图推荐算法介绍以及在EE问题上的应用

先荐 搜索与推荐Wiki 2020-03-30

前荔枝FM资深数据挖掘工程师庄正中为大家带来了主题为**《图推荐算法在E&E问题上的应用》**的分享，围绕以图为基础衍生的一类推荐算法介绍其原理和应用，如何构建推荐系统里面的图，如何用神经网络提升它的泛化能力以及它如何应对新用户和新内容问题。

授权转载：https://mp.weixin.qq.com/s/OhtuQOjCapRde-95SCf93Q

完整版PPT获取地址在文末

直播完整版内容请点击下方的直播回放观看。

由于大家参与度很高，一些很有价值的提问由于直播时长原因讲师未能及时解答，先荐协调讲师时间，针对直播当天未回答的提问进行了文字版解答。

以下是根据微信群内提问整理的文字版Q&A：

Q:千万级结点的图Embedding有什么快速的训练方法或者训练框架吗？

A：建议借鉴PinSage的大规模Graph Embedding方案，采用的训练框架为TensorFlow。

Q：Airbnb Embedding中的Listing Embedding方法能同时训练：(1)click click … book (2)click click … click这两种的session吗？用损失函数来区分吗？

A：可以。对于这种情况，原理是类似Doc2vec算法将book放到了其他click在Skip-gram模型的context中，即book样本会更多参与训练。

Q：CF图Graph Embedding后的向量是如何和item对应的呢？

A：训练出来是一个MxN的矩阵，这需要看事先你是如何定义的这个Embedding，需要与节点id通过之前的顺序关联起来。

Q："看了又看"场景可以只用类似于i2i这样的方法吗？不涉及用户侧特征，只用i2i效果怎么样？

A：可以，而且这是一般做法。想不u2i，i2i的拟合能力理论上稍弱一些，因为特征少了一些，不过可以解决大部分情况。

Q：DSSM和YouTube的双塔"Sampling-Bias-Corrected Neural Modeling for Large Corpus Item Recommendations"有什么区别？

A: 基本模型一样，但后者提出通过Batch SoftMax Optimization、Streaming Frequency Estimation提升训练效率、修正Sampling Bias。

Q: 图的存储有常用的格式吗？可否理解为Redis里key是一个商品id，value是一串逗号连接的商品id?

A: 内存里可以用map<string,float[]>，外存里可以使用你说的方式。

Q:Category Embedding Entity Embedding 为什么对应着多个Embedding，然后进行pooling?

A: 一个字段如果属于Multi-hot离散特征类型，一般用自己的Embedding，然后采用pooling得到Field Vector，联合编码做大Embedding通常只对1-hot离散特征和连续特征有效。

Q:Mixed graph给Word2vec的Sequence如何构建？和Similarity矩阵有何关系?

A: 请查阅Node2vec算法，Similarity矩阵即item与item的相似度，如果取每个item最相似的topK个item，它本质上可以变成一张图。

Q: 双塔里面的基础特征向量如何来的（例如年龄，地区，职业等）？

A: 随机初始化它们的Embedding，并参与模型的训练得到。

Q: 图像量如何存储，提高查询性能?

A: 建议放在内存里。

Q: 能不能直接用Session的序列，算出vec，而不是先构建Graph?

A: 可以，并且可以将Graph抽样的序列和Session原来的序列一起训练。

Q: Mix Graph,两种图怎么Random walk，边的权值同等对待?

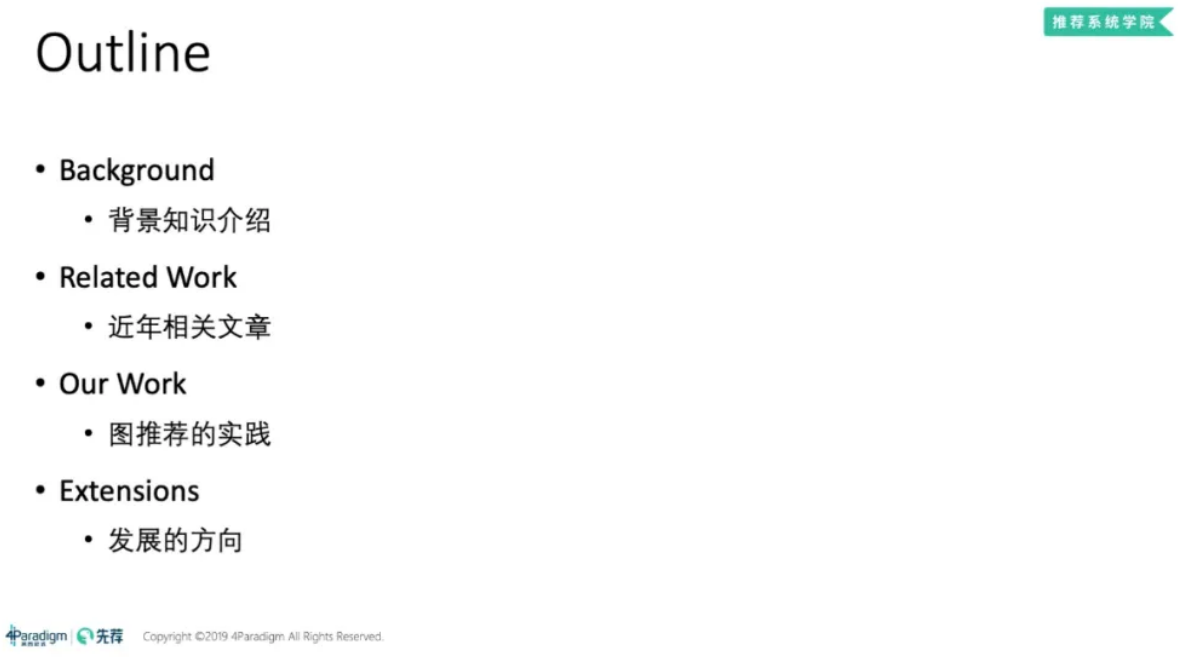
A: 请查阅Node2vec算法，边上如果记录的是频率，则是带权Random Walk。

Q: Graph Restrict Vector Search具体做什么用的?

A: 对于item2item这种场景，使用Vector可以加入用户个性化的一些信息。

部分PPT内容


推荐系统学院
第四范式（北京）技术有限公司
Copyright ©2018 4Paradigm All Rights Reserved.
图推荐算法
在E&E问题上的应用
庄正中
4Paradigm
先荐


推荐系统学院
Outline
• Background
• 背景知识介绍
• Related Work
• 近年相关文章
• Our Work
• 图推荐的实践
• Extensions
• 发展的方向
4Paradigm 先荐 Copyright ©2019 4Paradigm All Rights Reserved.

推荐系统学院

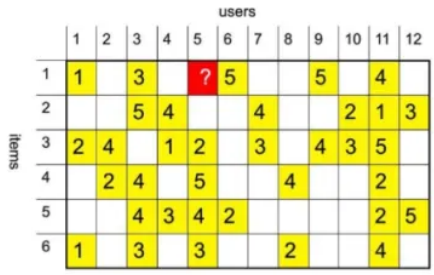
Background

- 推荐系统在E&E上的两大难点
- 新用户
  - 无行为，而用户行为却是最有效特征
  - 无信息，第三方信息获取成本高覆盖低
- 新内容
  - 无反馈，无用户反馈其真实价值
  - 难曝光，长尾内容很难进训练样本

推荐系统学院

Background

- 经典图模型—协同过滤
  - 一种统计学习的图模型，常用于“以物推物”的item-based CF，简单效果好
  - 离线计算出最相似的topk个内容，构成物与物相连的有向图保存在起来
  - 推荐时快速定位用户接触过的seeds列表：
    - 在“1推多”场景直接查找节点的邻边
    - 在“n推n”场景根据多节点对各邻边加权求和
    - 在多样性推荐中将seeds分组并执行多个“1推多”操作



	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3		?	5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

推荐系统学院

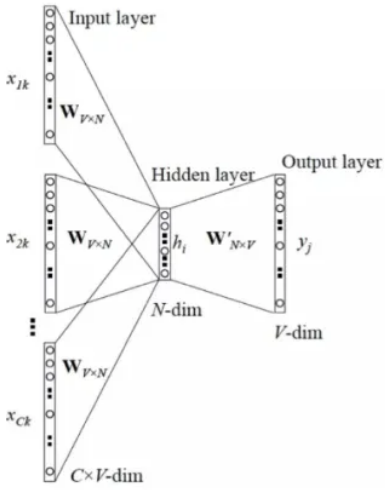
Background

- CF graph较弱的泛化能力
  - 偏热推荐
    因算法本身设计问题，造成马太效应推荐，长尾顾及不到
  - 内部环路
    容易形成内部环路，一些内容互为topk近邻
- 如何拓展图的泛化能力
  - 知识图谱
  - node2vec

推荐系统学院

Background

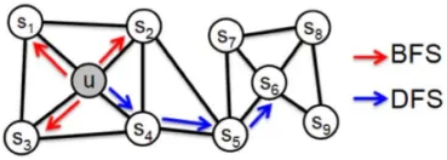
- word2vec
  - 一种将序列向量化的浅层神经网络
    通常只需要有效行为的序列就可以快速得到一个以id为单特征item2vec召回模型
  - 常见结构
    - 特征表达：CBow vs Skip-Gram
    - 加速训练：Huffman tree vs negative sampling
  - id向量集成了所有信息
  - word2vec vs CF
    - 序列保留了时序信息
    - 内容覆盖率提升，相关性略降低

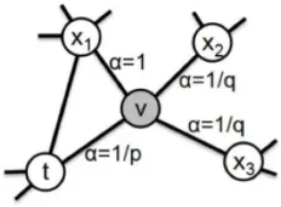


推荐系统学院

Background

- random walk—连接的桥梁
  - deep walk
    - 从一个节点开始采样，跳到下一个节点的概率完全取决于邻边的权重
  - node2vec
    - 额外定义了参数p、q：用于控制回退、BFS、DFS动作
    - prob(v1->v2) = $W_{v1,v2}$*α



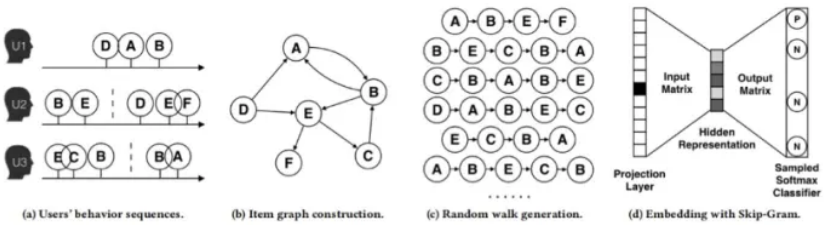


在对graph进行随机游走，生成了很多的ID序列，这些序列即可视为语料输入到word2vec算法中完成向量化

推荐系统学院

Related work

- Alibaba graph-embedding
  - reference《Billion-scale Commodity Embedding for E-commerce Recommendation in Alibaba(kdd 2018)》



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

推荐系统学院

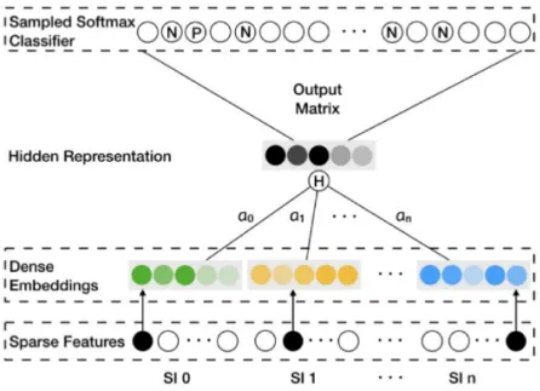
Related work

- Alibaba graph-embedding
  - 构造图的策略
    - 将每个用户浏览序列sequence，做session分隔，1个sequence->n个sessions，每个session构成一个sampled sequence
    - 将所有sampled sequence构成有向图，图边表示该方向流过的次数，得到的graph
    - *该步骤需要过滤掉：异常点击(时间太短的点击)、异常item(更新异常频繁的item)、异常用户(点击量异常多的user)

推荐系统学院

Related work

- Alibaba graph-embedding
  - 多特征融合的word2vec
    - 输出端与word2vec相同，采用对multi-class 做negative sampling加速训练
    - 输入端将物品特征扩展，从单个id特征扩展到引入多个物品特征field
    - 在hidden layer前置一个权重矩阵，来提升模型的拟合能力



> 获取完整版PPT，可在公众号后台回复【0330】