基于边提示异构图网络的多剧集发行受众扩展

Kai Song*
腾讯视频
中国北京
quantumks838@gmail.com

Shaofeng Wang*
腾讯视频
北京，中国
wangsf199610@gmail.com

Ziwei Xie*
腾讯视频
深圳，中国
zeavidxie@tencent.com

Shanyu Wang
腾讯视频
北京，中国
sammywang@tencent.com

Jiahong Li
腾讯视频
北京，中国
kristenli@tencent.com

Yongqiang Yang
腾讯视频
北京，中国
aronyang@tencent.com

arXiv:2304.05474v1 [cs.SI] 8 Apr 2023

摘要

在视频平台上进行用户定向和新节目的受众扩展时，关键在于如何生成它们的嵌入。这一过程应从用户和节目的角度实现个性化。此外，同时追求即时奖励（点击）和长期奖励（浏览时长），以及新节目面临的冷启动问题，也带来了额外挑战。由于数据具有天然的图结构，这类问题适合使用异构图模型进行处理。然而，现实世界中的网络通常包含数十亿个节点和各种类型的边。现有方法中很少有研究专注于处理大规模数据并利用不同类型的边，尤其是后者。在本文中，我们提出了一种基于边提示异构图网络的两阶段受众扩展方案，该网络能够同时考虑不同的双边交互和特征。在离线阶段，为构建图，我们选择用户ID以及节目的特定侧信息组合作为节点，并利用点击/共同点击关系和浏览时长构建边。随后计算嵌入并对用户群组进行聚类。当新节目出现时，可以在一致的空间内生成其嵌入以及后续匹配的用户。在在线阶段，利用包括点击用户和浏览用户在内的后验数据作为种子，以寻找相似用户。在公开数据集和我们的十亿级数据上的结果表明，我们的方法具有准确性和高效性。

CCS 概念

• **计算方法** → **学习潜在表示**；**神经网络**。

关键词

用户扩展，异构图网络，观看 -相似

*这项工作是在他们任职于腾讯视频期间完成的，且两位作者贡献相同。

1 引言

对于视频平台而言，高质量节目对其繁荣发展至关重要，或者用技术术语来说，对其日活跃用户数（DAU）和投资回报率（ROI）至关重要。

通常，新节目会以异步方式发布，每周更新若干集。每个节目会持续几周。因此，在某一天（例如 8月12 日），可能会同时存在多个节目。平台应及时将这些节目推送给对其感兴趣的用户。因此，个性化的用户定向在此发挥着重要作用。

在计算我们的投资回报率和其他核心指标时，尽管我们追求点击，但浏览时长、次日留存率（当天首次登录后于次日返回的用户数量）以及平台互动（投票、评论等）也都是重要因素。不同于电子商务或广告中的点击率/转化率等即时奖励，这些长期奖励在传统的即时奖励监督学习中并不能得到很好的描述。我们发现，点击与否分类模型或浏览时长回归模型会给核心指标的追求带来严重偏差（第5.3节中提到的基线模型）。而重量级多任务模型或强化学习框架在我们的数据上进行调优的成本又有些过高。

此外，在节目上线发布时，工程师通常直到上线前一天才会获知新节目的侧信息（类别、导演、演员、剧情等），因此缺少节目-用户交互，这被称为新节目冷启动问题。另外，不同类型的节目（例如爱情片或战争片）通常会在同一时期发布，但它们之间的点击/观看差异往往相当大。因此，基于历史数据并结合迁移学习 [6] 或元学习[27]的点击率模型或双塔模型都很难满足我们的需求。

为解决上述问题，本文尝试使用异构图网络模型。在图的构建中，用户ID和节目的侧信息（而非节目ID）被选作多属性节点。即时

并通过多属性边将用户与节目之间深层次的长期奖励纳入考量。它们被用于节点间权重计算，并作为明确的双侧特征，促使模型学习不同的关系。

通过这种方式，用户的嵌入向量与节目的侧信息组合被映射到同一空间中。当新节目（及其侧信息）到来时，可以立即查找到其嵌入向量。当出现新的组合时，简单的统计计数或监督学习模型可以帮助找到具有良好可解释性的邻近向量。在第5节中，我们展示了这种针对新节目的临时操作具有很高的准确率。

为了更好地描述节点之间的交互，本文提出了一种边提示异构图神经网络（EPGNN）。在需要进行归一化的边上添加显式交互项，用于引导网络学习由节点间不同关系构成的子图。公开数据集评估和在线结果验证了该方法的有效性。请参阅第5.1节和第5.3节。

与 [6, 31], 类似，我们的框架也是一种两阶段方案。在离线阶段，使用 EPGNN模型生成用户和节目的嵌入。随后，将数亿个用户嵌入聚类为 200,000 个群组，并得到 200,000 个质心的向量。每个节目的前N名（数千万）用户会被存储为基础部分。在在线阶段，在线服务每小时多次返回点击或观看特定节目的用户。随后，将每个节目的点击/观看用户作为种子，借助聚类后的群组和质心，在一分钟内召回数百万用户。详情见第4节。

我们的贡献如下：

- 用户与节目侧信息组合的嵌入在同一图空间中生成，这有助于解决我们的冷启动难题，并可直接用于新节目的发布。
- 本文提出了一种新型边提示异构图神经网络（EPGNN），该网络将由节点间不同关系构成的子图纳入考量。本文采用了一种即时采样策略。
- 一种两阶段（在线和离线）方案，采用异步更新频率设计。离线部分每日更新，而在线部分以低于一小时的频率更新。
- 采用聚类方法加速节目-用户召回，并在多个动态产品广告（mDPA）中重复使用。

2 相关工作

2.1 受众扩展

相似用户扩展技术在广告平台和用户增长中广受欢迎。它们以轻量和稳健性著称。在过去十年中，人们提出了各种各样的思路。

简单的基于规则的方法[17, 22]自然是基础起点。数据科学家根据分析结果制定几项用户筛选策略。例如，选择年龄在20至40岁之间的男性体育爱好者。在现代观点下，这些可以被视为单特征模型。与黑盒网络相比，其可解释性令人满意，同时性能也不差。在我们的实验（第5.3节）中，也测试了基于规则的方法。

在深度学习时代之后，在历史数据上训练监督式神经网络模型已十分常见，随后可以使用该模型生成嵌入，或利用迁移学习 [19] 或元学习 [27, 31]在新数据上进行微调。在 [16], Ma 等人介绍了几种根据计数和重要特征构建的临时相似度度量方法。Perlich 等人 [19] 专注于将迁移学习应用于定向展示广告时所采用的采样技术。在 [6], Pinterest 的deWet 等人使用历史用户及其相应的互动项目（Pins）训练了一个监督式分类器模型。由此获得了用户向量，并应用迁移学习进行更新。在实践中，分类器模型与基于嵌入的方法进行了融合。Zhu 等人 [31] 提出了类似的方案：在离线阶段，利用所有历史数据通过元学习训练一个通用模型。随后在在线阶段，从该通用模型演化出一个针对新广告活动的定制模型。

2.2 图神经网络（GNN）

Perozzi 等人提出的 Deepwalk [20] 和 Grover 等人提出的 node2vec [9]是两种著名的基于 word2vec[4]的同质图嵌入模型。前者在图上采用深度优先搜索（DFS）策略生成序列，后者则使用两个参数 p 和 q 来控制广度优先搜索（BFS）与 DFS 的组合。

在 [7],metapath2vec 模型中，随机游走被推广到了异构图。该模型提出了由元路径引导的游走策略，原则上能够捕捉多属性节点和边之间的相关性。

Velickovic 等人 [28]和 Wang 等人 [30] 将注意力机制应用于图数据，使模型能够在消息传递过程中隐式地为不同的相邻节点赋予不同的重要性。

Cen 等人 [2]提出了一个用于处理具有多属性边的异构图的框架。该模型按照元路径方案进行训练。

3 框架

图 1 展示了我们系统的框架。该系统被设计为类似于以往一些工作[6, 15]的两阶段架构。

左侧圆圈中是我们异构图的示意图，它是节点嵌入的核心。节点和边的不同颜色表示它们具有不同的属性。红色节点代表用户，橙色节点代表节目的侧信息组合。虚线表示浏览时长，实线表示点击或共同点击关系。边的粗细表示其重要性权重的不同。黑色实线表示从其拓扑环境向特定节点进行的消息传递

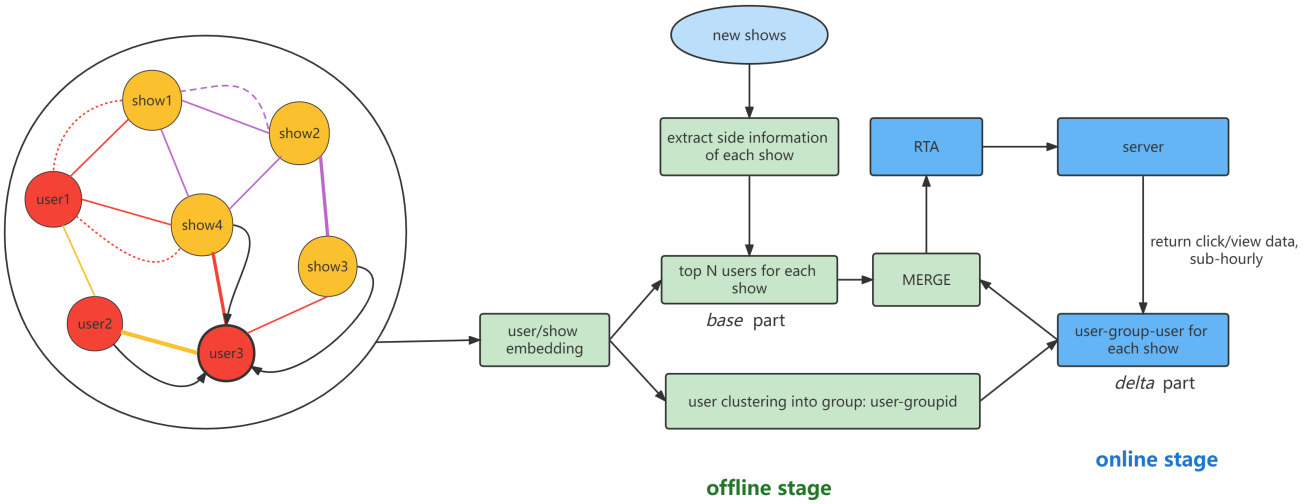


图1：总体框架

。图构建的具体细节将在第4节 4.1 中介绍。在我们的多节目发布中，图每天都会更新。

中间的绿色部分是离线阶段。当待发布的新节目按计划即将发布时，其简介信息（导演、类型等）通常会提前一天获知。可以通过查找组合向量空间或进行简单计数来生成它们的向量。对于前一种方法，可以训练一个有监督回归模型，其中节目的简介信息作为特征，嵌入作为标签。对于后一种方法，我们可以搜索历史节目，查找其简介信息与新节目重叠的节目。然后，通过对这些相似节目的嵌入取平均，获得新节目的嵌入向量。在离线阶段，为每个节目计算前N名用户候选者（数量达数千万）。这些候选者作为在线服务的 *base* 部分。另一方面，将所有用户嵌入聚类为200,000个细粒度分组，并获得用户与组ID之间的双射关系。

右侧的蓝色部分是在线阶段，通过我们内部的实时API（实时 API）服务实现。当每个节目开始获得展示时，点击/观看数据会在不到一小时内返回。这些数据被用作种子，以扩展潜在用户集合。实际上，这里不需要计算用户群组与节目之间的嵌入距离。在 200,000 个对象中的分布进行简单计数就已经足够精确。包含最多种子的 GroupID（及其对应的用户）会被选中。这也有助于减轻种子中的噪声。作为 *delta* 部分，它们会与 *base* 部分合并，以提供在线服务。

4 异构交互图神经网络

4.1 图构建

对于新节目，在发布之前没有点击/viewdata数据。因此，如果直接使用节目 ID作为节点，就无法将其纳入图中。我们改为使用其侧信息的组合，这使得我们的图模型对新节目的泛化能力成为内在属性，有助于缓解本案例中的冷启动问题。

节目约有25种侧信息，包括主类型、子类型、导演、编剧、主演、评分、制片人、热度等级等。需要指出的是，这些属性的不同取值数量庞大，其组合数量更是惊人。幸运的是，在实际应用中，大多数组合都没有意义（例如，某些演员与大多数外国导演合作的情况极为罕见）。最终选取了约135万种属性组合。

未选取其他指纹属性，以抑制噪声。

表1给出了我们的大型图中节点和边的详细信息，表2展示了每类边的度分布。由此可以看出，某些节目的热度显著较高。

需要指出的是，当然也可以将更多类型的节点纳入图中，包括节目中的演员和导演。此外，如果十亿级用户的规模过于庞大，可以先根据其特征或其他先验信息，将他们粗略地聚类为若干群组。然后用这些用户群组代替用户，与节目组合一起构建图。这样，图的规模便只有百万级，相当容易处理。

表1：数据详情

描述	node		同质边		异质边
	show	user	节目-节目	用户-用户	节目-用户
计算逻辑	节目组合	用户 ID	类型是否相同	点击同一节目	点击次数、浏览时长
数量	1.35+ 百万	3亿	2亿	10亿	30亿

表 2：度分布

边类型	分位数-1	分位数-2	分位数-3
节目-节目	>100 2.0%	>20 15.6%	>10 29.9%
用户-节目	>40 7.2%	>20 16.7%	>10 33.8%
用户-用户	>20 44.1%	>10 53.8%	>5 63.8%

4.2 边提示图神经网络（EPGNN）

图网络提供了一种优雅的框架，可在统一的空间中优雅地处理多路交互元素（节点）。异构图网络指的是一般情形，即图由两种或更多类型的节点或边，或两者兼有构成。

图上的消息传递范式大致可分为图嵌入和图网络两类。前者包括 DeepWalk[20], node2vec[9], EGES[29], 等经典模型。后者包括开创性的 GCN[13], GraphSAGE[10], GAT[28], 等。图网络还可根据是否通过积分变换（傅里叶变换、拉普拉斯变换或其他变换）将空间关系转换到谱域，进一步分为空间类型和谱类型。GraphSAGE 属于空间类型，而 GCN 属于谱方法。

本文提出了一种边提示图神经网络（EPGNN）模型，适用于具有多属性节点和多属性边的图。EPGNN 能够充分利用异构图节点之间的不同关系，并感知由任意关系构成的子图。与 GraphSAGE 类似，它属于空间类型。

受消息传递神经网络（MPNN）[8],我们将迭代设计如下：

节点 i 在第 $t+1$ 次迭代时的嵌入为：

$$v_i^{t+1} = b_i^t + f(v_i^t) + g(\{u_{ij}^t\}, \{v_j^t\}, v_i^t) \tag{1}$$

此处，v_i^{t+1}的形状为 $m_n \times 1$，可以使用该节点的侧信息（特征）进行初始化。

u_{ij}^t是边的嵌入，形状为 $m_e \times 1$，其维度不必与节点的维度相同。

b_i 是形状为 $m_n \times 1$的偏置项，随机初始化。

$f(v_i^t)$ 是用于处理节点 i 单侧信息的单体项，在不失一般性的情况下，可将其指定为多层神经网络。

$g(\{u_{ij}^t\}, \{v_j^t\}\ v_i^t)$ 是双体交互项，用于考虑节点 i 与其相邻节点（当然，也可以是邻居的邻居）：$v_j \in \mathrm{N}(v_i)$。通过双向信息对相邻节点的特征进行加权，网络可以学习由各种关系构成的子图，且彼此之间不会相互干扰。注意力技术可以整合进来：

$$g(\{u_{ij}^t\}, \{v_j^t\}, v_i^t) = \sum_j \alpha_{ij}(v_i^t, v_j^t)(\sum_k u_{ijk}^t W_{jk} v_j^t), \tag{2}$$

其中，α_{ij} 是注意力系数（一个实数），可以通过单个前馈层[28]定义：

$$\alpha_{ij} = \sigma\big(F_{\mathrm{nn}}[v_i^t, v_j^t]\big)\,, \tag{3}$$

其中，σ 是非线性映射（例如 Sigmoid 函数），F 表示变换参数。

W_{jk} is a trainable node-node matrix of shape $m_n \times m_n$.

u_{ijk}^t是向量 u_{ij}^t中的一个元素，其值表示某种双边交互，例如点击或浏览时长。

式（1）具有很强的通用性。若仅保留前两项，它就是一个深度神经网络。若保留三项（有无侧信息均可），并将 u_{ij}^t 设为常数，则大致等同于 GraphSAGE/node2vec。若保留三项但不包含双侧特征，并将 u_{ij}^t 设为标量，则会退化为 GATNE[2]。

采用对数 Sigmoid 损失函数 [4, 10]:

$$\mathcal{L}(v_i) = -\log(\sigma(v_k^T v_i)) - Q\mathbb{E}_{v_j \in P_{\mathrm{neg}}(v)} \log(\sigma(-v_j^T v_i)) \tag{4}$$

其中，第一项衡量相邻节点的相似度，第二项衡量图中一个节点与其负样本之间的距离。

其中 σ 是 Sigmoid 函数。$P_{\text{负采样}}(v)$ 是负采样分布。Q 是一个控制负项影响的实数。

我们的 EPGNN模型可以适用于转导式学习和归纳式学习。本质上，该模型以无监督的方式学习图信息与节点嵌入之间的映射。与 node2vec 等模型不同，我们无需在训练中包含特定节点即可生成其嵌入。节点承载的单边特征、边承载的双侧特征以及拓扑信息的处理过程可以与训练部分解耦，如算法 1 所示。而且，在大规模图场景中，所有这些过程都可以并行完成。

分层邻居采样策略

为了计算公式（1）中的第三项，需要一个节点及其相邻节点/边（或邻居的邻居等）。为

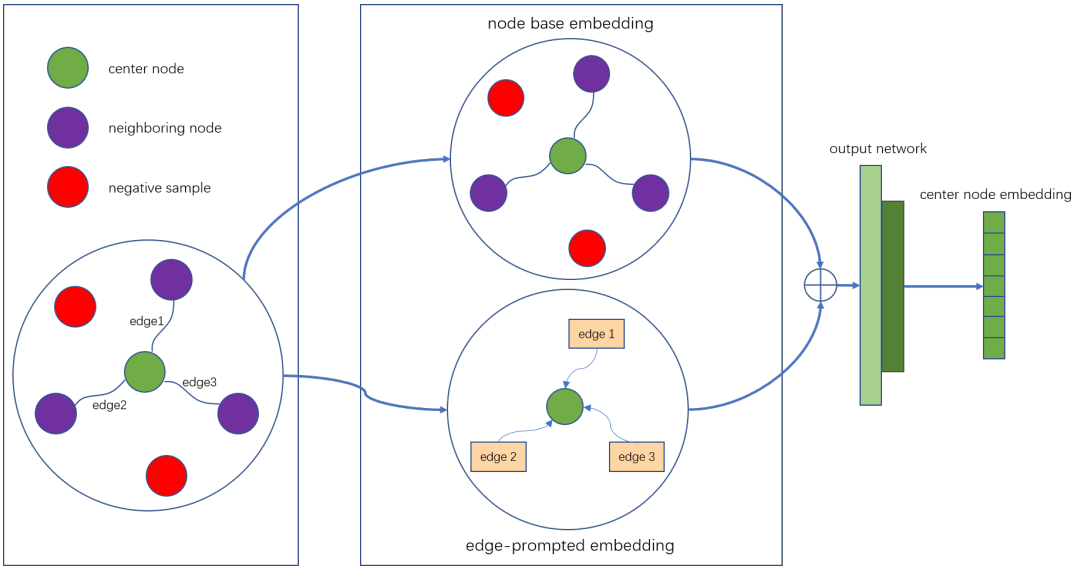


图 2：所提出的 EPGNN 模型结构

应对计算资源（中央处理器，尤其是内存）的限制，以及扩展到大型十亿节点图所面临的挑战，我们设计了一种分层采样策略。在一阶采样中，随机选取一个节点约 70%（或更多，只要内存允许）的相邻节点和边存入内存。

二阶采样以即时方式进行。对于每个批次，从已存储的 70% 相邻节点和边中，有放回地重新采样 50%（即原始度数的 35%）。为此，在保证迭代速度的同时，不会丢失节点局部范围内在内存允许范围内的信息。

然而，借助节点/边的属性和度分布（表 2），可以改进相邻节点/边的随机选择。

EPGNN 的前向传播算法

如算法 1 所示，我们的前向传播算法由四个部分组成：图构建、序列生成、多跳配对生成和模型训练。

在第6-9行中，选择月活跃用户和节目作为节点，以（源节点、目标节点、权重）的元组形式构建 G_1。权重按照第4.1节所述的方式计算。

第 10-15 行对应序列生成部分，其中采用了 metapath2vec[7] 算法。

LINE16-17通过在序列上进行滑动窗口操作生成正样本。与原始图数据 G_1相比，引入了多跳关系，而不仅仅是单跳邻居关系。滑动窗口大小可以控制这种长程交互。

第 18-21 行是对邻近节点和边的特征进行解耦准备。

第 22 至 27 行是 EPGNN 的训练部分。

第19行和第24行是我们上文所解释的分层邻居采样策略。

这个十亿规模的大规模图给存储和计算带来了巨大挑战。为了控制计算成本并加速 meta2path，我们转而采用基于业务背景的先验图划分。具体而言，领域知识和一系列数据分析表明，不同一级类别的节目之间的交互相对稀疏。原始图沿着这些边被划分为十个部分，然后并行输入 SequenceGeneration。

5 实验

5.1 性能分析

链接预测被广泛用于评估节点嵌入的质量。因此，我们首先在三个公开数据集上，将我们的模型与其他最先进的算法进行比较。亚马逊产品数据集[1][11, 18] 包含产品元数据和产品之间的链接；YouTube数据集[2] [24, 25] 由各种类型的交互组成；Twitter数据集[3] [5]也包含各种类型的链接。

选取三类算法作为竞争对手：网络嵌入方法、异质网络嵌入方法和多重异质网络嵌入方法。网络嵌入方法包括 DeepWalk、LINE[23], 和 node2vec，它们只能处理同质图。对于异质网络嵌入方法，我们重点关注代表性工作 metapath2vec。我们选择 PMNE[14], MVE[21],GATNE 作为多重异质网络的代表

[1]http://jmcauley.ucsd.edu/data/amazon/ [2] http://datasets.syr.edu/datasets/YouTube.html [3] https://snap.stanford.edu/data/higgs-twitter.html

表 3：三种数据集上不同方法的性能比较

	亚马逊			YouTube			Twitter		
	ROC-AUC	PR-AUC	F1	ROC-AUC	PR-AUC	F1	ROC-AUC	PR-AUC	F1
DeepWalk	94.20	94.03	87.38	71.11	70.04	65.52	69.42	72.58	62.68
node2vec	94.47	94.30	87.88	71.21	70.32	65.36	69.90	73.04	63.12
LINE	81.45	74.97	76.35	64.24	63.25	62.35	62.29	60.88	58.18
metapath2vec	94.15	94.01	87.48	70.98	70.02	65.34	69.35	72.61	62.70
PMNE	88.38	88.56	79.67	70.61	69.82	65.39	62.91	67.85	56.13
MVE	92.98	93.05	87.80	70.39	70.10	65.10	72.62	73.47	67.04
GATNE	**97.39**	**97.01**	92.62	84.36	82.06	76.95	74.96	74.93	68.29
EPGNN（我们的方法）	97.36	96.68	**92.93**	**85.00**	**82.73**	**77.28**	**81.02**	**80.11**	**73.18**

算法 1前向传播算法 for EPGNN

输入：
1: 节点的集合 V；节点特征的集合 F_n；边特征的集合 F_e；2: 元路径；窗口大小 k；游走长度 l；每个节点的游走次数 r；3: 一阶采样数量 s_1；二阶采样数量 s_2；4: EPGNN训练轮次 n；5: $walks = []$

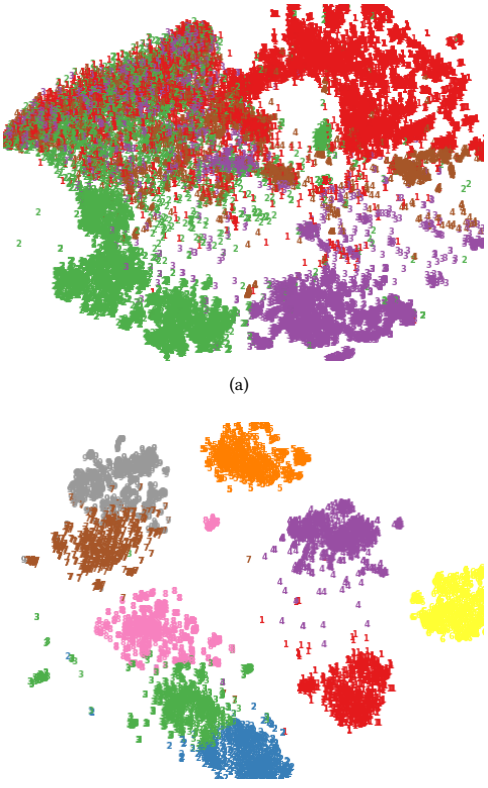
Output: 嵌入 gs $\{v\}$ 对于所有 v $\in V$;

6：边集合 E用户–节目 ={(vuserID,vshow)} 7：边集合 E用户–用户 ={(vuserID,vuserID)} 8：边集合 E节目–节目 ={(vshow,vshow)} 9：$G_1(V, E_1)$ =图生成(E用户–节目 E用户–用户 E节目–节目)，形式为 (vsource,vdestination,weight) 10：对于 $iter = 1$ 到 r，执行 11：对于所有节点 v，$\in V$ 执行 12：$walk$ =SequenceGeneration(G_1, v l) 13：将 $walk$ 追加到 $walks$ 14：结束循环 15：结束循环 16：$pair_set$ =配对生成($walks, k$) 17：$G_2(V, E_2)$ = 基于配对的样本生成 ($pair_set, G_1(V, E_1)$) 18：对于所有节点 v，$\in V$ 执行 19：一阶采样，从 v 采样 s_1 个邻居节点 {$neighbor$}v，以及相应的节点特征 {f_n} $\subseteq F_n$ 和边特征 {f_e} $\subseteq F_e$20：邻居 $_info$v ={neighbor f_n, f_e}v 21：结束循环 22：对于 $iter = 1$ 到 n，执行 23：对于 $pair_set$ 中的 (vi,vj) 24：二阶采样，从 $neighbor_info$v$_i$中采样 {$neighbor'$ f'_n, f'_e}v 个项目 25：通过 $Adam$(vi,vj,{neighbor' f'_n, f'_e}v$_i$) 更新 vv$_i$ 26：结束循环 27：结束循环 28：返回 {v};

嵌入方法，其中 GATNE 是我们的基线。我们复现了GATNE 的实验结果，并确保我们的 EPGNN模型训练设置与其一致。

三个公开数据集的实验结果如表3所示。我们的模型在不同数据集上的F1分数均优于所有类型的基线模型。可以看出，在YouTube和Twitter数据集上，我们方法的所有指标均为最佳。特别是在Twitter数据集上，我们的模型比基线模型GATNE高出约六个百分点。

5.2 生成的嵌入的后验分析和离线评估可视化



(a) (b)

图3：使用 t-SNE 将节目向量空间的维度从 32 降至 2 的可视化

图 3 使用 t-SNE[26] 算法将节目向量空间的维度从 32 降至 2，以此进行可视化。图 3 中相同的颜色表示来自相同类别/子类别的节目组合。上方图中展示了 5 个不同的粗粒度类别。可以看到，各类别之间的分离效果较好，这证明了我们生成的潜在向量质量较高。不过，右上角的区域有些模糊。我们推测类别之间确实存在重叠（例如，有些影片既属于“影片”，也属于“纪录片”）。为验证这一分析，我们根据节目的类型划分出 9 个更细粒度的类别，如图 3 下方所示。此时，各类别之间的分界变得清晰得多。

在线每小时更新部分相较于离线每日更新部分的提升

表4： ***delta*** **部分相较于** ***base*** **部分的点击率和有效观看提升结果**

Hour	点击率提升	有效观看提升
16:00	5.92%	4.14%
17:00	7.70%	3.57%
18:00	3.42%	9.09%
19:00	3.44%	4.99%
20:00	10.31%	16.86%
21:00	3.88%	10.87%
22:00	2.16%	4.96%
23:00	3.83%	9.16%

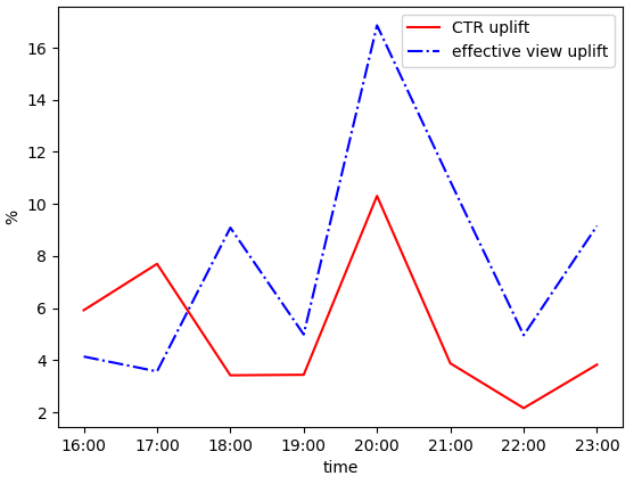


图4：在线按小时更新的 ***delta*** **部分相较于离线按日更新的** ***base*** **部分的提升曲线**

图4和表4展示了每日更新的*base* 部分与每小时更新的 *delta* 部分之间的对比。点击率和有效观看都是我们计算投资回报率时的核心指标。可以看出， *delta* 部分的点击率和有效观看远优于每日更新的 *base* 部分。点击率平均提升了3%，有效时长提升了8%。

实际上，我们分别优化了 *base*和 *delta* 部分的更新频率，并发现 *base* 可以每天更新一次，而 *delta* 增量应每小时更新一次。采用每小时频率的一个原因是，例如，用户在20:00的兴趣和行为可能与19:00时相似，但与18:00时相比则相对不同。

后验分析与离线评估

本文在在线服务之前展示后验分析和离线评估结果。

表5展示了不同模型的点击率和有效观看（时长超过特定秒数）的对比。基于规则的方法与参考文献 [17, 22]中的方法相似。定制化 DNN 方法与我们的 EPGNN 之间的区别仅在于生成嵌入的方式。它们采用了与第3节所述相同的框架。

我们可以看到，我们的 EPGNN 模型能够显著提升点击率和有效观看。

5.3 在线评估

表6对比了不同方法的在线结果。

基准方法选用基于 XGBoost 的通用点击率模型（未针对节目投放进行定制）。文中列出了三个关键指标：投资回报率（ROI）、日活跃用户数（DAU）和每用户平均收入（ARPU）。可以看到，基于规则的方法在投资回报率和日活跃用户数方面取得了显著提升，但每用户平均收入受到的负面影响较大。定制化深度神经网络方法的投资回报率和日活跃用户数甚至优于基于规则的模型，而且每用户平均收入的降幅更小。

与前两种方法相比，本文提出的 EPGNN 在投资回报率方面提升最大，超过基于规则的模型的两倍。我们模型的日活跃用户数与另外两种模型处于同一水平，且每用户平均收入没有下降。

5.4 聚类嵌入的有效性

使用 Faiss 对超过 300 [12] 200, 百万名 000 用户进行聚类，并生成群组。表7展示了条目-群组-用户召回方案的有效性。可以看到，对于四个节目，条目-群组-用户方案的召回结果至少覆盖了精确条目-用户方案的 91%。因此，向量聚类带来的加速使我们能够处理更多条目。事实上，我们已将这种方法用于 mDPA（多种动态产品广告；在我们的案例中数量约为 10,000 个）的定向投放，并取得了令人满意的结果。

6 结论

本研究为视频平台的多节目发布中的用户扩展提供了一种工业级解决方案。我们提出了一种边提示异构图神经网络模型。为兼顾即时奖励和长期奖励，我们使用点击/协同点击和浏览时长来构建边。为解决新节目发布中的冷启动问题，我们选择将节目的侧信息组合而非其 ID 作为一种节点。

表5：离线点击率和有效观看结果

Time	点击率提升 相较于基于规则的方法	有效观看提升 相对于基于规则的方法	点击率提升 相对于定制化深度神经网络	有效观看提升 相对于定制化深度神经网络
第1天 19:00	69.02%	50.16%	27.45%	54.27%
第1天 20:00	98.52%	57.51%	69.86%	81.62%
第1天 21:00	109.60%	55.34%	87.75%	96.53%
第1天 22:00	130.12%	58.32%	104.91%	144.25%
第1天 23:00	24.53%	22.91%	13.85%	21.14%
第2天 00:00	7.24%	9.49%	5.75%	13.44%
第2天 01:00	9.15%	13.90%	8.92%	15.91%
第2天 02:00	7.86%	10.36%	8.06%	11.77%
第2天 03:00	4.15%	8.18%	2.76%	9.43%
第2天 04:00	4.12%	6.29%	6.78%	9.65%
第2天 05:00	8.90%	13.43%	7.20%	14.48%
第2天 06:00	7.62%	11.84%	8.47%	13.38%
第2天 07:00	7.20%	9.82%	8.23%	15.29%
第2天 08:00	6.75%	11.99%	5.09%	15.18%
第2天 09:00	38.87%	50.46%	14.28%	51.35%

表6：不同模型的在线结果

模型	ROI提升	DAU提升	ARPU提升
基于规则	11.43%	26.81%	-9.26%
定制化深度神经网络	19.76%	**28.18**%	-3.60%
EPGNN	**28.57%**	27.34%	**3.00%**

表7：用户-用户召回与用户-群组-用户召回结果对比

show	节目-用户召回	节目-群组-用户召回交集		精确率
节目1	59,945,504	62,244,815 55,558,198		98.1%
节目2	67,877,920	71,002,776 61	,884,203	91.2%
节目3	60,866,837	62,675,794	56,886,038	93.5%
节目4	70,312,486	72,610,760	68,409,003	97.3%

该方案设计为两个阶段。在离线阶段，由用户和节目侧信息组合构成、带有多属性边的图每天都会更新并训练，以在同一空间中生成节点嵌入。通过计算用户与每个待发布节目之间的距离，获得所有节目的 *base* 用户集合。同时，将用户嵌入聚类为细粒度分组，并计算每个群组质心的嵌入。

在线阶段，每个节目的点击/观看数据都会由在线服务以不足一小时的间隔返回。这些点击/观看用户作为种子集，并通过用户-群组-用户技术寻找更多潜在受众（作为 *delta* 集）。*base* 和 *delta* 部分合并后用于在线服务。

在未来的工作中，我们的一项尝试是将动态图与元学习相结合，以便利用每小时增量数据加快更新。这种方法面向视频平台上的高频活动，这类活动通常只持续几个小时，变化也十分剧烈。在这种情况下，缓慢更新的图或多任务模型等复杂方法很难应用。另一个方向是在图上进行多样性控制。多个节目会同时发布。如果曝光量大多集中在其中一个节目上，其他节目就无法积累受众。我们可以引入行列式点过程（DPP）[3] 来寻求平衡。最后，对于视频平台的新用户而言，他们的信息十分稀疏，因此其在图中的局部环境也很粗糙：可用于消息传递的边太少。我们已验证，贝蒂数[1]是有用的工具。

7 致谢

我们感谢腾讯视频增长团队的 Dong Cui、Jinlong Ran 和 Xiao Xiong 提供的帮助并参与讨论。他们积极参与了多节目发布的形式化工作。在线服务由我们团队的实时 API 实现。感谢 Teng Hu 在实时 API 服务方面给予的帮助。我们也感谢 Lei Fan 博士鼓励我们发表这项工作。

参考文献

[1] ZixuanCang 和Guo-WeiWei。2017年。通过元素特异性持久同调分析和预测突变引起的蛋白质折叠能量变化。《生物信息学》33，22（2017年），3549–3557。 [2] Yukuo Cen、XuZou、Jianwei Zhang、Hongxia Yang、Jingren Zhou 和Jie Tang。2019年。属性多路复用异构网络的表示学习。载于《第25届ACM SIGKDD国际知识发现与数据挖掘会议论文集》。1358–1368。 [3] Laming Chen、Guoxin Zhang和Eric Zhou。2018年。用于行列式点过程的快速贪心映射推断，以提升推荐多样性。《神经信息处理进展》第31卷（2018年）。 [4] Kenneth Ward Church。2017年。Word2Vec。《自然语言工程》23，1（2017年），155–162。 [5] Manlio De Domenico、Antonio Lima、Paul Mougel 和 MircoMusolesi。2013年。科学谣言的结构。《科学报告》3，1（2013年），1–9。[6] Stephanie deWet和 Jiafan Ou。2019年。寻找行为相似的用户：用于扩展广告主受众的迁移学习。载于《第25届ACM SIGKDD国际知识发现与数据挖掘会议论文集》。2251– 2259。[7] Yuxiao Dong、NiteshV Chawla 和 AnanthramSwami。2017年。metapath2vec：异构网络的可扩展表示学习。载于《第23届ACMSIGKDD国际知识发现与数据挖掘会议论文集》。135–144。[8] Justin Gilmer、Samuel S Schoenholz、Patrick F Riley、Oriol Vinyals 和George E Dahl。2017年。用于量子化学的神经消息传递。载于《国际机器学习会议》。PMLR，1263–1272。 [9] Aditya Grover和 JureLeskovec。2016年。node2vec：面向网络的可扩展特征学习。载于《第22届ACMSIGKDD国际知识发现与数据挖掘会议论文集》。855–864。 [10] Will Hamilton、Zhitao Ying 和 JureLeskovec。2017年。大规模图上的归纳式表示学习。《神经信息处理进展》第30卷（2017年）。[11] Ruining He和Julian McAuley。2016年。起伏变化：使用单类协同过滤对时尚趋势的视觉演化进行建模。载于《第25届万维网国际会议》。507–517。[12] Jeff Johnson、Matthijs Douze 和Hervé Jégou。2019年。利用图形处理器进行十亿规模相似性搜索。《IEEE大数据汇刊》7，3（2019年），535–547。 [13] ThomasN Kipf 和 MaxWelling。2016年。利用图卷积网络进行半监督分类。arXiv预印本 arXiv:1609.02907（2016年）。 [14] Weiyi Liu、Pin-Yu Chen、Sailung Yeung、ToyotaroSuzumura和 Lingli Chen。2017年。原理化多层网络嵌入。载于《2017年IEEE国际数据挖掘会议研讨会（ICDMW）》。IEEE，134–141。 [15] ZhiningLiu、Xiao-Fan Niu、Chenyi Zhuang、Yize Tan、Yixiang Mu、Jinjie Gu和 GuannanZhang。2020年。用于营销中金融定向的两阶段受众扩展。载于《第29届ACM信息与知识管理国际会议论文集》。2629–2636。[16] Qiang Ma、Eeshan Wagh、Jiayi Wen、Zhen Xia、RobertOrmandi和 Datong Chen。2016年。对相似受众进行评分。载于《2016年IEEE第16届国际数据挖掘会议研讨会（ICDMW）》。IEEE，647–654。 [17] Ashish Mangalampalli、Adwait Ratnaparkhi、Andrew O Hatch、Abraham Bagher- jeiran、RajeshParekh 和 VikramPudi。2011年。面向转化的用户定向长尾营销活动的基于特征对的关联分类方法，用于相似建模。载于《第20届万维网国际会议伴随会议论文集》。85–86。 [18] Julian McAuley、Christopher Targett、QinfengShi和AntonVanDen Hengel。2015年。基于图像的风格与替代品推荐。载于《第38届ACM信息检索研究与开发国际会议论文集》。43–52。[19] Claudia Perlich、Brian Dalessandro、Troy Raeder、OriStitelman和 Foster Provost。2014年。用于定向展示广告的机器学习：迁移学习的实践。《机器学习》95，1（2014年），103–127。[20] Bryan Perozzi、Rami Al-Rfou 和 StevenSkiena。2014年。Deepwalk：社会表示的在线学习。载于《第20届ACMSIGKDD知识发现与数据挖掘国际会议论文集》。701–710。 [21] Meng Qu、Jian Tang、Jingbo Shang、XiangRen、Ming Zhang和 JiaweiHan。2017年。用于多视图网络表示学习的基于注意力的协作框架。载于《2017年ACM信息与知识管理会议论文集》。1767–1776。 [22] Jianqiang Shen、Sahin Cem Geyik和Ali Dasdan。2015年。在线广告中的有效受众扩展。载于《第21届ACM SIGKDD国际知识发现与数据挖掘会议论文集》。2099–2108。 [23] Jian Tang、Meng Qu、Mingzhe Wang、MingZhang、Jun Yan 和QiaozhuMei。2015年。LINE：大规模信息网络嵌入。载于《第24届万维网国际会议论文集》。1067–1077。[24] LeiTang 和Huan Liu。2009年。揭示多维网络中的跨维度群组结构。载于SDM动态网络分析研讨会。568–575。

[25] Lei Tang、Xufei Wang 和 Huan Liu。2009。通过异构交互分析发现群组。载于*2009*年第九届*IEEE*国际数据挖掘会议。IEEE，503–512。[26] Laurens Van der Maaten 和 Geoffrey Hinton。2008。使用 t-SNE 可视化数据。机器学习研究期刊9，11（2008）。[27] Joaquin Vanschoren。2018。元学习：综述。*arXiv*预印本 *arXiv:1810.03548*（2018）。[28] Petar Velickovic、Guillem Cucurull、Arantxa Casanova、Adriana Romero、Pietro Lio 和 Yoshua Bengio。2018。图注意力网络。统计1050（2018），4。[29] Jizhe Wang、Pipei Huang、Huan Zhao、Zhibo Zhang、Binqiang Zhao 和 Dik Lun Lee。2018。阿里巴巴电商推荐中的十亿级商品嵌入。载于第*24*届*ACM SIGKDD*国际知识发现与数据挖掘会议论文集。839–848。[30] Xiao Wang、Houye Ji、Chuan Shi、Bai Wang、Yanfang Ye、Peng Cui 和 Philip S Yu。2019。异构图注意力网络。载于万维网会议。2022–2032。[31] Yongchun Zhu、Yudan Liu、Ruobing Xie、Fuzhen Zhuang、Xiaobo Hao、Kaikai Ge、Xu Zhang、Leyu Lin 和 Juan Cao。2021。通过元混合专家与评论者学习扩展推荐和广告受众。载于第*27*届*ACM SIGKDD*会议论文集：知识发现与数据挖掘。4005–4013。