基于子图邻域学习的网络视频突发事件挖掘

张承德 周 璇
(中南财经政法大学信息工程学院 武汉 430073)

摘 要 基于异构信息网络的跨媒体关联挖掘成为新的研究热点。一般情况下,视频中非线性视觉信息和失范的文本会使得模态间关联极其稀疏。现有方法多采用嵌入多条语义路径来增强媒体间关联。然而,这种方法往往忽略了路径中局部子图结构内节点间的关联,导致节点的子图邻域信息被遗漏,节点嵌入无法捕捉与邻域节点的关联性,进而引起网络视频突发事件挖掘效果不佳。因此,本文提出了一种基于子图邻域学习的跨媒体语义关联增强方法。具体来说,该方法将异构图分解为不同类型子图,在不同子图中捕捉邻域节点的关联,得到关联丰富后的节点最终嵌入。首先,将不同模态节点初级特征映射到统一空间后,将异构图分解为同构和异构子图,以获取节点基于元路径的同构邻居和一阶异构邻居;然后,通过特定注意力机制分别嵌入基于同构和异构子图的邻域节点,捕获子图内节点邻域信息;最后,通过图级注意力聚合同构和异构子图间交互和语义信息,得到邻域关联后的节点最终嵌入,在下游任务中实现网络视频突发事件的准确挖掘。通过在 10 个真实数据集上的实验验证,本文方法展现了较高的可靠性,且所提模型在性能上超越了现有方法。

关键词 跨媒体;网络视频;事件挖掘;子图;子图邻域学习
中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2025.01134

Cross-Media Video Event Mining Based on Subgraph Neighborhood Learning

ZHANG Cheng-De ZHOU Xuan
(School of Information Engineering, Zhongnan University of Economics and Law, Wuhan 430073)

Abstract Cross-media association mining based on heterogeneous information networks has received wide-spread attention. Typically, the non-linear visual information and the inaccurate textual information within videos result in extremely sparse associations between them. Existing methods often enhance these associations by embedding multiple semantic paths. However, these approaches overlook the associations between nodes within local subgraph structures, leading to the omission of neighborhood information. As a result, node embeddings fail to capture the association with neighboring nodes, ultimately leading to poor performance in web video event mining. To address this issue, this paper proposes a cross-media semantic association enhancement method based on subgraph neighborhood learning. Specifically, this method decomposes heterogeneous graph into different types of subgraphs, captures the associations of neighboring nodes within these subgraphs, and obtains the final node embeddings. Initially, node attributes are projected into a shared latent space using type-specific linear transformations. Subsequently, the heterogeneous graph is divided into multiple subgraphs, including homogeneous and heterogeneous structures based on predefined metapaths. Subsequently, tailored attention are independently applied to each subgraph to capture the neighborhood information of nodes. Finally, information from different subgraphs is fused through graph-level attention, aggregating the interactions and semantic information. The learned representations are evaluated by web video event mining.

收稿日期:2024-09-19;在线发布日期:2025-03-12。**张承德(通信作者)**,博士,副教授,主要研究领域为多媒体信息检索、数据分析。E-mail:chengdezhang@zuel.edu.cn。**周 璇**,硕士研究生,主要研究领域为多媒体信息检索、数据挖掘。

Through experiments on 10 real-world datasets, the proposed method in this paper has demonstrated high reliability and outperformed existing methods in terms of performance.

Keywords　cross-media; webvideo; event mining; subgraph; subgraph neighborhood learning

1　引　言

随着互联网技术的飞速发展和智能终端设备的广泛普及,视频已逐渐超越文字成为分发力度更强的事件传播媒介[1]。据统计,YouTube 平台每分钟新增超过 500 小时视频内容,每日观看次数高达 7000 亿次[2]。在事件传播过程中,海量视频数据常常让用户无所适从,难以在短时间内获悉事件全貌。因此,网络视频事件挖掘变得十分必要。

通常,视频中关键帧代表的视觉信息以非线性叙事方式呈现,时空跳跃且情节破碎,这导致生成的视觉近似关键帧(Near-Duplicate Keyframes, NDKs)分布极为分散。而文本信息则存在标题与内容不匹配的失范问题。因此,文本信息在所有 NDKs 中分布的稀疏性不可避免,这导致构建的异构网络不完整且关联缺失,给节点嵌入带来巨大挑战。然而,在异构网络中局部子图捕捉节点间隐藏关联的能力不容忽视[3]。如图 1 所示,子图(a)呈现了一个关联稀疏的异构图,蓝色虚线框和紫色虚线框分别表示节点 NDK_1 和节点 NDK_3 基于路径的局部子图,在局部子图中呈现了路径上节点的邻域结构信息。因此,如何利用子图中丰富的结构信息来捕捉稀疏异构网络中丢失的关联是一个巨大挑战。

目前,针对稀疏异构网络提出了多种图挖掘方法来捕获其中丢失的关联,大致可分为以下三类:(1)基于图的端到端学习方法。该方法通过图卷积网络[4]端到端整合高阶邻域信息或者利用图注意力网络[5]捕获关键信息。然而,这些方法在预测之前,通常将节点及其邻域信息压缩到单个嵌入向量。这种情况下,只有两个节点和一条边被激活,忽视了其他节点及其邻域信息。(2)元图嵌入。该方法将异构图分解为多个同构子图,通过图卷积聚合同构子图内节点信息[6]。然而,该方法忽略了节点的异构交互,导致节点异构邻域信息完全丢失。(3)基于元路径的方法。元路径是由一系列不同类型节点及其关系组成的交替序列,用于捕获网络中的语义信息[7]。然而,该方法过度依赖专家或经验定义的路径,当路径连接稀疏或有噪声时,性能表现较差。此外,元路径嵌入方法沿着包含不同节点的路径以相同方式嵌入语义信息,忽视了路径中不同节点语义差异,导致语义混淆[8]。最后,该方法忽视了元路径之外节点的邻域结构信息,导致节点丢失了对邻域关联的捕捉。以图 1(b)中节点 NDK_3 为例,元路径 NDK_3-V_2-NDK_2 仅能捕捉 NDK_2 与 NDK_3 是同一视频的多种视觉信息,却无法捕捉 NDK_2 的邻域关联节点,导致嵌入 NDK_3 节点时遗漏了 NDK_2 的邻域文本信息。上述方法虽然能够聚合高阶信息和路径中节点语义信息,但是他们都忽略了节点及其邻域信息。在异构信息网络中,子图在提取和理解图中局部结构特征方面起着至关重要的作用,局部丰富的结构化信息超越了单一路径和端到端学习的限制,能深入挖掘图结构的内在关联[3]。如图 1(c)中,子图中粉色区域表示元路径 NDK_1-V_1-NDK_3 中节点 NDK_3 的邻域信息,元路径仅表示 NDK_1 与 NDK_3 是来源同一视频的不同视觉信息,且描述了同一事件,但是通过子图结构却可以找到 NDK_1 与 NDK_3 的邻域信息 T_3 丢失的关联,进而丰富稀疏的异构网络。因此,嵌入局部子图结构中邻域信息的思想有助于增强视觉和文本信息之间的语义关联。为了提高网络视频事件挖掘效果,本文尝试基

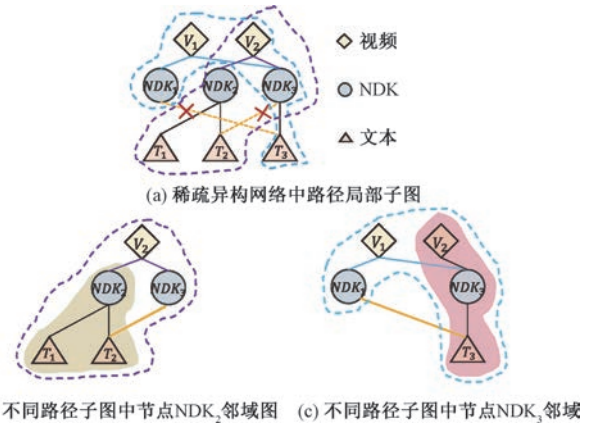

图 1　一个关联稀疏异构图局部子图的例子(上半部分蓝色虚线框和紫色虚线框分别表示节点 NDK_1 和节点 NDK_3 基于路径局部子图. 而图的下半部分阴影部分则表示局部子图中节点邻域关联,通过子图中节点邻域关联能够找到稀疏异构图中丢失的多源关联关系)

于子图邻域学习的方法来丰富稀疏的异构网络。

但是,基于子图邻域学习的网络视频突发事件挖掘仍然面临三个挑战:

(1)复杂的异构信息。通过异构网络建模网络视频通常包含复杂的异质信息,包括单词(Term)与视觉近似关键帧集合(NDKs)之间的语义交互,以及单词和视频底层特征之间的交互。如何处理这些复杂的交互关系并无保留地聚合这些异构信息是一个紧迫问题。

(2)多样化的邻域交互。子图能够独立关联节点及其邻域节点,将异构图分解为多个具有不同交互语义的独立区域。然而,每个独立子图中包含了多样化的交互信息。此外,节点在不同子图中与不同类型节点交互。因此,除了将多样化邻域交互区域化之外,还需要设计聚合方式来区分节点在不同子图中与邻域的交互差别。

(3)跨子图语义交叉。节点在不同子图内与其邻域有丰富交互关系。此外,包含同一节点的不同子图间也有丰富交叉语义关联。在异构网络中,由于不同子图代表特定类型的交互关系和语义信息,因此,如何挖掘跨子图间交互,学习交叉语义特征是一项挑战性问题。

为了解决这些挑战,本文提出了基于子图邻域学习的网络视频突发事件挖掘方法,通过子图中节点邻域来捕获和聚合节点对之间的交互关系。首先,本文将节点邻域推广到基于元路径引导的同构邻域和基于关系的异构邻域。在此基础上,本文构建交互模块,将异构图分解为节点同构子图和异构子图,用于捕获和聚合节点邻域交互信息。最后,本文设计了子图邻域聚合模块,主要由两部分组成:(1)子图内特定注意力机制,用于衡量每个子图内节点与其邻域节点的关联;(2)子图间图级注意力机制,用于聚合节点在不同子图间的语义嵌入。经由以上三个步骤,可以获取节点的最终嵌入,实现网络视频突发事件的准确挖掘。本文的主要创新和贡献如下:

(1)提出了一种基于子图邻域交互学习的框架来捕获和聚合异构图中的复杂节点及其丢失的邻域信息。它旨在通过子图将异构图中复杂邻域交互分而治之,从而将基于邻域的网络嵌入扩展到子图邻域交互学习嵌入,利用不同子图间结构信息丰富媒体间关联。

(2)提出了差异化子图模块来建模多样化邻域交互。针对节点与不同邻居节点间多样化的邻域交互关系,分别诱导同构和异构子图关联节点高阶邻域信息和直接邻域信息,将节点的同构和异构邻域分开处理,以捕获同构邻域和异构邻域对节点嵌入性能的贡献。

(3)提出了子图间交叉交互语义学习方法。本方法应用注意力机制嵌入不同子图间节点的交互关系和语义信息,利用邻域信息找到节点与丢失节点间的关联,丰富异构信息网络,对不同特定交互的子图计算其语义贡献权重,实现聚合子图间交互语义的最优节点嵌入。

2 相关工作

2.1 事件挖掘

事件挖掘是话题检测与追踪(Topic Detection and Tracking,TDT)[9]下的重要分支。在信息爆炸的时代背景下,用户对全面、快速获取感兴趣主题信息的需求日益增长。事件挖掘技术因其能够高效解析话题中的具体事件而受到广泛关注,并已成为学术界的研究热点[10]。随着社交媒体的快速发展,事件挖掘也从面向单一文本、图像和视频扩展到包括文本、图像、视频的跨媒体挖掘。

对于文本信息,早期事件挖掘方法主要基于概率的潜在主题模型及其众多变体,如概率潜在语义分析(pLSA)[11]和潜在狄利克雷分配(LDA)[12]。这些模型将文档视为潜在主题的混合体,而主题则是词汇的概率分布。然而,词汇稀疏的短文本文档词汇概率分布极低,导致在事件挖掘中表现不佳[13]。近年来,短文本事件挖掘引起了广泛关注,其主要目的是克服短文本稀疏问题。文献[14]从知识库中引入与短文本相关的先验知识,如常识性知识和概念,以缓解短文本数据稀疏性。文献[15]运用主题模型增强特征的方法对稀疏文本信息进行增强。虽然这些方法可以推断出潜在的主题,但它们都是基于语义扩展的短文本事件挖掘,没有充分挖掘短文本本身语义信息,缺乏对文本的语义理解[16]。图嵌入技术通常将节点隐射到低维空间,用密集向量表达节点信息,是捕获节点语义的一种方法[4]。然而,当前 GNN 中将不同模态节点特征简单拼接或线性投影,无法提取节点异构内容之间的“深度”交互。该文献主要方法通过递归神经网络对异构特征进行深度交互编码,从而生成节点的嵌入特征表示。

对于图像和视频,现有方法大多利用视觉信息实现事件挖掘。文献[17]提出了一种基于注意力机

制的模型,能够同时提取视频的局部和全局语义特征。该方法通过计算关键帧之间的相似度来衡量视频的相关性,并将相似度超过阈值的视频划分到同一事件中。文献[18]以视觉特征为主要特征追踪事件主线,提出 NDK 这一概念,该方法在视觉相似度计算和镜头捕捉与追踪中表现优异。文献[19]主要通过视觉关键帧的相似性分析对 NDK 进行可视化聚类,以挖掘突发事件。文献[20]提出了一种优化视觉特征的方法,通过结合样本类的中心特征和邻域特征来增强不同事件之间的区分度。上述方法能够充分利用视觉信息进行挖掘。然而,对于大多数由用户随意拍摄和编辑的网络视频,由于光照、运动、拍摄角度等因素的干扰,视频的视觉语义特征往往不够准确,导致检测效果不理想。以上方法主要关注单一媒体,忽略了跨媒体挖掘。

对于跨媒体数据,目前方法主要基于跨媒体数据的互补性进行事件挖掘。文献[21]提出了一种创新的网络视频事件挖掘框架,该框架基于注意力图结构学习,能够生成新的邻接矩阵并重新构建节点间的关联关系。文献[22]将视频理解特征和视频中标题和文本特征进行融合,实现文本与视觉信息之间的关联增强,利用跨媒体信息的互补性弥补话题检测中文本稀疏问题。文献[23]提出了一种跨媒体循环神经网络,该网络通过对多达五种媒体类型数据的细粒度信息进行联合建模,深入挖掘了不同媒体内部的细节特征及其上下文关系。文献[24]通过联合微调的多模态预训练的文本和图像编码器将图像嵌入文本空间,利用图像信息弥补稀疏文本信息实现事件挖掘。文献[25]将密集的突发标签与近似重复关键帧(NDKs)进行融合来挖掘网络视频事件。这些方法都体现了多源信息关联思想以改进和规范特征表示[26]。跨媒体信息的互补性为事件挖掘带来了较好的效果,因此,本文建模异构信息网络来关联跨媒体数据,利用异构信息网络中子图邻域交互学习来挖掘跨模态语义关联,实现网络视频突发事件精准挖掘。

上述代表性方法的特点和局限性见表 1。在基于文本的事件挖掘方法中,网络视频的短文本特性使得文本语义稀疏,难以捕捉视频间语义联系。视觉信息从另一维度丰富了稀疏的文本信息。然而,这些方法严重依赖视觉相似性检测,导致难以捕获事件中具有显著差异的特征。对于跨媒体事件挖掘方法,异构网络因其表征各类实体及其复杂关系的能力成为建模网络视频中多种媒体间关联的有效解决方案[21]。然而,聚合邻域节点的网络嵌入主要以相同方式聚合邻域信息,而忽略了节点同构和异构邻域的语义差异,给节点嵌入带来了语义混淆问题,致使网络视频突发事件挖掘表现不佳[27]。为了解决这一问题,本文通过子图化邻域节点,探索节点与不同类型邻居差异化交互信息。通过子图内注意力机制,衡量每个子图内节点邻居对其交互信息的作用;通过子图间注意力机制,聚合节点在不同子图间

表 1　不同方法的类别、特点与局限性

类　别	方　法	特　点	局限性
使用文本信息	基于 LDA 的在线主题检测[12]	文档被视为潜在主题的混合体,而主题则是词汇的概率分布。	未考虑词汇随时间的变化,这可能会限制它在捕捉主题变化方面的能力。
	从外部引入主题相关先验知识以进行短文本聚类[14]	这个过程克服了短文本数据稀疏问题,使用来自外部的常识知识来辅助短文本聚类。	依赖先验知识和短文本之间转移主题。但是,这种依赖性可能会导致在传输过程中主题失真或信息丢失。
使用视觉信息	通过基于注意力的视频表示和分类进行复杂事件检测[17]	互补地表示突出的物体和周围环境。从显著性对象和整个特征图中提取两个卷积神经网络(CNN)特征,即局部显著性特征和全局特征。	强调视频帧之间的时空关系,而忽略了上下文信息,可能导致复杂场景中的主题检测不准确。
	通过检测视频中视觉特征的方法跟踪事件主题[18]	采用图聚类技术利用视觉信息进行相似性检测,将突发特征分组为有意义的突发事件。	只能检测非常相似的关键帧。这种方法忽略了属于同一事件不同的视觉呈现形式。
融合文本与视觉信息	基于注意力图结构学习的跨媒体视频事件挖掘[21]	为了求解稀疏的文本信息,该方法从异构图中创建关系子图并提取节点关系。然后,通过注意力融合这些图表以丰富信息。	忽略了局部结构信息,无法捕获不同尺度的跨媒体数据之间的局部隐式相关性。
	联合微调的多模态预训练的文本和图像编码器将图像嵌入文本空间[24]	利用图像信息弥补稀疏文本信息,将图像信息嵌入编码至文本空间中,利用多源信息融合实现事件挖掘。	忽略了图像与文本信息对齐问题,导致对主题语义的理解不准确。

的语义嵌入,获得节点的最终嵌入,实现网络视频突发事件的准确挖掘。

2.2 图神经网络

GNN 的目的是学习每个节点的低维向量表示,并用于下游网络挖掘任务,如顶点分类[28]和链接预测[29]。从广义上讲,GNN 是卷积运算对任意图结构数据的泛化。现有方法主要分为两类:基于谱的方法和基于空间的方法。

基于谱的方法的主要思想是在谱域(即傅里叶域)中进行卷积运算以学习图中节点表示。受到图谱理论的启发,Bruna 等人[30]首次设计了傅里叶域中可学习图卷积运算。Defferrard 等[31]采用切比雪夫多项式对图函数进行近似,有效减少了谱域图卷积的计算复杂度,显著提升了运算效率。Kipf 等[32]使用线性图函数简化了卷积操作,并成为最流行的模型。当前,基于谱的方法需要输入整个图结构来执行卷积操作,其在可扩展性和稳定性方面表现较差。另一方面,执行图卷积操作的图函数选择必须满足图的拉普拉斯定理,且该函数选择受限于图结构。

相反,空间图神经网络采用传递消息[33]范式从节点的局部邻域提取、聚合和转换信息,通过堆叠多个网络层对高阶邻域进行建模。Atwood 等[34]将图卷积建模为一种信息扩散机制,假设节点能够以特定概率将自身信息传播至邻近节点。Hamilton 等[35]使用聚合器函数来组合采样邻居的信息。随着注意力机制的广泛应用,Velikovic 等[5]将自注意力机制应用在消息传递过程中动态计算节点和其邻居对的聚合权重,根据重要性聚合相邻节点的消息。

上述图神经网络模型仅适用于节点和边类型相同的同构图。本文构建的异构图中,不同模态节点特征不在同一空间,这些 GNN 模型无法直接对其进行建模。

2.3 异构图嵌入

异构网络嵌入是连接网络原始数据和网络应用任务的桥梁。它能够在保留网络拓扑信息和语义信息的基础上,将网络中的节点映射为低维稠密向量,降低异构网络中由于复杂结构产生的维度灾难[36]。现有的异构网络嵌入建模主要分为两类:一类是基于元路径的嵌入方法,另一类是基于边类型聚合邻居信息。

基于元路径的嵌入方法中,文献[37]提出了一种元路径引导的随机游走策略,用于生成节点序列,并将其输入到 skip-gram[38]模型中以获取每个节点的嵌入表示。该方法能够有效捕捉节点之间的特定语义关系。文献[39]基于元路径的随机游走算法,捕捉相同类型节点序列语义信息,并应用 DeepWalk[40]模型进行节点表示学习。尽管这些方法都专注于图结构的传统图表示学习模型,但它们忽视了节点的属性信息以及其他类型的节点信息。为了更充分地捕捉异构图中的丰富信息,研究人员开始使用元路径来发现目标节点的邻居节点,并聚合邻居节点和元路径的语义信息对目标节点进行编码。文献[41]通过元路径提取同构子图,并利用注意力机制聚合同构子图节点语义信息,最后再基于语义注意聚合多条路径信息。而文献[42]则通过注意力机制将元路径和元结构进行整合,以学习节点的嵌入表示。虽然上述方法将异构节点的属性信息和语义信息充分整合,但这些模型仅聚合了来自元路径端点的信息,而忽略了中间节点的丰富结构和属性信息,导致了信息的丢失。为了保留中间节点的信息,文献[43]提出了 MAGNN 模型,利用元路径内聚合的方式来聚合每个元路径中的所有节点信息。文献[8]也采用了类似的方法,通过对元路径上的所有节点信息进行平均聚合来嵌入节点信息。尽管这些方法对元路径上的不同类型节点均执行了聚合策略,但它们忽视了相同类型节点和不同类型节点之间的连接差异。相比之下,本文提出的模型采用了更具针对性的策略,分别从同构和异构邻居中学习节点信息,以解决因节点类型差异而引起的语义混淆问题。

基于边类型聚合邻居信息的方法中,文献[44]将异构网络中不同边子图的邻接矩阵进行矩阵乘法,在异构网络中不同边子图的邻接矩阵进行矩阵乘法,在异构图中生成有语义价值的元路径邻域图,并使用 GCN 对图进行编码,得到异构网络中节点的嵌入。然而,异构信息网络中部分链路可能是随机生成的,这些链路无法传递有效信息,甚至可能带来额外的噪声干扰。文献[45]利用异构卷积,将对应于不同边类型的关系子图组合成新的图结构,并采用训练图卷积消息传递范式的方法,以优化和提取有效的元路径,用于异构图表示学习。文献[46]将整个异构图输入到图神经网络中,以聚合相邻节点信息并捕获异构图的结构信息,通过去噪操作过滤潜在的噪声节点,以进一步增强节点表示。文献[47]将异构图分解成若干个二部图,应用 LINE[35]模型来学习每个二部图的嵌入。该方法虽然捕捉了异构图中不同类型节点之间的语义信息,没有直接

语义关联但语义密切相关的节点被忽略,导致同构语义信息的丢失。因此,如何整合图结构中不同邻域节点是异构网络嵌入研究的挑战之一。

基于元路径的网络嵌入方法能够将元路径和邻居节点的信息融合到目标节点的嵌入向量中。然而,手动确定元路径可能会导致这些方法的性能不稳定。基于边类型聚合邻居信息的方法大多将异构图映射为同构子图的策略,以获取目标节点的高阶邻居信息。然而,这种做法不仅会忽略异质邻居提供的信息,而且可能导致语义混淆。实际上,节点的高阶信息需要通过堆叠网络层来捕获,当异构图的层数增加时,节点将不断积累越来越多的噪声信息。因此,本文提出了基于子图结构邻近嵌入的方法。首先将异构图解耦为关系异构子图和基于元路径的同构子图;其次,在不同子图结构中对目标节点的异构邻居和同构邻居分别聚合;最后,从同质和异质邻居中全面捕获结构、语义和属性信息,生成目标节点的嵌入。

3　网络视频事件挖掘方法

本文提出的视频事件挖掘方法如图 2 所示,框架包含 4 个步骤,分别是异构网络构建、异构图变换、局部聚合和全局聚合。

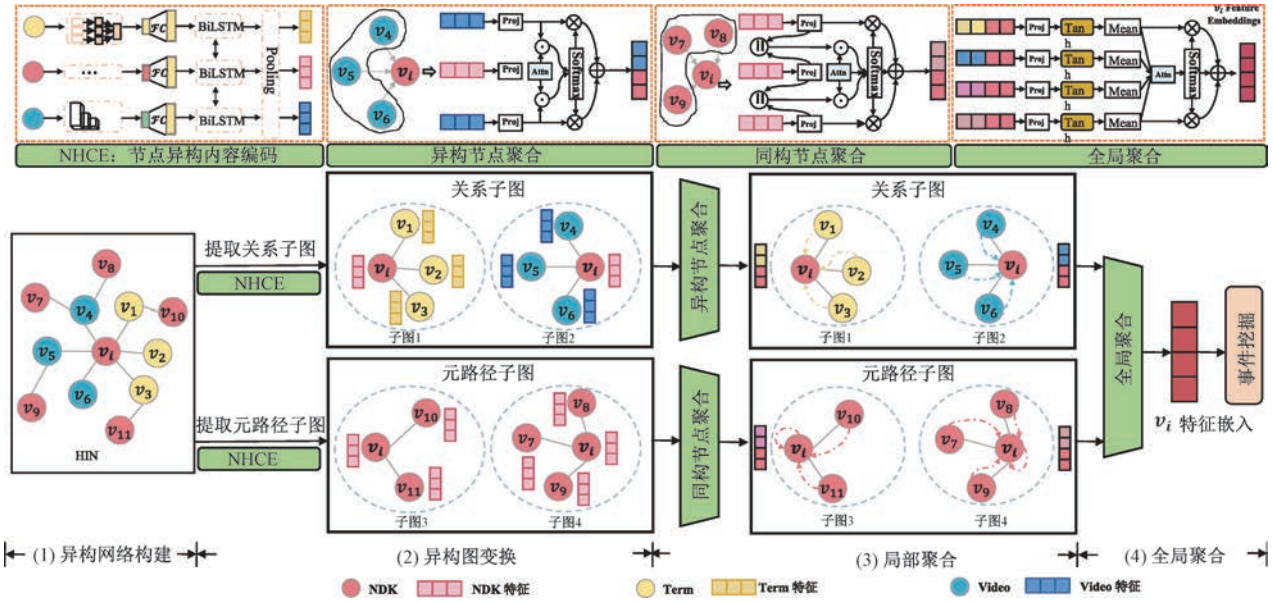

图 2　视频事件挖掘框架(该框架包含四个部分:(1)异构网络构建:本文以 NDK(视觉近似关键帧集合)、Term(文本)、Video(视频)为节点类型,并定义 NDK-Term、NDK-Video 为边关系构建多模态异构网络。(2)异构图变换:对异构网络进行解耦,生成基于特定关系的异构子图和基于元路径的同构子图,同时对节点的异构内容特征进行编码。(3)局部聚合:利用注意力机制,聚合异构子图和同构子图的邻域信息,提取局部语义关系。(4)全局聚合:通过语义注意力机制,整合多种语义信息和关系信息,从而生成节点的最终嵌入表示。)

3.1　概念定义

本文所提方法利用异构网络来关联网络视频中的文本信息与视觉信息,并利用子图结构对异构网络进行嵌入。本节将对所涉及到的基本概念进行介绍。

定义 1. *NDK*(Near-Duplicate Keyframes)是从多个视频中提取的相似关键帧组成的集合,这些帧集中体现了视频镜头的关键视觉信息。受场景转换、拍摄视角、光线变化以及后期剪辑等多种因素的作用,各 NDK 集合之间展现出明显的视觉区别。借助 NDK,能够有效追踪视觉特征相近的视频片段,这一技术已在事件挖掘研究中被广泛采用[48]。

定义 2. 异构信息网络(Heterogeneous Information Network,HIN). 给定异构信息网络 $G=(V,E)$,其中 V 是节点集合,E 是边集合。异构信息网络与两种类型的映射函数有关,包括节点类型映射函数 $\varphi:V\rightarrow A$ 和边类型映射函数 $\psi:E\rightarrow R$,其中 A 和 R 分别表示节点和边类型的集合,满足 $|A|+|R|$[43]。本文的多模态异构信息网络包括三种不同媒体类型节点和两种边类,$A\in(Video,NDK,Term)$,$R\in(Video\text{-}NDK,Term\text{-}NDK)$。其中,当 $|A|+|R|=2$ 时,异质图将变为同质图。

定义 3. 元路径(Meta Path). P 在异构信息网

络中定义成形式 $A_1 \overrightarrow{R_1} A_2 \overrightarrow{R_2} \cdots \overrightarrow{R_l} A_{l+1}$（缩写为 $A_1A_2A_{l+1}$）。它描述了节点类型 A_1 到 A_{l+1} 的复合关系 $R=R_1 \circ R_2 \circ \cdots \circ R_l$。其中。表示两个节点之间的连接关系。

定义 4. 基于元路径的子图(Meta Path-based Subgraph). 给定元路径 P，其起始节点类型为 A_1，结束节点类型为 A_{l+1}，基于元路径的图 G_P 是由所有节点对 $v \in A_1$ 和 $u \in A_{l+1}$ 构建的图，它们通过元路径 P 连接。其中，若元路径 P 起始节点和结束节点类型相同，则定义 G_P 是同构子图，否则为异构子图。例如，基于元路径的子图如图 3(c)所示。

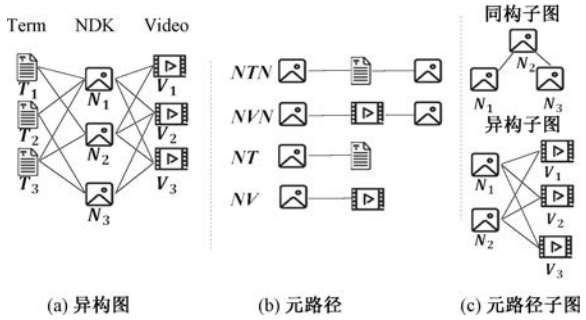

图 3 异构信息网络示意图

3.2 异构网络构建

首先，分析文本在 NDK 中的分布特性。对过滤掉无关停用词后的每个主题构建一个度量矩阵，其中每一行对应一个 NDK，每一列对应一个单词。通过 TF-IDF 方法计算每个单词在所有 NDK 中的分布特性，并在最后一列标注每个 NDK 所属的实际事件。生成的索引矩阵可以表示为一个二维表格，其元素定义如下：

$$NT_{i,j}=\frac{TF_{i,j}}{N_j}\times\log\frac{N}{DF_i} \tag{1}$$

其中，$TF_{i,j}$ 为单词 T_i 在 NDK_j 中出现的频率，N_j 表示 NDK_j 中分布的所有单词数量，DF_i 表示包含单词 T_i 的 NDK 数量，N 表示 NDK 总数。从索引矩阵中筛选出 TF-IDF 大于 0 的值对应为 NDK-Term。

其次，计算视频在 NDK 中的分布特征。对每一个视频提取关键帧并进行 NDK 聚合后为每个主题形成一个度量矩阵，每个 NDK 为行，每个视频为列。该矩阵中的每个值用于判断 NDK 之间是否存在共同的视频。若两个 NDK 中包含更多来自同一视频的关键帧，则它们之间建立语义关联的可能性更高，从而有助于发现更丰富的场景信息。如果两个 NDK 中有更多的关键帧来自同一个视频，这意味着它们之间更有可能建立语义交互，即帮助找到更多的场景信息。

3.3 异构图变换

3.3.1 异构节点内容编码

本文构建的异构信息网络涵盖三种节点类型，分别是 V_N（表示 NDK），V_T（表示 Term）和 V_V（表示 Video）。这些节点携带着非结构化的异构内容。例如，在图 3 中可见，(Term 节点)承载文本信息，而(NDK 节点)则携带视觉信息。然而，不同类型节点之间的异构内容无法直接交互。受文献[4]的启发，本文提出了节点内容编码器，旨在解决异构信息网络中不同节点内容间的异构性问题。

本文提出了一个模块，用于从节点 $v_i \in V$ 中提取异构内容 C_V，并利用神经网络 f_1 将其映射为固定维度的嵌入表示。为提取文本内容特征，文献[49]提出了将每个单词或短语转换为向量表示的方法。对于图片内容特征的提取，则利用 CNNs[50] 进行预训练以获取关键帧视觉特征。至于视频内容特征的提取，文献[51]提供了相应的模型。基于此，我们得到异构内容 C_V。本文将 C_V 中第 i 个内容的特征表示为 $x_i \in R^{d_f \times 1}$（d_f：内容特征维度）。为了使不同节点类型之间的异构内容“深度”交互，并获得更大表达能力，本文采用基于双向 LSTM(Bi-LSTM)[52] 的架构，以捕获节点深层特征交互。首先使用不同的 $\mathcal{FC}$ 层来转换不同节点内容特征；其次，使用 Bi-LSTM 来捕获“深层”特征交互；最终，通过对所有隐藏状态应用平均池化操作，生成节点 $v_i \in V$ 的内容嵌入表示。节点内容嵌入计算如下：

$$f_1(v)=\frac{\sum_{i\in C_v}\left[\overrightarrow{LSTM}\{\mathcal{F}C_{\theta_x}(x_i)\}\oplus\overleftarrow{LSTM}\{\mathcal{F}C_{\theta_x}(x_i)\}\right]}{|C_v|} \tag{2}$$

其中，$f_1(v)\in R^{d\times 1}$（d：内容嵌入维数），$\mathcal{F}C_{\theta_x}$ 表示特征转换器，参数 θ_x 为全连接神经网络，$\oplus$ 表示向量加法。

3.3.2 子图提取

(1)关系子图提取

在异构信息网络中，节点的一阶邻居关系反映了节点间直接交互关系。然而，在异构图中直接应用图卷积层来聚合邻居信息并不能充分捕获节点之间细粒度的语义结构关系。因此，为了更好地捕获目标节点的一阶语义关系和隐藏结构关系，本文对目标节点的关系子图进行了提取。这一过程包含三个步骤。

首先根据边的类型对异构网络进行解耦。令 $\{\boldsymbol{M}_r \in R^{m\times n} \mid r=1,2,\cdots,|R|\}$ 表示生成所有关系子图的基本邻接矩阵,其中 m 和 n 是关系节点的个数。其次,从所有关系邻接矩阵中筛选出以 NDK 节点作为目标节点的关系邻接矩阵。最后,关系子图提取。关系邻接矩阵反映了与目标节点相关联的异质节点一阶邻接关系。如图 4(i)展示了关系子图的提取过程。

(2)元路径子图提取

在上一步中,虽然提取了目标节点的一阶邻居语义信息和结构信息,但异构图中隐藏的高阶信息往往被忽视。元路径因其强大的语义捕获能力而常被用于提取异构图中的高阶信息。然而,元路径仅仅是简单的语义游走,并未充分表达在游走过程中产生的空间结构信息。因此,基于元路径的子图结构不仅反映了目标节点的高阶语义信息,而且能够获取更细粒度的语义关联和空间结构关系。如图 4 所示,展示了元路径子图提取的三个过程。

首先,提取指定关系的邻接矩阵。在原始异构图中包含 V_N(表示 NDK),V_T(表示 Term)和 V_V(表示 Video)三种节点,以及四种关系 R_{NT}(NDk-Term),R_{NV}(NDk-Video),R_{TN}(Term-NDK)和 R_{VN}(Video-NDK)。提取这四种关系的邻接矩阵,邻接矩阵的提取可以表示为

$$\boldsymbol{M}_{v_i,v_j} = \boldsymbol{ExtractSubgraph}(V_{v_i},V_{v_j},R_{ij}) \tag{3}$$

其中,$\boldsymbol{ExtractSubgraph}(V_{v_i},V_{v_j},R_{ij})$ 表示从原始异构图中提取关系为 R_{ij} 的邻接矩阵。v_i 和 v_j 分别表示不同节点类型。

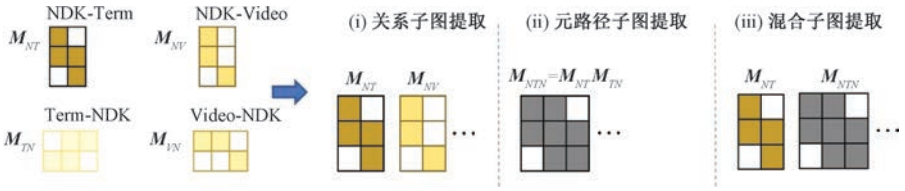

图 4　子图提取

其次,在得到不同关系的邻接矩阵后,构建基于邻接矩阵的元路径。由于网络拓扑结构与邻接矩阵一一对应,通过邻接矩阵相乘将提取不同语义的元路径。邻居聚集过程可以表示为:

$$\boldsymbol{M}^P = \hat{\boldsymbol{M}}_{v_0,v_1}\hat{\boldsymbol{M}}_{v_1,v_2}\cdots\hat{\boldsymbol{M}}_{v_{j-1},v_j} \tag{4}$$

其中,$P=v_0v_1v_2\cdots v_l$ 是 l 跳元路径,$\hat{\boldsymbol{M}}_{v_i,v_j}$ 是节点类型 v_i 和 v_j 之间邻接矩阵。

最后,元路径子图提取。在前一步得到不同语义的元路径后,提取由首尾节点类型相同的邻接矩阵相乘后得到的元路径子图为同质子图,相乘之后的邻接矩阵即为同质节点之间的邻接关系。该元路径子图反应了节点的不同邻居之间的高阶关系。

经过上述步骤后,从异构图中提取出包含关系结构的异构子图和包含节点间高阶关系的同构子图。具体表示为

$$G=\{G_{h_o}\cup G_{h_e}\} \tag{5}$$

其中,G 是原始异构图,G_{h_o} 是同构子图集合,G_{h_e} 是异构子图集合。

3.4　局部聚合

3.4.1　同构子图节点聚合

在上一节完成节点类型的特征交互后,本文采用自注意力机制[53]来学习节点间的权重关系。给定同构子图 $S_n^{h_o}$,对子图中的节点对 (i,j) 施加注意力层,节点级注意力 e_{ij} 可以学习节点 v_j 对节点 v_i 的重要性。

在计算节点级注意力 e_{ij} 之前,我们首先应用共享权重矩阵 $\boldsymbol{W}_r \in R^{d\times d'}$ 来转换 v_i 的输入特征,其中 d 和 d' 分别是输入层的维度和嵌入后维度:

$$\boldsymbol{z}_i = \boldsymbol{W}_r \cdot \boldsymbol{h}_i \tag{6}$$

其中 $\boldsymbol{z}_i \in R^{d'}$ 是 v_i 的中间嵌入。然后,基于子图的节点对 (i,j) 的重要性可以表述如下:

$$\boldsymbol{e}_{ij} = LeakyReLu(\boldsymbol{W}_{ij}[\boldsymbol{z}_i \| \boldsymbol{z}_j]) \tag{7}$$

其中,$\|$ 是向量拼接,$W_{ij} \in R^{2d'\times d'}$ 是权重矩阵,鉴于相同子图结构中连接方式相似,所有基于子图的节点对共享注意力机制。上述公式(7)表明,给定子图 $S_n^{h_o}$ 中节点对 (i,j) 的权重取决于其特征。

为了提取子图结构信息,我们首先计算节点 $j \in s_n^i$ 的重要性 $\boldsymbol{e}_{ij}$,其中 s_n^i 表示节点 v_i 基于子图 $s_n^{h_o}$ 的邻居(包括自身)。然后以 $softmax$ 函数对重要性值归一化,得到权重系数 α_{ij}:

$$\alpha_{ij} = softmax(\boldsymbol{e}_{ij}) = \frac{\exp(\boldsymbol{e}_{ij})}{\sum_{k\in\mathcal{N}_{i,r}}\exp(\boldsymbol{e}_{ik})} \tag{8}$$

其中,$\mathcal{N}_{i,r}$ 表示关系 r 关联的同构子图中节点 v_i 的邻居节点集合,α_{ij} 是 v_i 和 v_j 之间的注意力权重系数去线性组合邻居节点的特征。我们通过结合 v_i 和 v_j 之间的相互作用,进一步使其适应事件挖掘任

务,以便可以找到分散的视觉信息。

随后,节点 v_i 的嵌入表示可以通过聚合邻居节点的投影特征来表征,对应的权重系数如下公式所示:

$$\boldsymbol{h}_i^{s_n}=\sigma\Big(\sum_{j\in\mathcal{N}_{i,r}}\alpha_{ij}\cdot\boldsymbol{z}_j\Big) \tag{9}$$

其中,$\boldsymbol{h}_i^{S_n}$ 是子图 $s_n^{h_o}$ 中节点 v_i 的最终向量。每个节点的嵌入都是通过聚合其邻居节点的信息得到的。由于注意力权重 α_{ij} 是针对特定子图计算得出的,因此它们具有语义特异性,并能够对特定的语义信息进行编码。

为了减少由于网络异质性产生节点输出特征差异以及增强模型的表达能力,本文还将其扩展到多头注意力,使训练过程更加稳定。其流程制定如下:

$$\boldsymbol{h}_i^{s_n}=\Big\|_{k=1}^{k}ReLu\Big(\sum_{j\in\mathcal{N}_{i,r}}\alpha_{ij}\cdot\boldsymbol{z}_j\Big) \tag{10}$$

其中,$\boldsymbol{h}_i^{s_n}$ 是子图 $s_n^{h_o}$ 中节点 v_i 的学习嵌入,α_{ij} 是 v_i 和 v_j 之间的注意力权重,$z_j\in R^{d'}$ 是 v_j 的中间嵌入,$\mathcal{N}_{i,r}$ 表示关系 r 关联的同构子图中节点 v_i 的邻居节点集合。

给定同质子图集 $\{\boldsymbol{s}_1^{h_o},\cdots,\boldsymbol{s}_n^{h_o}\}$,通过将节点特征输入节点级注意力机制,可生成 n 组与同质子图语义相关的节点嵌入,表示为 $\{\boldsymbol{h}^{s_1^{h_o}},\cdots,\boldsymbol{h}^{s_n^{h_o}}\}$。

3.4.2 异构子图节点聚合

受 GraphSAGE[35] 的启发,由于目标节点邻居的无序性,节点聚合函数则对一组无序向量进行操作。在异构子图聚合中,本文提出了 3 个候选聚合器。

(1)均值聚合

基于均值操作的聚合中,对邻居节点特征进行平均池化来作为目标节点的输出特征。对于异构子图 $S_n^{h_e}$ 中的节点 v_i,$\mathcal{N}_{i,r}$ 表示关系 r 关联的异构子图中节点 v_i 的邻居节点集合。均值聚合操作具体如下:

$$\boldsymbol{h}_{i,r}=\sigma(MEAN(\{\boldsymbol{z}_j,\forall j\in\mathcal{N}_{i,r}\})) \tag{11}$$

其中,$\boldsymbol{h}_{i,r}$ 表示 v_i 通过关系 r 的学习嵌入,σ 是激活函数,$\mathcal{N}_{i,r}$ 表示关系 r 关联的异构子图中节点 v_i 的邻居节点集合,$z_j\in R^{d'}$ 是 v_j 的中间嵌入。

(2)池化聚合

基于池化操作的聚合中,为提取目标节点邻居节点的重要特征,我们采取最大化池化操作,通过一个全连接层和 ReLU 激活函数对目标节点的邻居节点特征进行聚合,选取池化后特征中每个维度的最大值作为目标节点的输出特征。对于异构子图 $\boldsymbol{S}_n^{h_e}$ 中的节点 v_i,$\mathcal{N}_{i,r}$ 表示关系 r 关联的异构子图中节点 v_i 的邻居节点集合。最大化池化操作具体过程如下:

$$\boldsymbol{h}_{i,r}=\max(\{\sigma(\boldsymbol{W}_{pool}\boldsymbol{z}_j+\boldsymbol{b}_{pool},\forall j\in\mathcal{N}_{i,r})\}) \tag{12}$$

其中,σ 是非线性激活 ReLU 函数,max 是最大化池化算子,$W_{pool}\in R^{d'}$ 和 $b_{pool}\in R^{d'}$ 是学习的参数,$z_j\in R^{d'}$ 是 v_j 的中间嵌入,$\mathcal{N}_{i,r}$ 表示关系 r 关联的异构子图中节点 v_i 的邻居节点集合。

(3)注意力聚合

基于注意力机制的聚合中,为计算异质子图 $\boldsymbol{S}_n^{h_e}$ 中 v_i 的关联特异性嵌入 $h_{i,r}$,我们首先应用共享权重矩阵 $\boldsymbol{W}_r\in R^{d\times d'}$ 来转换 v_i 的输入特征,其中 d 和 d' 分别是输入层的维度和嵌入后维度:

$$\boldsymbol{z}_i=\boldsymbol{W}_r\cdot\boldsymbol{h}_i \tag{13}$$

其中,$z_i\in R^{d'}$ 是 v_i 的中间嵌入。然后,我们计算 v_i 和 v_j 之间的非归一化注意力系数 e_{ij},并使用 softmax 函数对 e_{ij} 进行归一化:

$$\boldsymbol{e}_{ij}=LeakyReLu(\boldsymbol{W}_{ij}\cdot\boldsymbol{z}_j) \tag{14}$$

$$\alpha_{ij}=softmax(\boldsymbol{e}_{ij})\frac{\exp(\boldsymbol{e}_{ij})}{\sum_{k\in\mathcal{N}_{i,r}}\exp(\boldsymbol{e}_{ik})} \tag{15}$$

其中,$\boldsymbol{W}_{ij}\in R^{2d'\times d'}$ 是权重矩阵,$\mathcal{N}_{i,r}$ 表示关系 r 关联的异构子图中节点 v_i 的邻居节点集合,α_{ij} 是 v_i 和 v_j 之间的注意力权重,作为系数来线性组合相邻节点的特征。

本文还将其扩展到多头注意力,使训练过程稳定。其流程制定如下:

$$\boldsymbol{h}_i^{s_n}=\Big\|_{k=1}^{k}ReLu\Big(\sum_{j\in\mathcal{N}_{i,r}}\alpha_{ij}\cdot\boldsymbol{z}_j\Big) \tag{16}$$

其中,$h_i^{s_n}$ 是子图 $s_n^{h_e}$ 中节点 v_i 的学习嵌入,α_{ij} 是 v_i 和 v_j 之间的注意力权重,$z_j\in R^{d'}$ 是 v_j 的中间嵌入,$\mathcal{N}_{i,r}$ 表示关系 r 关联的异构子图中节点 v_i 的邻居节点集合。

给定关系异质子图集 $\{s_1^{h_e},\cdots,s_n^{h_e}\}$,通过将节点特征输入节点级注意力机制,可生成 n 组基于异质子图语义特定的节点嵌入,表示为 $\{h^{s_1^{h_e}},\cdots,h^{s_n^{h_e}}\}$。

3.5 全局聚合

在上一步中我们得到子图内节点的特征后,接下来对基于子图的节点特征进行全局聚合。鉴于子图中节点属于不同结构并表示不同语义关系,为衡量不同子图结构和语义重要性,本文使用全局注意

力机制来学习他们之间的权重。然后,利用这些权重对不同结构中节点的特征进行聚合。

3.5.1　全局聚合

为了评估不同子图的结构和语义重要性,本文首先通过共享非线性权重矩阵 $\boldsymbol{W}_R \in \boldsymbol{R}^{d' \times d'}$ 对子图中的节点嵌入进行转换。随后,我们利用转换后的嵌入与语义级注意力向量 $\boldsymbol{q} \in \boldsymbol{R}^{d'}$ 之间的相似性,来量化每个子图节点嵌入的重要性。最后,我们对所有子图嵌入的重要性进行了平均计算,每个子图的重要性(记作为 $\omega_{i,r}$)如下所示:

$$\omega_{i,r} = \frac{1}{|V_r|} \sum_{i \in V_r} \boldsymbol{q}^{\mathrm{T}} \cdot \tanh(\boldsymbol{W}_R \cdot h_{i,r}^{l+1} + b) \quad (17)$$

其中,V_r 代表与关系 r 相关联的节点集合,$\boldsymbol{R}^{d'}$ 为偏置向量,$\boldsymbol{W}_R \in \boldsymbol{R}^{d' \times d'}$ 为权重矩阵,$\boldsymbol{q} \in \boldsymbol{R}^{d'}$ 为语义级注意力向量。

在得到每个子图的重要性之后,接下来要对其进行归一化处理,本文利用 softmax 函数来操作。在对 $\omega_{i,r}$ 归一化后,将得到最终的语义级注意力权重系数 $\beta_{i,r}$:

$$\beta_{i,r} = \mathrm{softmax}(\omega_{i,r}) = \frac{\exp(\omega_{i,r})}{\sum_{r \in R_i} \exp(\omega_{i,r})} \quad (18)$$

其中,R_i 是 v_i 的关联关系,$\beta_{i,r}$ 表示 v_i 的贡献。我们可以得出,当 $\beta_{i,r}$ 值越高时,子图 s_n 重要性越高。

接下来,我们将上述公式中计算得到的权重系数用于融合特定语义的嵌入,并得到最终嵌入 $\boldsymbol{h}_i^G$,如下所示:

$$\boldsymbol{h}_i^G = \sum_{i=1}^{|R_i|} \beta_{i,r} \cdot \boldsymbol{h}_{i,r}^{s_n} \quad (19)$$

最终的嵌入表示是通过聚合所有基于子图结构和语义信息的嵌入生成的。然后,我们将这一嵌入应用于网络视频事件挖掘任务中。该模型的伪代码如下:

算法 1. 网络视频突发事件挖掘整体框架算法

输入:异构图 $G = (V, E)$,节点编码后整体特征 $f_1(v)$,元路径 P ,邻接矩阵 $\boldsymbol{M}_{v_i, v_j}$

输出:最终嵌入 h_i^G

1. FOR 每个节点 $v \in V$

2. 通过元路径 P 和邻接矩阵 $\boldsymbol{M}_{v_i, v_j}$ 构建同构子图 G_{h_o} 和异构子图 G_{h_e}

3. FOR 每个子图 $G_s \in \{G_{h_o}, G_{h_e}\}$

4. 使用子图特定的聚合方法计算节点嵌入:

5. $\boldsymbol{h}_i^{s_n} \leftarrow \Vert_{k=1}^{k} ReLu\left(\sum_{j \in N_{i,r}} \alpha_{ij} \cdot \boldsymbol{z}_j \right)$

6. END FOE

7. 计算每个子图的权重 $\beta_{i,r}$ 并融合不同子图的嵌入:

8. $\boldsymbol{h}_i^G \leftarrow \sum_{i=1}^{|R_i|} \beta_{i,r} \cdot \boldsymbol{h}_{i,r}^{s_n}$

9. END FOR

3.5.2　事件挖掘

经过前面的节点聚合和子图嵌入,得到每个 NDK 节点的向量表示矩阵 $\boldsymbol{h}_i^G$ 。

本文以反向传播算法和梯度下降算法为基础最小化交叉熵损失。所有标注节点的真实标签与预测标签之间的交叉熵损失定义如下:

$$Loss = -\sum_{v \in V_L} y_v \log(C \cdot \boldsymbol{h}_i^G) \quad (20)$$

其中,V_L 表示标签顶点集,y_v 表示顶点 v 的真实值,C 表示分类器参数,$\boldsymbol{h}_i^G$ 表示嵌入向量。最终,在标记节点的指导下,我们通过训练模型以最小化损失函数($Loss$),从而优化所提出的模型。在此基础上,我们进一步学习节点的嵌入表示,以用于事件挖掘任务。

3.5.3 复杂度分析

我们提出的模型需要计算三个进程的时间。首先,节点特征转换的时间复杂度为 $O(d_A \times d')$ 。其中 d_A 表示节点特征的初始维度,d' 表示隐藏层的维度。其次,将原始图分解为多个子图的时间复杂度与原始图 $\boldsymbol{G}$ 中的节点数呈线性关系,其时间复杂度为 $O(2 \mid A \mid\mid P \mid\mid V_A \mid)$ 。其中 A 是异构图中节点类型集合,P 为生成子图的元路径。V_A 是节点类型 A 的所有集合。然后,在节点聚合模块中,该模块的时间复杂度为 $O(2nd_A d')(c_1 + c_2)$ 。其中 c_1 和 c_2 分别表示同构图和异构图的个数,d_A 表示节点特征的初始维度,d' 表示隐藏层的维度。两种不同类型的子图在节点聚合时可以并行展开。最后,当聚合多个子图的信息时,时间复杂度是 $O(\mid A \mid\mid V_A \mid\mid S_n \mid)$ 。其中 S_n 为子图个数。

4　实验分析

4.1　数据集

在对 CNN、Time 和新华网推荐的前十大新闻主题进行综合评估后,本文选取了 10 个重要主题以构建实验数据集。具体数据如表 2 所示。我们将这些话题发布到 YouTube 上并爬取相关视频,手动选择并确定视频所属的话题。通常,每个话题都涵盖多个不同的事件。数据集的具体信息如下:共包含了 6 187 个网络视频、19 610 个 NDK 以及 15 910

个术语。这些术语主要来源于网络视频的标题和标签。所选话题覆盖了体育、社会、军事、科技、经济等多个领域。此外，不同话题的持续时间存在显著差异：有些话题可能持续数年（如乌克兰危机、叙利亚战争），而有些话题则较为短暂（如马拉多纳去世、2016 年里约奥运会）。文中数据真实有效，能够全面验证方法的有效性。

表 2 实验数据集

编号	话题	视频数	NDKs 数	文本的单词数	事件
1	金融危机	1 025	7 692	3 946	16
2	科比坠机	551	1 215	1 076	7
3	嫦娥 5 号	602	1 614	874	7
4	里约奥运会	547	1 123	1 226	4
5	克里米亚危机	627	1 174	1 518	8
6	加泰罗尼亚独立	619	1 261	1 435	9
7	乌克兰危机	481	1 036	1 005	6
8	叙利亚战争	758	2033	1 787	5
9	伊朗危机	575	1 292	1 565	6
10	马拉多纳之死	402	1 170	1 478	5
总计		6 187	19 610	15 910	73

4.2 评价指标

本文采用精确率、召回率和 $F1$ 来验证其有效性。

$$\text{精确率}=\frac{|Y_i^+|}{X_i} \tag{21}$$

$$\text{召回率}=\frac{|Y_i^+|}{Y_i} \tag{22}$$

$$F1=\frac{2\times\text{精确率}\times\text{召回率}}{\text{精确率}+\text{召回率}} \tag{23}$$

X_i 为要检测的目标事件，Y_i 为与 X_i 最相似的真实事件，Y_i^+ 为正确检测到的事件，精确率为正确检测数据所占比例，召回率为正确样本所占比例。由于 $F1$ 值能够更好地平衡召回率与精度。因此，本文以 $F1$ 核心指标。

4.3 对比实验

本文与多种经典方法进行了对比与分析，这些方法可以分为两类。第一类以文本和视觉信息的利用方式划分，具体包括：仅使用文本信息的模型（LDA）[12]、仅使用视觉信息的模型（VSDA）[17]以及同时结合文本和视觉信息的模型（MMBT[24]、CoNe[20]）。第二类以网络嵌入方法划分，包括同构网络嵌入（DeepWalk[40]）和异构网络嵌入（MP2vec[37]、MHGCN[54]、muxGNN[55]、HeGAE-AC[56]）。以下是对比方法的具体说明。

（1）LDA[12]：该方法主要以文本信息挖掘事件，在推测文档主题分布后，计算视频主题分布情况。我们将视频中的短文本信息输入到 LDA 模型，得到文本－事件分布向量，然后进行事件挖掘。

（2）VSDA[17]：该方法基于注意力模型提取视频局语义特征和全局语义特征，它通过计算视频关键帧之间的相似度来获得视频之间的相关性，将相似度高于阈值的视频归入同一个事件。

（3）MMBT[24]：该方法通过联合微调多模态预训练文本和图像编码器将图像嵌入文本空间，利用图像信息弥补稀疏文本信息实现事件挖掘。

（4）CoNe[20]：该方法不仅通过其类中心对每个样本进行监督，同时考虑样本与相似邻居样本的相关性，以生成更具适应性和精细化的目标，从而将相同视觉表达归入同一事件。

（5）DeepWalk[40]：作为一种基于随机游走的同构网络嵌入方法，它通过截断随机游走捕获网络的局部结构信息。在本研究中忽略了图的异构性，模型将 NDKs、Terms 和 Videos 之间的关系转化为向量形式，从而实现节点的低维表示。

（6）MP2vec[37]：该方法主要以 NDK、文本和视频间异构关系建模。该方法主要以元路径引导的随机游走算法生成节点嵌入表示。

（7）MHGCN[54]：该方法设计了一个多元异构图卷积模块来自动提取元路径信息，并通过多层卷积聚合不同长度元路径的语义信息，得到节点的最终嵌入并应用于下游任务。

（8）muxGNN[55]：这是针对异构网络的嵌入模型，该方法设计特定的耦合注意力机制整合图中一阶关系和高阶关系，以学习异构图中不同类型节点和边关系的重要性，并输出目标节点的最终嵌入。

（9）HeGAE-AC[56]：该方法通过设计一个异构图自编码器，将图的拓扑结构和节点的高阶属性信息整合到低维嵌入中。具体而言，它利用解码器重建目标节点的属性嵌入，并最终借助图神经网络模型生成目标对象（如 NDKs）的嵌入表示。

4.4 实验设置

在本文中，我们采用随机初始化的方式设置模型参数，并利用 Adam[57] 优化器对模型进行训练。具体超参数配置如下：学习率设定为 0.001，权重衰减系数为 0.0005，正则化参数为 0.001，注意力机制中的头数 K 设为 8，注意力 dropout 比率为 0.6，提前停止的耐心值（patience）设为 20。对于 MMBT[24] 和 CoNe[20]，我们通过验证集来调整和优化它们的超参数。在图神经网络领域，针对 MHGCN[54]、muxGNN[55] 和 HeGAE-AC[56]，本文划分了完全一

致的训练集、验证集与测试集,以确保实验的公平性。对于基于随机游走的 DeepWalk[40] 和 Metapath2vec[37],窗口大小为 5,游走步长为 100,每个节点执行 40 次游走,负采样数为 5。为保障对比的公平性,所有算法的嵌入维度均统一设置为 64。

4.5　实验结果及分析

本文对表 3～表 5 中的精确率(表 3)、召回率(表 4)以及 $F1$ 值(表 5)进行了对比分析,其中最优结果已加粗标注。基于这些表格数据,可以得出以下观察与结论。

表 3　精确率对比

话题	LDA	VSDA	MMBT	CoNe	Deepwalk	MP2vec	MHGCN	muxGNN	HeGAE-AC	本文
1	0.31	0.65	0.60	0.54	0.07	0.35	0.50	0.45	0.39	**0.83**
2	0.37	0.72	0.50	0.52	0.46	0.62	0.43	0.58	0.41	**0.77**
3	0.06	0.75	0.52	0.63	0.62	0.46	0.38	0.63	0.48	**0.97**
4	0.55	0.84	0.78	0.55	0.47	0.52	0.67	0.70	0.59	**0.85**
5	0.16	0.75	0.27	0.23	0.42	0.61	0.60	0.67	0.39	**0.94**
6	0.44	0.71	0.34	0.50	0.53	0.59	0.48	0.52	0.55	**0.90**
7	0.53	0.70	0.60	0.53	0.57	0.65	0.61	0.72	0.79	**0.79**
8	0.36	0.74	0.56	0.51	0.70	0.58	0.62	0.67	0.82	**0.83**
9	0.67	0.88	0.45	0.23	0.82	0.83	0.65	0.70	0.65	**0.98**
10	0.34	0.79	0.49	0.34	0.74	0.63	0.62	0.75	0.77	**0.95**
平均	0.38	0.75	0.51	0.46	0.54	0.58	0.55	0.64	0.58	**0.88**

LDA[12] 在精确度、召回率和 $F1$ 表现均不佳,尤其是平均召回率只有 0.15。这是因为 LDA[12] 模型对长文本更敏感,而在短文本上效果较差.网络视频中的文本信息通常由标题和标签组成,相较于传统新闻文本,文本信息较少,且存在大量噪声。文本信息稀缺且导致文本共线频率低,进而影响了该模型的检测效果。因此,仅仅使用文本信息进行网络视频事件挖掘效果不佳。

VSDA[17] 仅利用视觉信息进行相似度检测,在精确度方面表现较好,平均精确度达到 0.75。相比仅使用文本信息,视觉呈现的特征信息更客观全面,能够检测出具有相似视觉信息的视频事件。然而,在召回率方面效果不佳,平均召回率仅为 0.18。这是因为同一事件通常由视觉信息丰富的不同场景构成,视频关键帧的相似性检测只能检测出同一事件中非常相似的画面,而丰富的不同场景通常会漏检。基于关键帧的相似性检测无法将同一事件中不同视觉场景关联起来。然而,视觉信息客观全面、精确率高的特点使得我们可以利用视觉信息来弥补文本信息的稀疏性。

MMBT[24] 将视觉信息嵌入到文本空间中,以连接预训练的文本和关键帧编码器。与 VSDA[17] 相比,MMBT[24] 在平均召回率和 $F1$ 上有所提高,表明利用跨媒体信息融合增强了视觉信息与文本信息间的关联。网络视频中视觉信息场景丰富,呈现非线性的特点,MMBT[24] 在融合文本信息中可以将丰富的视觉信息串联起来,进而在召回率中有所提高,降低了漏检率。然而,召回率和精确率之间存在较大差异,最佳精确率为 0.78,而最差值仅为 0.27。这主要是因为网络视频通常由上传者二次编辑,时空跳跃且情节破碎,其中描述事件相关核心视觉信息的比例较小,而与事件关联性较小的信息占比较大,从而导致了视觉信息增强文本信息时引入了一定噪声。因此,收集与主题相关的不同核心视觉信息变得至关重要,这有助于从视觉特征中获取更多有用信息,从而提高事件挖掘的效果。

CoNe[20] 利用交叉熵损失来确保类内样本具有相同的目标。随后,它利用经过训练的特征来构建更精细的目标,以提高类内样本的紧凑性。表 3 显示,最佳精度为 0.63,最差精度仅为 0.23。这归因于同一主题的不同关键帧之间的视觉差异。在这种情况下,视觉特征会失去适应性,无法涵盖主题类别内的差异。总体而言,其准确性不如其他方法。

DeepWalk[40] 是一种经典的同构网络嵌入学习方法,我们在研究中重新实现了该方法,用于生成我们构建的网络中 NDK 节点的嵌入表示。通过随机游走,DeepWalk[40] 能够将分散的 NDK 节点联系起来,为网络视频事件挖掘任务中的节点分类提供了基础支持。据表 4 的结果,尽管 DeepWalk[40] 的表现未达到预期,但其性能仍优于仅使用文本信息的 LDA[12] 模型和仅依赖视觉特征的 VSDA[17] 模型。这一结果表明,网络嵌入方法在融合文本与视觉信息方面展现出一定的优势。

表 4 召回率对比

话题	LDA	VSDA	MMBT	CoNe	Deepwalk	MP2vec	MHGCN	muxGNN	HeGAE-AC	本文
1	0.14	0.13	0.32	0.37	0.09	0.15	0.44	0.45	0.59	**0.70**
2	0.16	0.22	0.52	0.54	0.42	0.45	0.34	0.48	0.31	**0.56**
3	0.11	0.14	0.78	0.54	0.50	0.37	0.31	0.59	0.78	**0.94**
4	0.23	0.13	0.55	0.71	0.37	0.40	0.40	0.60	0.68	**0.68**
5	0.30	0.21	0.69	0.52	0.37	0.53	0.56	0.67	0.54	**0.83**
6	0.12	0.11	0.64	0.59	0.40	0.37	0.36	0.52	0.80	**0.73**
7	0.14	0.30	0.70	0.51	0.39	0.41	0.33	0.52	0.72	**0.77**
8	0.09	0.08	0.38	0.50	0.52	0.44	0.39	0.59	0.53	**0.69**
9	0.13	0.19	0.57	0.52	0.44	0.42	0.32	0.46	0.61	**0.70**
10	0.04	0.28	0.59	0.35	0.61	0.53	0.46	0.68	**0.69**	0.68
平均	0.15	0.18	0.57	0.52	0.41	0.41	0.39	0.56	0.63	**0.73**

MP2vec[37]在多个话题中精确率和召回率均优于DeepWalk[40]。但是，MP2vec[37]的整体性能表现并不理想，平均 $F1$ 分数仅为 0.44。这表明仅依赖异构信息进行节点嵌入是不够的，节点之间的丰富语义关联不容忽视。因此，本研究同时考虑网络的结构信息和语义信息，以学习更稳健的跨媒体关联，从而更有效地挖掘网络视频事件。

MHGCN[54]设计了一个图卷积模块来自动地提取元路径信息，相比于手动选择元路径的模型Deepwalk[40]和 MP2vec[37]而言，其在精确率、召回率和 $F1$ 上均有所提升，因为其规避了手动选择元路径导致的信息遗漏的问题。但是，该模型针对元路径上的不同节点类型进行无差别地聚合，导致了节点特征嵌入的偏差。因此，本文提出将异构网络解耦为关系异构子图和基于元路径的同构子图，分别对两种类型的子图嵌入表示，将异构图中不同节点类型之间的差异进行区别对待，使得目标节点学习到不同关系节点邻居对其的贡献。

表 5 *F1* 对比

话题	LDA	VSDA	MMBT	CoNe	Deepwalk	MP2vec	MHGCN	muxGNN	HeGAE-AC	本文
1	0.20	0.22	0.42	0.44	0.08	0.18	0.45	0.45	0.47	**0.75**
2	0.22	0.34	0.51	0.53	0.42	0.48	0.35	0.53	0.25	**0.62**
3	0.08	0.24	0.62	0.58	0.54	0.38	0.34	0.60	0.59	**0.96**
4	0.32	0.23	0.65	0.62	0.40	0.44	0.42	0.63	0.50	**0.75**
5	0.21	0.33	0.39	0.32	0.38	0.55	0.58	0.67	0.45	**0.86**
6	0.19	0.18	0.44	0.54	0.43	0.42	0.37	0.52	0.65	**0.81**
7	0.22	0.42	0.65	0.52	0.44	0.47	0.37	0.57	0.75	**0.78**
8	0.14	0.14	0.45	0.50	0.57	0.48	0.43	0.62	0.64	**0.75**
9	0.22	0.31	0.50	0.32	0.52	0.49	0.37	0.54	0.63	**0.80**
10	0.07	0.41	0.54	0.34	0.66	0.56	0.50	0.71	0.73	**0.74**
平均	0.19	0.28	0.52	0.47	0.44	0.45	0.42	0.58	0.57	**0.78**

muxGNN[55]将图注意力应用在异构信息网络嵌入中。与Deepwalk[40]和MP2vec[37]模型相比，其平均精确率、召回率和 $F1$ 都有所提升。但是，muxGNN[55]直接在异构图上执行多层图卷积来学习高阶语义信息，节点的高阶信息通过堆叠网络层数来捕获时，这会导致语义混淆的问题。因此，本文提出通过对异构信息网络诱导为不同的同构子图和异构子图，分别对同构子图和异构子图信息进行聚合来学习节点的高阶信息和局部信息，从同质和异质邻居中全面捕获结构、语义和属性信息。

HeGAE-AC[56]通过设计异构图自编码器将图的拓扑结构和高阶节点的属性信息进行聚合。与Deepwalk[40]和 MP2vec[37]模型相比，其平均精确率、召回率和 $F1$ 值都有所提升。该方法虽然聚合了节点的高阶信息，但是节点的邻域信息被忽略，这会导致邻域信息丢失。因此，本文提出将异构图分解为同构和异构子图，以获取节点基于元路径的同构邻居和一阶异构邻居，利用子图捕获路径以外节点局部关联。

融合文本与视觉信息能够从不同角度呈现社会事件，提供更全面且互补的信息，从而为网络视频事件挖掘带来便利。然而，现有的方法如 DeepWalk[40]、MP2vec[37]、MHGCN[54]、muxGNN[55]和HeGAE-AC[56]等为代表的深度学习模型均利用了异构网络关联网络视频中文本和视觉信息，在网络

视频事件挖掘中也取得了一定的效果,但是受异构网络嵌入方式的影响,挖掘效果仍存不佳。这一现象在于 Deepwalk[40] 作为一种同构嵌入模型,随机游走过程路径中间异构邻域节点被丢弃,忽略了中间节点丰富的结构和语义信息,导致了信息损失,节点嵌入不完整。MP2vec[37] 和 MHGCN[54] 弥补了这一缺陷,这两种异构邻域信息融合表示学习方法考虑了路径中间节点,但它们只是简单地对包含不同类型节点的路径以相同方式聚合,忽略了路径中邻域节点的差异性,导致信息混淆,模型挖掘效果不佳。muxGNN[55] 和 HeGAE-AC[56] 嵌入方法规避了元路径的缺陷,但是对节点高阶信息的聚合都是堆叠网络层数实现,这不仅导致信息混淆还使得节点丢失了局部结构中的邻域信息,使得异构网络中跨模态间节点关联依然稀疏,网络视频事件挖掘效果不佳。

对于本文方法,我们将网络嵌入方法用于网络视频事件挖掘,将异构图解耦为不同的多个子图,并对子图结构进行基于注意力的局部聚合和全局嵌入。通过子结构来捕获原始异构图中节点的高阶信息和局部信息,并学习到融合语义信息和结构信息的节点向量表示,最终实现网络视频事件挖掘。实验结果表明,本文方法表现优异,取得了 0.94 的最高召回率和 0.96 的 $F1$ 值,同时平均精度显著提高,且在各话题上均展现出稳定的性能。首先,所提模型不仅融合了网络结构信息,还将节点的语义与属性特征整合到嵌入中,从而实现了节点嵌入与事件类别结构的一致性。其次,该方法通过同构子图学习高阶邻域节点,有效避免了语义混淆;同时借助异构子图捕捉一阶邻域关联,进一步防止了信息遗漏。本文对同构子图和异构子图基于注意力的聚合嵌入了不同子图间节点交互关系和语义信息,利用邻域信息找到节点与丢失节点间的关联,丰富稀疏的异构网络,实现最优节点嵌入,最终在下游网络视频事件挖掘中效果最佳。

4.6　消融实验

为进一步验证模型各组成部分的有效性,本节设计并报告了一系列消融实验,并对实验结果进行了详细分析。

(1)异构节点内容编码分析

在嵌入节点特征的过程中,本文首先对节点的初始特征进行了“深度”编码,以增强跨媒体信息的互补性。为验证模型中此部分设计的有效性,我们在消融实验中直接使用异构节点的初始内容特征作为模型的输入特征。实验结果如表 6 所示。通过分析所有数据集的平均结果可以发现,引入节点内容编码后,模型性能在平均精确率、平均召回率及平均 $F1$ 上均有显著提升,验证了其在捕获异构信息间关联的优势,不同节点类型之间的异构内容的“深度”交互能够有效提升模型的表达能力。

表 6　节点内容编码对模型性能的消融实验结果

组件成分	平均精确率	平均召回率	平均 $F1$
无节点内容编码	0.79	0.66	0.72
有节点内容编码	0.87	0.74	0.78

具体而言,视频中非线性视觉信息和失范的文本使得节点间关联较为稀疏。然而,视频和视觉近似关键帧集合($NDKs$)中视觉信息的客观描述可以弥补文本信息的稀疏性,从而增强文本与视觉语义之间的关联强度;同时,文本信息将非线性视觉信息串联起来,起到连接作用。因此,对节点初始内容特征进行“深度”编码不仅提升了节点间信息的关联强度,还实现了跨媒体信息的有效互补。

(2)异构子图嵌入方式分析

为了评估不同邻域节点嵌入的重要性,本文比较了三种聚合方式对节点嵌入效果的差异。具体而言:①异构子图均值聚合:基于均值操作的聚合方式,通过对邻居节点特征进行平均池化来生成目标节点的输出特征。②异构子图池化聚合:基于池化操作的聚合方式,通过选取池化后特征中每个维度的最大值作为目标节点的输出特征。

图 5 展示了异构子图嵌入方式对模型性能的消融实验结果。实验表明,在所有数据集上,无论是平均精确率、平均召回率还是平均 $F1$ 值,基于注意力的聚合方式均优于平均聚合和最大池化聚合。这一结果验证了本文提出的基于注意力的聚合方式在处理异构子图信息时的显著优势。每个异构子图中,目标节点与其邻居节点在语义信息和重要性方面存在较大差异。例如,在包含文本信息的子图中,文本邻居节点包含核心关键词和非核心关键词,他们对目标节点的影响并不相同。传统的平均聚合或最大池化聚合方式无法有效区分这些节点的重要性,导致关键信息的权重被稀释或忽略。相比之下,基于注意力的聚合方式通过为每个邻居节点分配不同的权重,能够根据节点之间的语义和结构关系动态地捕捉其重要性。这种自适应的权重分配机制使得模型能够更加精准地聚合关键信息,从而提升节点嵌入的表示能力和模型的整体性能。

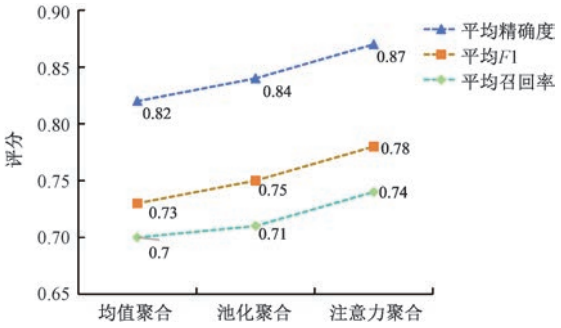

图 5 异构子图嵌入方式对模型性能的消融实验结果

(3)嵌入方式分析

为充分挖掘异构网络中的丰富关联信息,本文将异构图解耦为同构子图和异构子图,分别用于捕获高阶邻域信息与一阶邻域信息,以生成节点最终嵌入。为了验证同时嵌入同构子图和异构子图的有效性及其对模型性能的贡献差异,本文设计了以下消融实验:①仅嵌入同构子图:移除异构子图,仅利用同构子图学习高阶邻域信息。②仅嵌入异构子图:移除同构子图,仅利用异构子图学习一阶邻域信息。

实验结果如图 6 所示,模型在同时嵌入同构子图和异构子图时,平均精确率、平均召回率和平均 $F1$ 值均显著提升。这表明,同构子图能够有效捕获高阶邻域信息,而异构子图在提取节点直接语义关联的有效性也得到了进一步验证。此外,从图中可以看出,仅嵌入异构子图的模型在多个性能指标上表现上更优,对于节点嵌入性能的贡献更大。其原因在于,异构子图能够直接捕获节点与邻居之间的语义关联,而同构子图只能学习间接的高阶关系,丢失了节点一阶邻域信息。

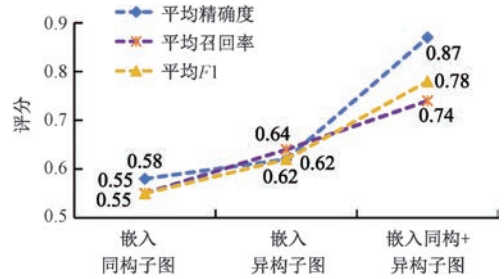

图 6 不同嵌入方式对模型性能的消融实验结果

5 总 结

针对跨媒体关联学习中文本在视觉近似关键帧(NDKs)的稀疏性问题,本文提出了一个新的基于子图邻域学习的关联增强方法。首先,本文将节点邻域推广到基于元路径引导的同构邻域和基于关系的异构邻域。然后,在此基础上,将异构图分解为节点同构邻域子图和异构邻域子图,用于捕获和聚合节点邻域交互信息。最后,本文设计了子图内特定注意力机制和子图间图级注意力机制对子图邻域进行聚合。经由以上三个步骤,可以获取节点的最终嵌入,实现网络视频突发事件的准确挖掘。实验结果表明,本文提出的方法能有效提高网络视频热点话题检测效果。

致谢 本课题得到国家社会科学基金一般项目(22BXW081)资助。

参 考 文 献

[1] Newman, N. Journalism, media and technology trends and predictions 2018. Oxford: Reuters Institute for the Study of Journalism, 2017

[2] Li H, Ke Q, Ong G M, et al. Video joint modelling based on hierarchical transformer for Co-Summarization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3): 3904-3917

[3] Qin D, Tang X, Huang Y, et al. Subgraph autoencoder with bridge nodes. Expert Systems with Applications, 2024, 257:125069

[4] Zhang C, Song D, Huang C, et al. Heterogeneous graph neural network//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage, USA, 2019: 793-803

[5] Velickovic P, CuaYull G, Casanava A, et al. Graph attention networks//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-12

[6] Chen K J, Lu H, Liu Z, et al. Heterogeneous graph convolutional network with local influence. Knowledge-Based Systems, 2022, 236: 107699

[7] Lai P Y, Dai Q Y, Lu Y H, et al. MIGP: Metapath integrated graph prompt neural network. Neural Networks, 2024, 179: 106595

[8] Li X, Ding D, Kao B, et al. Leveraging meta-path contexts for classification in heterogeneous information networks//Proceedings of the 2021 IEEE 37th International Conference on Data Engineering (ICDE). Chalkida, Greece, 2021: 912-923

[9] Liu H, Chen Z, Tang J, et al. Mapping the technology evolution path: a novel model for dynamic topic detection and tracking. Scientometrics, 2020, 125: 2043-2090

[10] Zhang C, Liu D, Wu X, et al. Near-duplicate segments based news web video event mining. Signal Processing, 2016, 120: 26-35

[11] Hofmann T, et al. Probabilistic latent semantic analysis//Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI' 99). Stockholm, Sweden, 1999: 289-296

[12] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3 (Jan):

993-1022

[13] Wang J, Wang Z, Zhang D, et al. Combining knowledge with deep convolutional neural networks for short text classification//Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17). Melbourne, Australia, 2017: 2915-2921

[14] WU M. Commonsense knowledge powered heterogeneous graph attention networks for semi-supervised short text classification. Expert Systems with Applications, 2023, 232: 120800

[15] Vo D T, Ock C Y. Learning to classify short text from scientific documents using topic models with various types of knowledge. Expert Systems with Applications, 2015, 42(3): 1684-1698

[16] Chen J, Hu Y, Liu J, et al. Deep short text classification with knowledge powered attention//Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI'19). Honolulu, USA, 2019: 6252-6259

[17] Jian M, Wang J, Yu H, et al. Integrating object proposal with attention networks for video saliency detection. Information Sciences, 2021, 576: 819-830

[18] Ke Y, Sukthankar R, Huston L, et al. Efficient near-duplicate detection and sub-image retrieval//Proceedings of the ACM Multimedia 2004. New York, USA, 2004:5

[19] Yao J, Cui B, Huang Y, et al. Bursty event detection from collaborative tags. World Wide Web, 2012, 15: 171-195

[20] Zheng M, You S, Huang L, et al. Xu, CoNe: Contrast your neighbours for supervised image classification. arXiv preprint arXiv:2308.10761,2023.

[21] Zhang C, Lei Y, Xiao X, et al. Cross-media video event mining based on attention graph structure learning. Neurocomputing, 2022, 502: 148-158

[22] Zhang Cheng-De, Liu Yu-Xuan, Xiao Xia, et al. Hot topic detection of web video based on cross-media semantic association enhancement. Journal of Computer Research and Development, 2023, 60(11): 2624-2637 (in Chinese)
(张承德,刘雨宣,肖霞等.跨媒体语义关联增强的网络视频热点话题检测.计算机研究与发展, 2023, 60(11):2624-2637)

[23] Zhuo Yun-kan, Qi Jin-wei, Peng Yu-xin. Cross-media deep fine-grained correlation learning. Journal of Software, 2019, 30(4): 884-895 (in Chinese)
(卓昀侃,綦金玮,彭宇新.跨媒体深层细粒度关联学习方法.软件学报,2019, 30(4):884-895)

[24] Kiela D, Bhooshan S, Firooz H, et al. Supervised multimodal bitransformers for classifying images and text. arXiv preprint arXiv:1909.02950,2019.

[25] Chen T, Liu C, Huang Q. An effective multi-clue fusion approach for web video topic detection//Proceedings of the 20th ACM International Conference on Multimedia. Nara, Japan, 2012: 781-784

[26] Pu Zhan-Xing, Ge Yong-Xin. Few-shot action recognitionin video basedon multi-feature fusion. Chinese Journal of Computers,2023,46(3):594-608 (in Chinese)
(蒲瞻星,葛永新.基于多特征融合的小样本视频行为识别算法.计算机学报, 2023,46(3):594-608)

[27] Xiao X, Du M, Xu S, et al. Cross-media web video event mining based on multiple semantic-paths embedding. Neural Computing and Applications, 2024, 36(2): 667-683

[28] Xiao Guo-qing, Li Xue-qi, Chen Yue-dan, et al. A survey of large-scale graph neural networks. Chinese Journal of Computers, 2024, 47(1): 148-171(in Chinese)
(肖国庆,李雪琪,陈玥丹等.大规模图神经网络研究综述.计算机学报, 2024, 47(1):148-171)

[29] Lou Zheng-Zheng, Zhu Jun-Jiao, Zhang Wan-Chuang, et al. Role-guided neural recommendation in user-generated content scenarios. Chinese Journal of Computers, 2024,47(1).1288-1303 (in Chinese)
(娄铮铮,朱军娇,张万闯等.用户生成内容场景下角色导向图神经推荐方法.计算机学报, 2024,47(1).1288-1303)

[30] Bruna J, Zaremba W, Szlam A, et al. Spectral networks and locally connected networks on graphs. arXiv preprint arXiv: 1312.6203,2013.

[31] Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering//Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 3844-3852

[32] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016.

[33] Mo X, Wan B, Tang R. TemporalHAN: Hierarchical attention-based heterogeneous temporal network embedding. Engineering Applications of Artificial Intelligence, 2024, 133: 108376

[34] Atwood J, Towsley D. Diffusion-convolutional neural networks//Advances in Neural Information Processing Systems (NIPS). Barcelona, Spain, 2016: 1993-2001

[35] Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs//Advances in Neural Information Processing Systems (NIPS). Long Beach, USA, 2017: 1025-1035

[36] Fu C, Yu P, Yu Y, et al. MHGCN+: Multiplex heterogeneous graph convolutional network. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3): 1-25

[37] Dong Y, Chawla N V, Swami A. metapath2vec: Scalable representation learning for heterogeneous networks//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada, 2017: 135-144

[38] Mikolov T. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781,2013.

[39] Shi C, Hu B, Zhao W X, et al. Heterogeneous information network embedding for recommendation. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(2): 357-370

[40] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2014: 701-710

[41] Wang X, Ji H, Shi C, et al. Heterogeneous graph attention network//Proceedings of the World Wide Web Conference (WWW'19). San Francisco, USA, 2019: 2022-2032

[42] Mei G, Pan L, Liu S. Heterogeneous graph embedding by aggregating meta-path and meta-structure through attention mechanism. Neurocomputing, 2022, 468: 276-285

[43] Fu X, Zhang J, Meng Z, et al. MAGNN: Metapath aggregated graph neural network for heterogeneous graph embedding//Proceedings of the Web Conference 2020 (WWW'20). Taipei, China, 2020: 2331-2341

[44] Yun S, Jeong M, Kim R, et al. Graph transformer networks//Advances in Neural Information Processing Systems (NeurIPS). Vancouver, Canada, 2019: 1322-1332
[45] Chang Y, Chen C, Hu W, et al. Megnn: Meta-path extracted graph neural network for heterogeneous graph representation learning. Knowledge-Based Systems, 2022, 235: 107611
[46] Dong X, Zhang Y, Pang K, et al. Heterogeneous graph neural networks with denoising for graph embeddings. Knowledge-Based Systems, 2022, 238: 107899
[47] Chen H, Yin H, Wang W, et al. PME: projected metric embedding on heterogeneous networks for link prediction//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK, 2018: 1177-1186
[48] Yang Y, Tian Y, Huang T. Multiscale video sequence matching for near-duplicate detection and retrieval. Multimedia Tools and Applications, 2019, 78(1): 311-336
[49] Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding. arXivpreprint arXiv: 1810.04805, 2018
[50] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA, 2015: 3431-3440
[51] Arnab A, Dehghani M, Heigold G, et al. ViViT: A video vision transformer//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Virtual, 2021: 6836-6846
[52] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991, 2015
[53] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Advances in Neural Information Processing Systems (NeurIPS). Long Beach, USA, 2017: 5998-6008
[54] Yu P, Fu C, Yu Y, et al. Multiplex heterogeneous graph convolutional network//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD). Virtual, 2022: 2377-2387
[55] Melton J, Krishnan S. muxGNN: Multiplex graph neural network for heterogeneous graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 11067-11078
[56] Chen Y, Liu Y. HeGAE-AC: Heterogeneous graph auto-encoder for attribute completion. Knowledge-Based Systems, 2024, 287: 111436
[57] Kingma D P, Ba J L. Adam: A method for stochastic optimization//Proceedings of the 3rd International Conference on Learning Representations (ICLR). San Diego, USA, 2015: 1

ZHANG Cheng-De, Ph. D., associate professor. His main research interests include multimedia information retrieval and data analysis.

ZHOU Xuan, Master candidate. Her main research interests include multimedia information retrievaland data mining.

Background

Recently, event mining based on cross-media association learning in web videos has garnered significant attention. Existing methods primarily rely on calculating text or video similarity to identify related videos within a video repository. However, the non-linear visual content and sparse textual annotations in web videos make it challenging to establish semantic connections among videos depicting the same event. Heterogeneous information networks (HINs) have become a popular tool for representing cross-media data in web videos due to their strong ability to represent complex relationships. Nevertheless, the fragmented visual content and sparse textual annotations inevitably result in the sparse distribution of text across visually approximate keyframes (NDKs), leading to incomplete heterogeneous networks with missing associations. This poses significant challenges for effective node embedding. Existing methods enhance cross-media associations by embedding multiple semantic paths, but they often overlook node correlations within local subgraph structures. This results in node embeddings that fail to capture neighborhood information, ultimately degrading the performance of event mining.

To address these challenges, we propose a method for cross-media semantic association enhancement based on subgraph neighborhood learning. This approach decomposes the heterogeneous graph into various types of subgraphs to capture neighborhood node associations and generate the final node embeddings. First, multimodal node features are mapped into a unified feature space. The heterogeneous graph is decomposed into homogeneous and heterogeneous subgraphs to extract homogeneous neighbors based on meta-paths and first-order heterogeneous neighbors. Then, neighborhood node information within the subgraphs is embedded using a specific attention mechanism. Finally, graph-level attention is used to aggregate interactive semantics between homogeneous and heterogeneous subgraphs. This produces node embeddings enriched with neighborhood associations, enabling accurate mining of web video emergencies. Through experiments on 10 real-world datasets, the proposed method is shown to surpass existing models, providing strong evidence for its reliability.

The research presented in this paper was supported by the General Program of the National Social Science Foundation of China (No. 22BXW081).