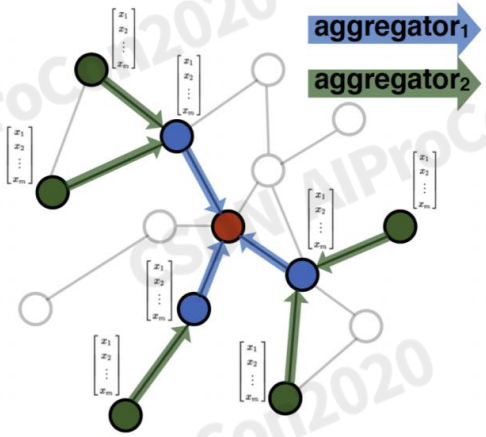
Embedding技术在推荐系统中的落地

图神经网络（Graph Neural Network）—GraphSAGE

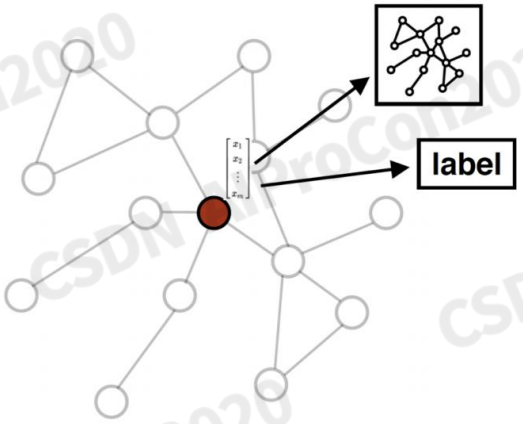
GraphSAGE — Graph Sample and Aggregate
PinSAGE — Pinterest GraphSAGE



1. Sample neighborhood



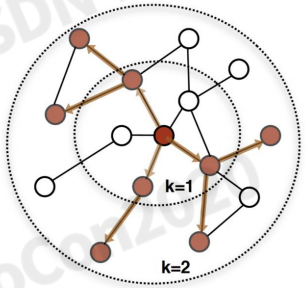
2. Aggregate feature information from neighbors



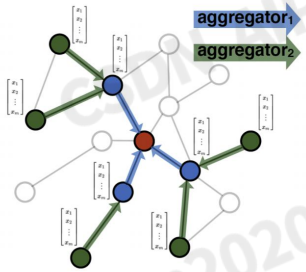
3. Predict graph context and label using aggregated information

图神经网络（Graph Neural Network）

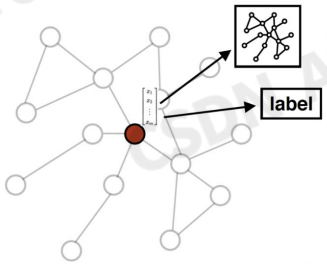
从原始数据
到形成GNN样本子图



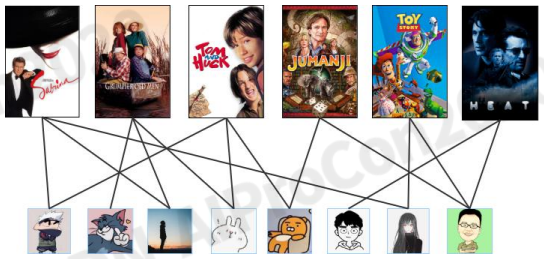
1. Sample neighborhood

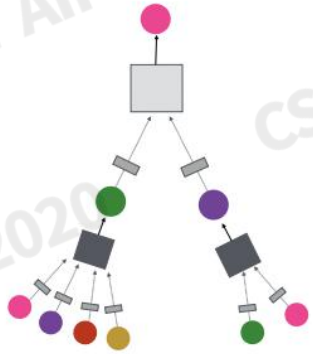


2. Aggregate feature information from neighbors

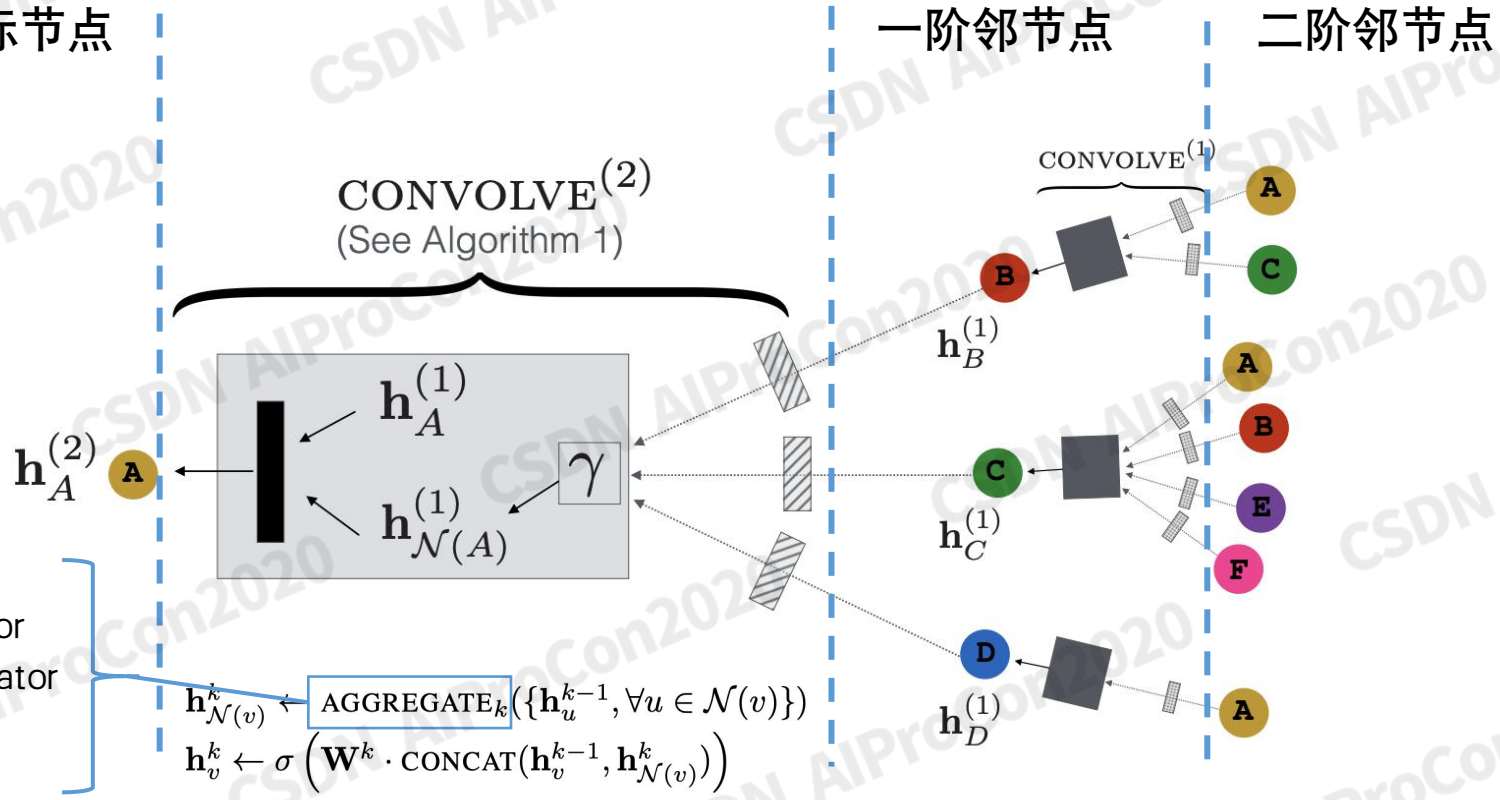


3. Predict graph context and label using aggregated information



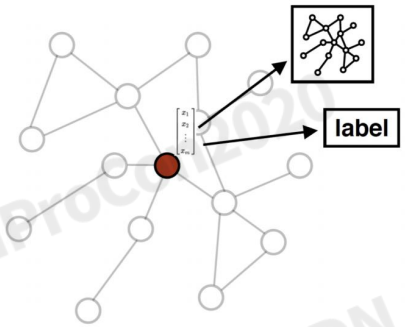


图神经网络（Graph Neural Network）—GNN结构



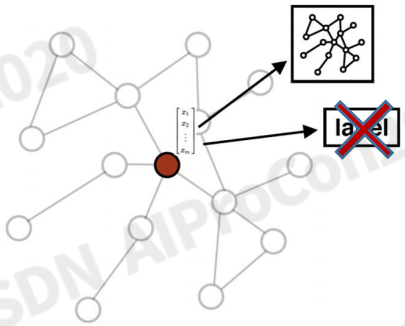
Mean Aggregator
Pooling Aggregator
Attention Aggregator
LSTM Aggregator
MLP Aggregator

GNN的无监督与有监督学习



分类问题

GNN以Logistics regression或Softmax作为输出层

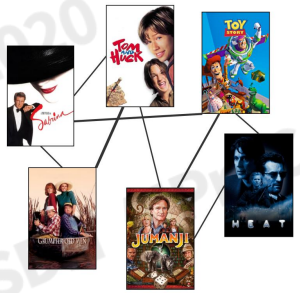


尽量让邻接节点的Embedding相似

$$J_{\mathcal{G}}(\mathbf{z}_u) = -\log\left(\sigma(\mathbf{z}_u^\top \mathbf{z}_v)\right) - Q \cdot \mathbb{E}_{v_n \sim P_n(v)} \log\left(\sigma(-\mathbf{z}_u^\top \mathbf{z}_{v_n})\right)$$

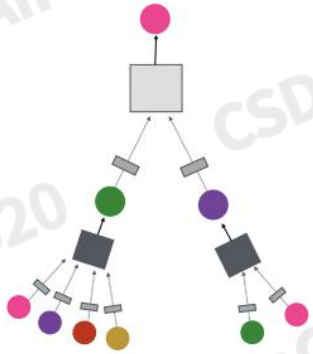
PinSAGE的工程技巧

1. 在邻节点获取中，采用基于Random Walk的重要性采样
2. Hard负样本抽取
3. 基于邻节点权重的重要性池化操作
4. 利用图片、文字信息构建初始特征向量
5. 阶段性存储节点最新Embedding，避免重复计算



Item间的连接权重不同

Hard负样本



相同节点避免重复计算

从Item2vec到GNN—Embedding在推荐系统中的发展和应用

什么是Embedding？

基于序列数据的Embedding方法

- Word2vec
- Item2vec

基于Random Walk的Graph Embedding方法

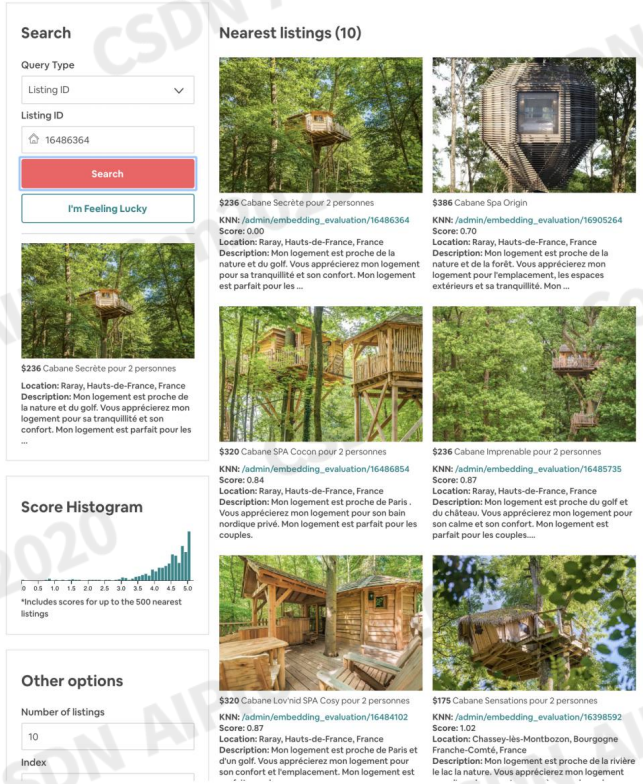
- DeepWalk
- Node2vec

图神经网络

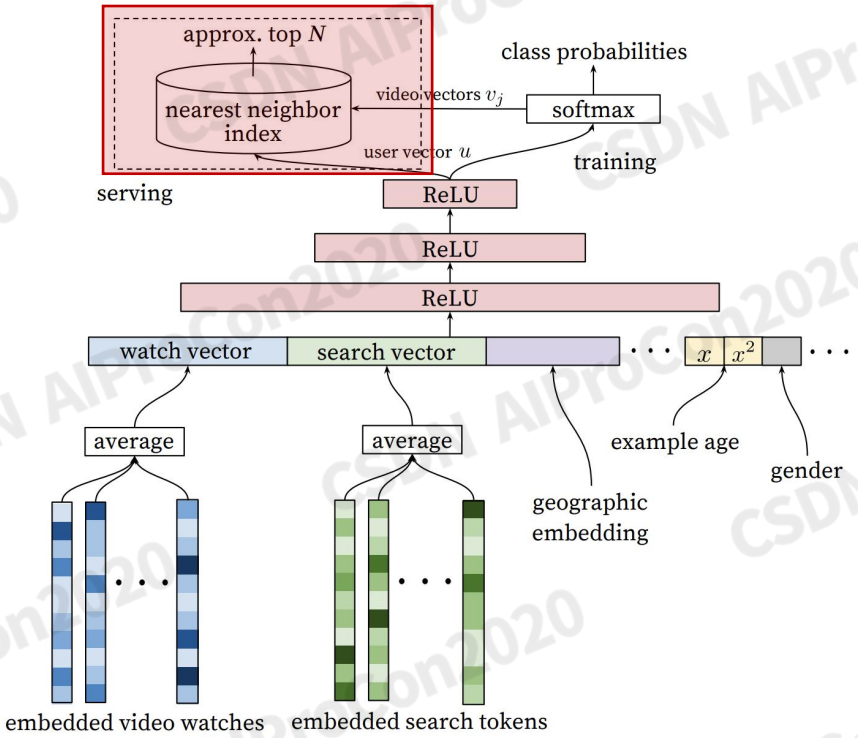
- GraphSAGE
- PinSAGE

Embedding技术在推荐系统中的落地

Embedding技术在推荐系统中的落地

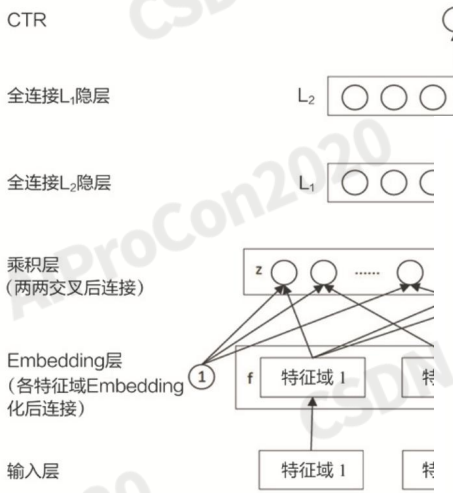


相似物品推荐

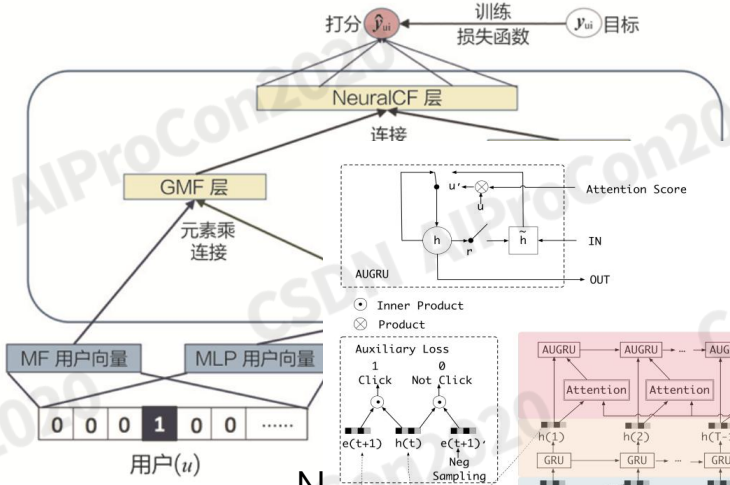


Embedding+ANN召回层

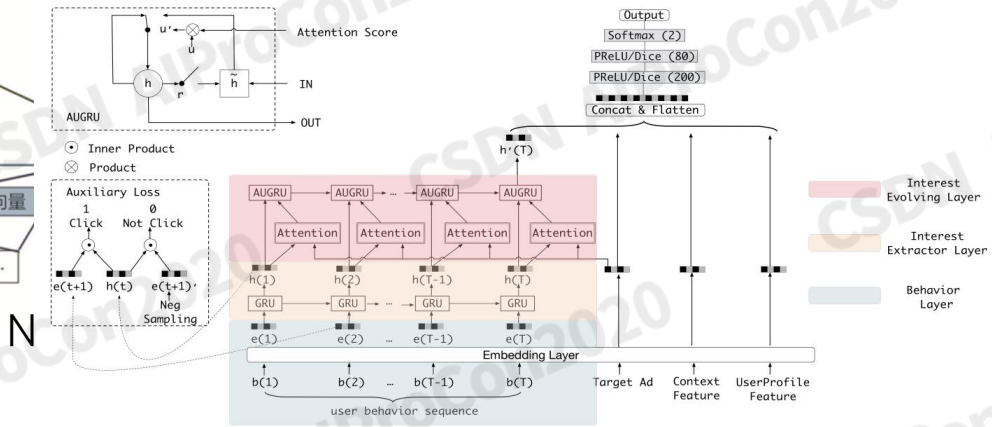
Embedding技术在推荐系统中的落地



PNN



N



DIEN

Embedding作为推荐系统主模型的输入向量

谢谢！

什么是Embedding？

基于序列数据的Embedding方法：Word2vec、Item2vec

基于Random Walk的Graph Embedding方法：DeepWalk、Node2vec

图神经网络：GraphSAGE、PinSAGE

Embedding技术在推荐系统中的落地