

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

百万AI

2020
AI 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](#)



7月3日-4日

CSDN

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

AI

2020
ProCon
万人开发者大会

7月3日-4日

CSDN

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

AI

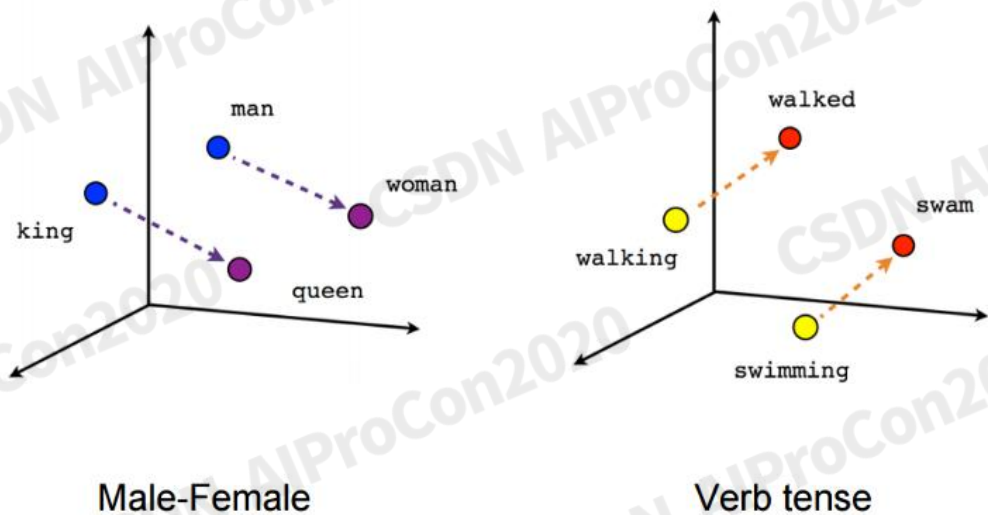
2020
ProCon
万人开发者大会

7月3日-4日

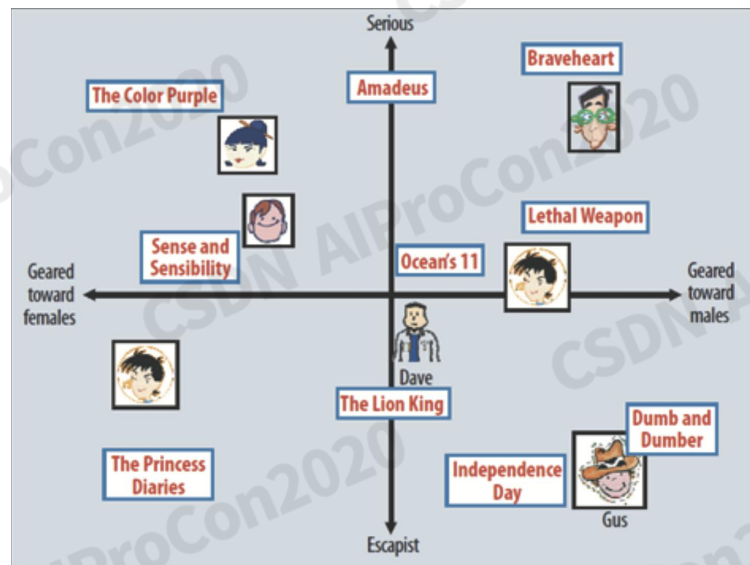
CSDN

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。



词向量



视频-用户向量



7月3日-4日

CSDN

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

AI

2020
ProCon
万人开发者大会

7月3日-4日

CSDN

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

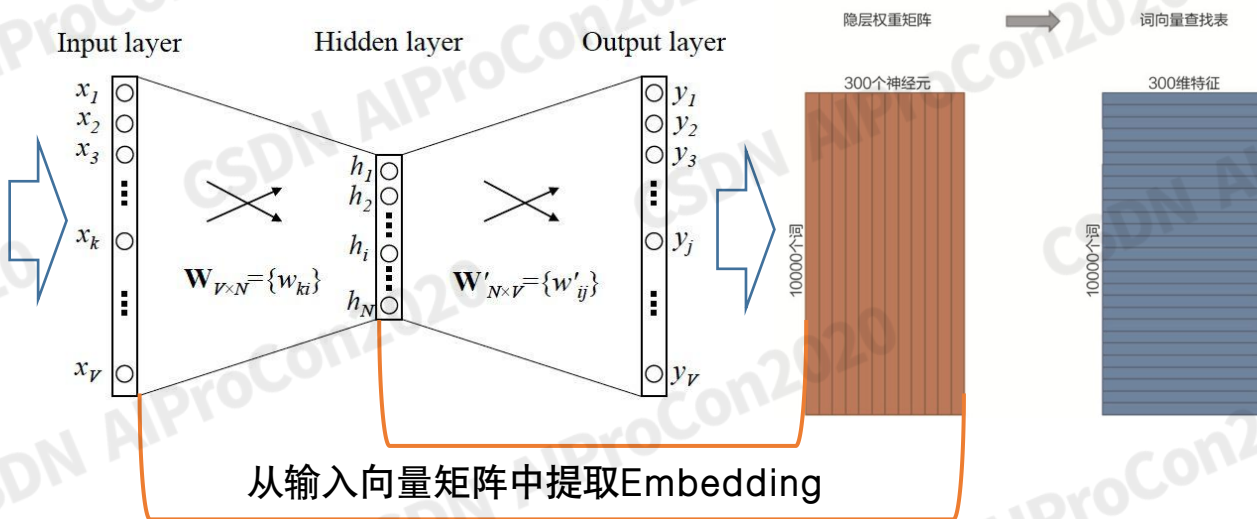
Embedding -> GNN, 推荐系

统

推荐系统 -> Embedding, 发

展

发展 -> 推荐系统, 应用



7月3日-4日

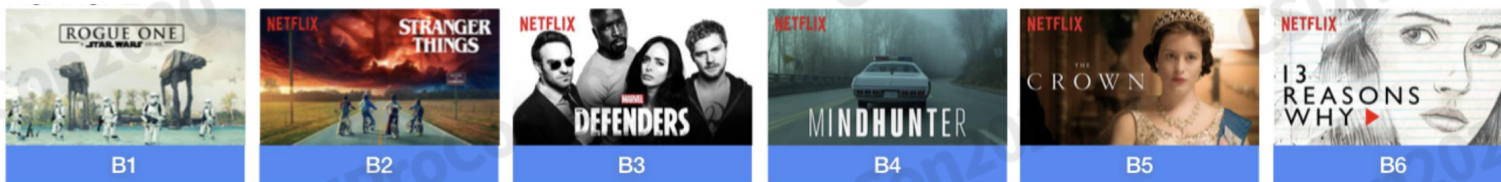
CSDN

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

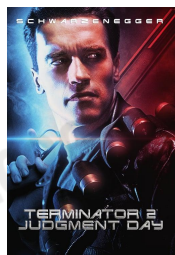
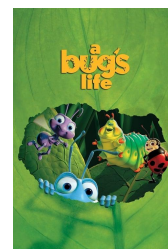
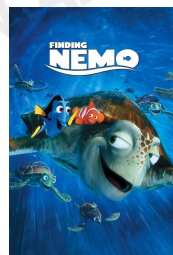
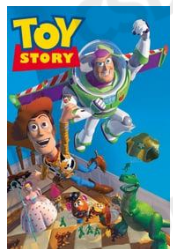
$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t) \Rightarrow \frac{1}{K} \sum_{i=1}^K \sum_{j \neq i}^K \log p(w_j | w_i)$$



7月3日-4日

CSDN

基于序列数据的Embedding方法——Movie2vec的例子



百万人学AI

2020
AI ProCon
万人开发者大会

7月3日-4日

CSDN

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

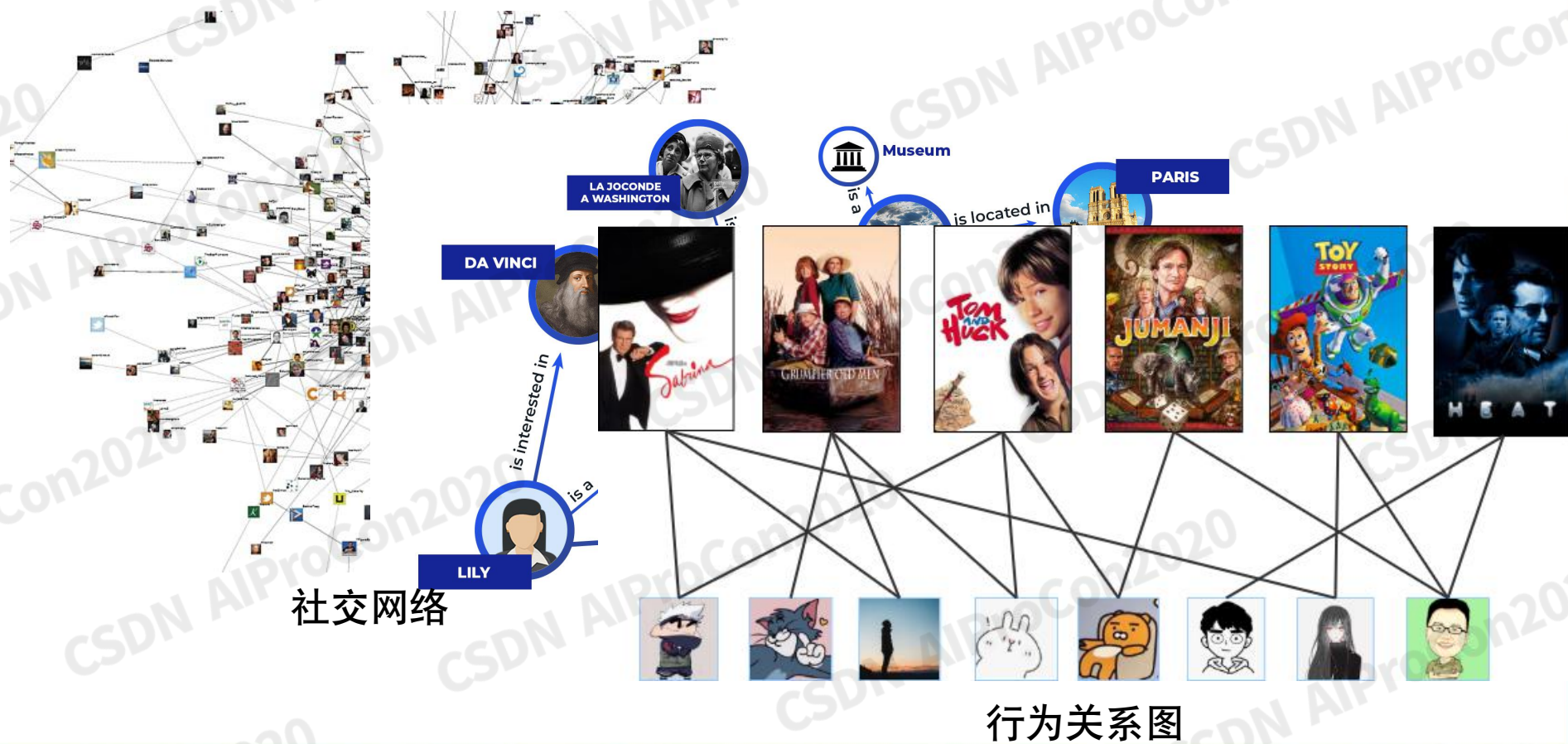
AI

2020
ProCon
万人开发者大会

7月3日-4日

CSDN

互联网数据更多是以图的形式存在



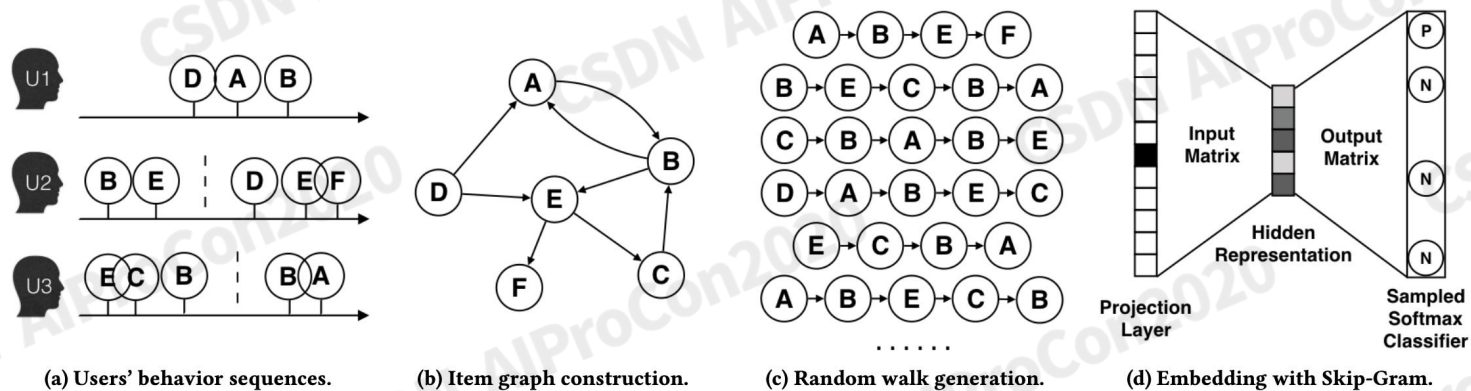
百万AI

2020 ProCon 万人开发者大会

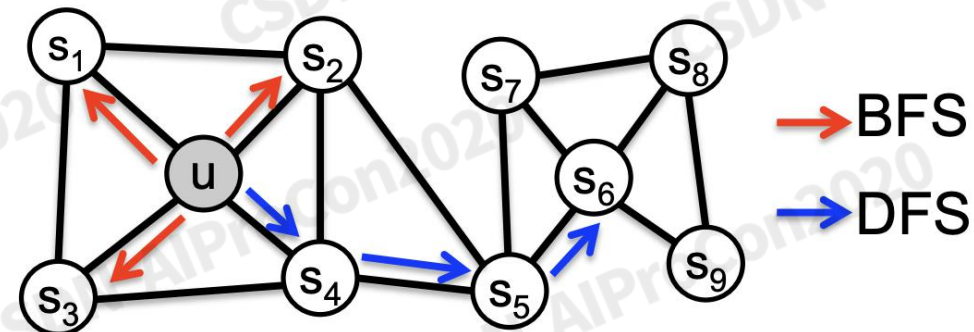
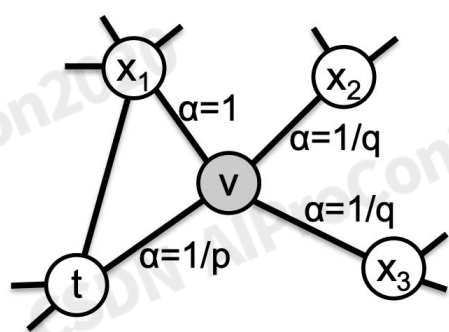
7月3日-4日

CSDN

基于Random Walk的Graph Embedding方法



DeepWalk



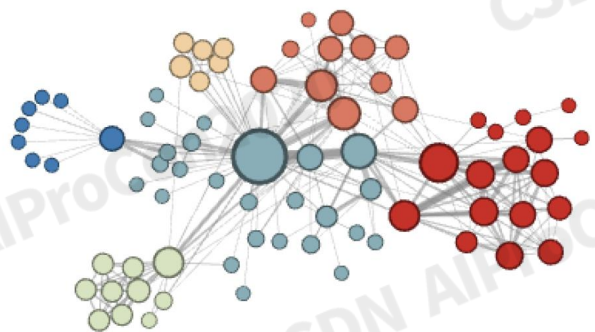
Node2vec



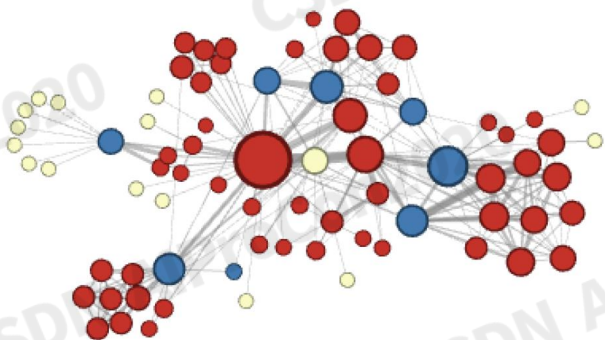
7月3日-4日

CSDN

基于Random Walk的Graph Embedding方法



DFS
节点的同质性



BFS
网络的结构性



百万AI

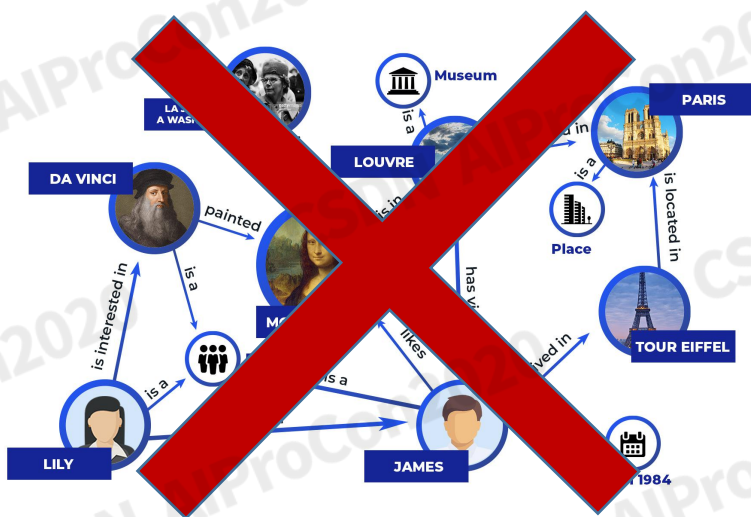
2020
ProCon
万人开发者大会

7月3日-4日

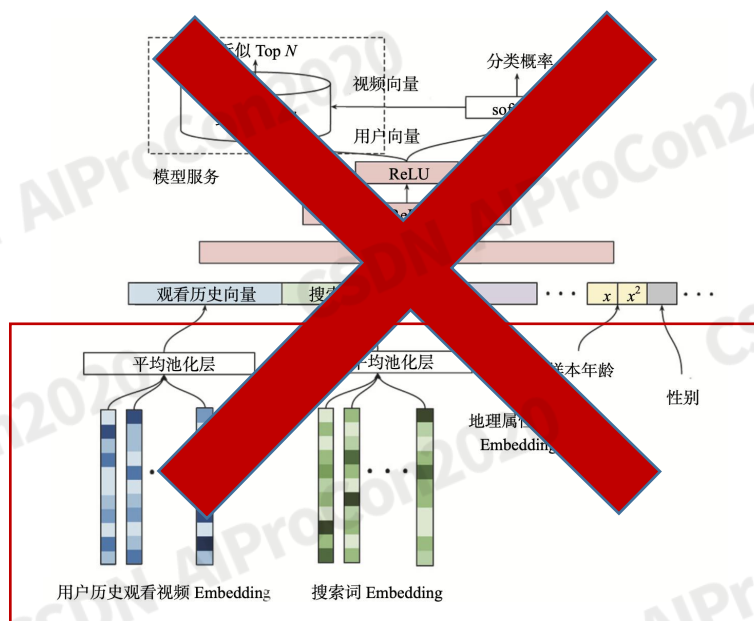
CSDN

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

AI

2020
ProCon
万人开发者大会

7月3日-4日

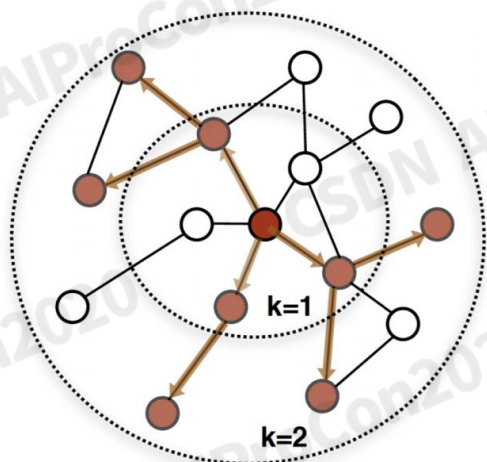
CSDN

图神经网络 (Graph Neural Network) — GraphSAGE

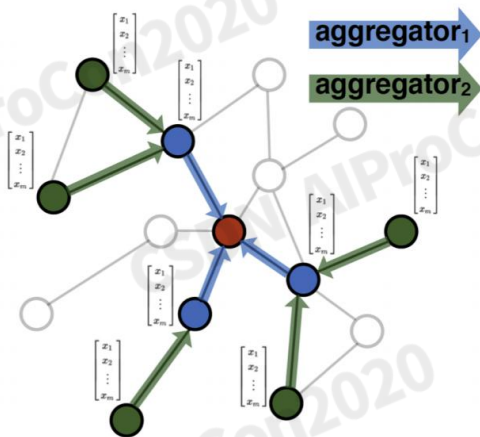
GraphSAGE
PinSAGE

—
—

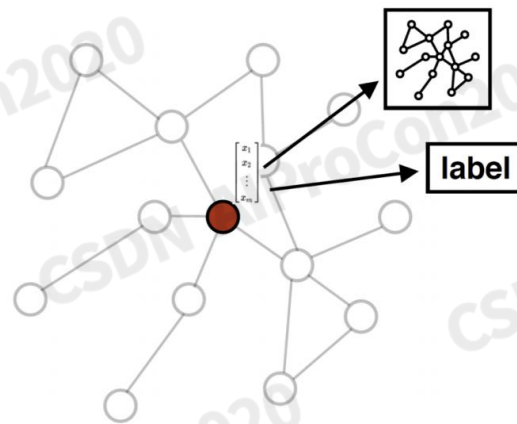
Graph Sample and Aggregate
Pinterest GraphSAGE



1. Sample neighborhood



2. Aggregate feature information from neighbors



3. Predict graph context and label using aggregated information



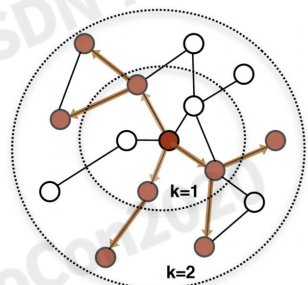
2020
AI ProCon
万人开发者大会

7月3日-4日

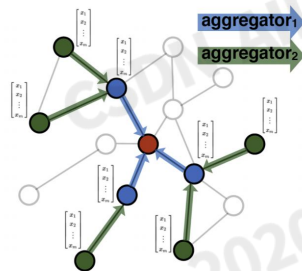
CSDN

图神经网络 (Graph Neural Network)

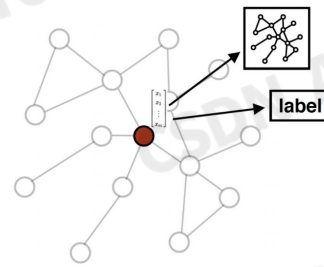
从原始数据
到形成GNN样本子图



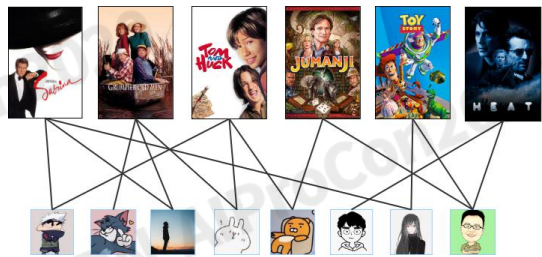
1. Sample neighborhood



2. Aggregate feature information from neighbors



3. Predict graph context and label using aggregated information



AI

2020
ProCon
万人开发者大会

7月3日-4日

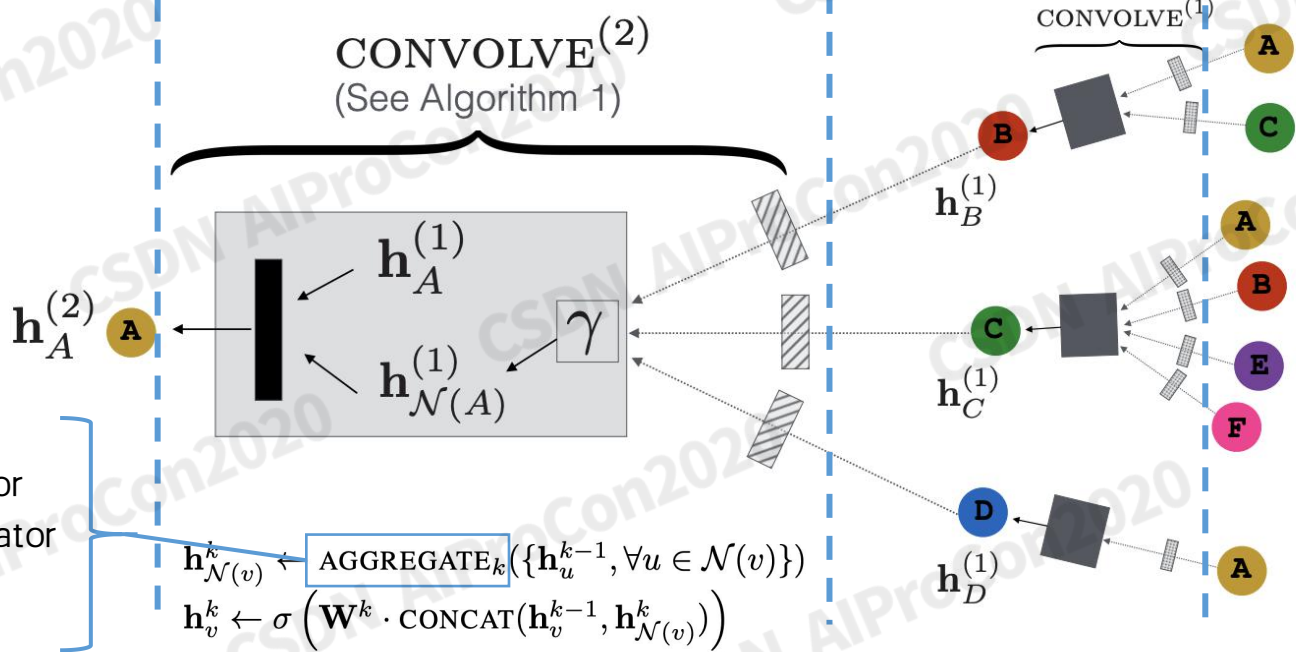
CSDN

图神经网络 (Graph Neural Network) — GNN结构

目标节点

一阶邻节点

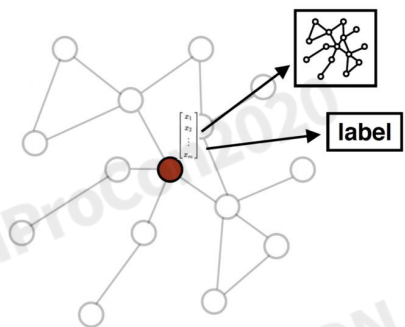
二阶邻节点



7月3日-4日

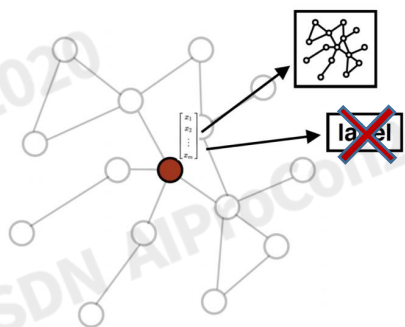
CSDN

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

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

百万
AI

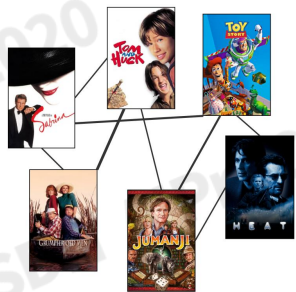
2020
ProCon
万人开发者大会

7月3日-4日

CSDN

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

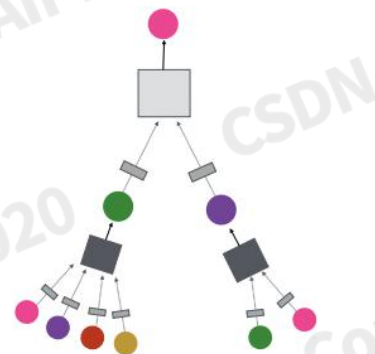


Item间的连接权重不同



Query Positive Example Random Negative Hard Negative

Hard负样本



相同节点避免重复计算



7月3日-4日

CSDN

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

AI

2020
ProCon
万人开发者大会

7月3日-4日

CSDN

Embedding技术在推荐系统中的落地

Search

Query Type
Listing ID
Listing ID
16486364
Search
I'm Feeling Lucky

Nearest listings (10)

\$236 Cabane Secrète pour 2 personnes
KNN: /admin/embedding_evaluation/16486364
Score: 0.00
Location: Raray, Hauts-de-France, France
Description: Mon logement est proche de la nature et du golf. Vous apprécierez mon logement pour sa tranquillité et son confort. Mon logement est parfait pour les ...

\$326 Cabane SPA Cocon pour 2 personnes
KNN: /admin/embedding_evaluation/16486854
Score: 0.84
Location: Raray, Hauts-de-France, France
Description: Mon logement est proche de Paris. Vous apprécierez mon logement pour son bain nordique privé. Mon logement est parfait pour les couples.

\$326 Cabane Imprenable pour 2 personnes
KNN: /admin/embedding_evaluation/16485735
Score: 0.87
Location: Raray, Hauts-de-France, France
Description: Mon logement est proche du golf et du château. Vous apprécierez mon logement pour son calme et son confort. Mon logement est parfait pour les couples...

\$320 Cabane Lovinid SPA Cosy pour 2 personnes
KNN: /admin/embedding_evaluation/16484102
Score: 0.87
Location: Raray, Hauts-de-France, France
Description: Mon logement est proche de Paris et d'un golf. Vous apprécierez mon logement pour son confort et l'emplacement. Mon logement est

\$175 Cabane Sensations pour 2 personnes
KNN: /admin/embedding_evaluation/16398592
Score: 1.02
Location: Chassey-lès-Montbozon, Bourgogne Franche-Comté, France
Description: Mon logement est proche de la rivière le lac la nature. Vous apprécierez mon logement

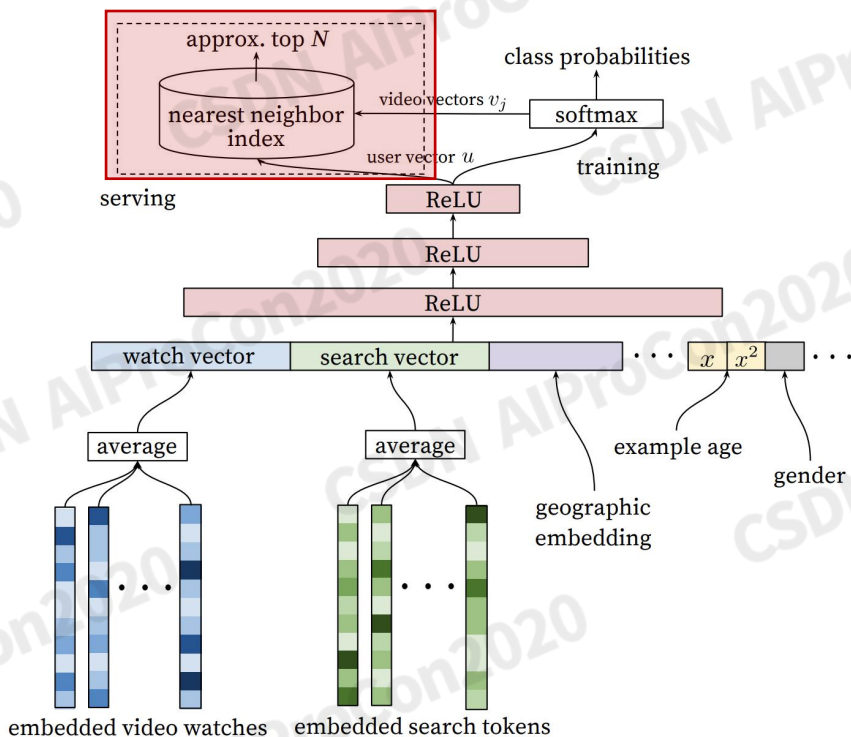
Score Histogram

Includes scores for up to the 500 nearest listings

Other options

Number of listings
10
Index

相似物品推荐



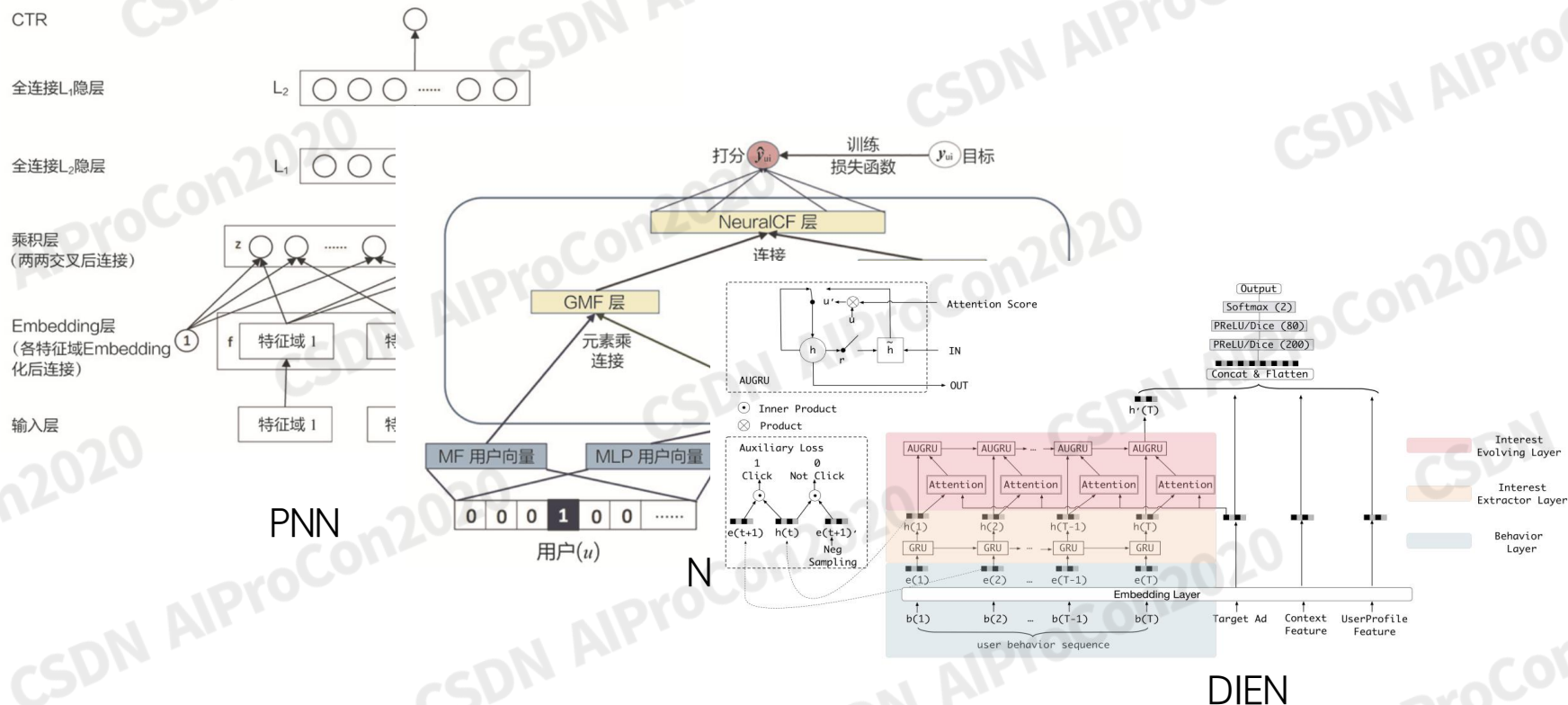
Embedding+ANN召回层



7月3日-4日

CSDN

Embedding技术在推荐系统中的落地



Embedding作为推荐系统主模型的输入向量



7月3日-4日

CSDN

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



7月3日-4日

CSDN