



# 多媒体智能计算若干研究进展

朱文武\*

清华大学计算机科学与技术系, 北京信息科学与技术国家研究中心, 北京 100084

\* 通信作者. E-mail: wwzhu@tsinghua.edu.cn

收稿日期: 2025-02-27; 修回日期: 2025-06-22; 接受日期: 2025-09-08; 网络出版日期: 2025-09-12

科技创新 2030 “新一代人工智能”重大项目 (批准号: 2020AAA0106300)、国家自然科学基金原创探索计划项目 (批准号: 62250008) 和北京信息科学与技术国家研究中心项目 (批准号: BNR2023TD03006) 资助

**摘要** 多媒体智能计算是融合人工智能与多媒体计算的交叉学科, 核心目标是通过大规模多媒体数据的智能表征、分析与协同计算, 实现多模态数据的深度理解、高效服务. 多媒体智能计算面临两个重大挑战: (1) 多媒体数据实体规模随时间指数增长, 关联组合指数爆炸, 因此数据实体关联规模面临呈双重指数增长态势的难题; (2) 多媒体数据量指数增长, 网络结构异构复杂, 用户需求动态并发, 导致用户访问媒体所需网络资源随时间指数增长. 针对多媒体大数据实体关联规模双重指数增长导致的关联与规律难以发现的难题, 本文介绍多媒体大数据关联表征理论与方法的新进展, 通过揭示图拓扑空间和向量表征空间的映射规律, 发现了图非对称传递性向量化度量的统一数学表达, 提出了结构保持的深度关联表征方法, 建立了微观局部结构与宏观全局结构融合的图关联表征学习模型, 实现了图关联表征学习从浅层到深层、从微观结构到宏观结构的跨越. 针对异构网络环境下大规模多媒体内容指数增长与网络资源有限的矛盾难题, 本文探究用大数据分析理论解决该难题的多媒体网络资源计算新范式, 揭示了媒体内容、用户行为和网络拓扑的相互作用机理, 提出了基于用户行为的媒体内容传播预测与大规模网络拓扑模型, 建立了大规模多媒体的资源跨域协同计算新理论, 实现了网络资源配置从被动固化到主动感知的跨越. 研究成果推动多媒体智能从语义驱动到关联驱动、从孤立优化到全局协同的跨越, 对公共安全、智慧城市、互联网经济等场景具有重要应用价值.

**关键词** 多媒体大数据, 跨媒体智能, 多媒体资源计算, 大数据分析

## 1 引言

随着人工智能的迅猛发展, 大规模多媒体数据呈指数增长态势, 多媒体智能计算作为融合人工智能与多媒体技术的交叉学科, 已成为支撑数字内容服务、智能媒体分析、实时流媒体传输等国家重大应用的核心技术领域. 其核心目标是通过大规模多媒体数据的智能表征、分析与协同计算, 实现多模态数据的深度理解、高效传输与动态服务<sup>[1~4]</sup>. 多媒体智能计算综合了人工智能与多媒体计算的相关核心技术. 从内涵来看, 多媒体智能计算以跨模态数据表征学习为基础, 结合网络资源动态优化机制, 构建从数据感知到服务决策的全链路智能框架; 其外延则涵盖多媒体内容生成与理解、跨模态检

**引用格式:** 朱文武. 多媒体智能计算若干研究进展. 中国科学: 信息科学, 2025, 55: 2153–2164, doi: 10.1360/SSI-2025-0091

Zhu W W. Research progress on multimedia intelligence computing. Sci Sin Inform, 2025, 55: 2153–2164, doi: 10.1360/SSI-2025-0091

索与推荐、分布式媒体服务等多个应用场景,并与计算机视觉、自然语言处理、边缘计算等学科深度交叉,形成多维度的技术生态。

当前,多媒体智能计算面临两大核心挑战:其一,多媒体数据的规模与复杂性呈双重指数增长。随着多模态传感器技术的普及,图像、视频、音频、文本等异构数据在实体数量和关联关系上均呈现爆炸式增长,传统基于单模态或浅层关联的表征方法难以刻画跨媒体数据的复杂语义与拓扑耦合关系;其二,网络资源供需矛盾日益尖锐。在用户需求动态化、服务场景多样化的驱动下,海量多媒体内容的高效分发与低延迟传输需求与异构网络环境中资源受限、拓扑动态的特性形成根本性冲突。

本文指出上述两大挑战分别对应多媒体智能计算的两个基础性问题——跨媒体表征学习与多媒体网络资源计算。前者旨在解决数据内在关联规律难以挖掘的问题,后者则致力于突破资源分配与服务质量间的动态平衡难题,二者共同构成了多媒体智能计算理论体系的核心支柱。进一步,本文提出通过跨媒体大数据表征学习与大规模多媒体网络资源计算来解决多媒体数据规模复杂性双指数增长以及多媒体网络资源供需矛盾尖锐这两大挑战。

**跨媒体大数据表征学习是多媒体智能的基石。**跨媒体大数据表征学习是从大规模多元异构数据中发现并利用深层次、多尺度、全周期关联规律的核心,其本质是一个由海量异构实体与复杂关联结构相互耦合所形成的科学难题。传统理论方法一直受限于实体语义与关联结构相独立的研究范式,通过文本、图像等固有表征空间构建关联结构,忽略了实体语义与关联结构的相互作用及耦合效应,形成的不完备表征空间难以刻画跨空间异构实体的复杂关联关系与本质关联规律,导致跨媒体大数据关联不能识别、规律无法发现、态势难以预测的困境。如何突破传统实体表征对关联结构的表达局限,揭示跨媒体/跨空间复杂关联关系的驱动要素和形成机理,构造实体语义与关联结构相互耦合的跨媒体/跨空间统一关联表征,是长期困扰学术界的一个基本难题。本文以大规模异构网络作为跨媒体/跨空间异构实体和复杂关联的科学抽象,揭示了网络拓扑空间与向量表征空间的映射机理,建立了基于大规模异构网络的跨媒体/跨空间大数据关联统一表征,构建了外在拓扑结构保持、内在演化性质对等、总体语义关联耦合的大规模网络表征学习理论体系。

**大规模多媒体网络资源计算则是实现智能服务落地的关键保障。**大规模多媒体的网络资源计算是在异构网络环境下满足大规模用户对海量多媒体内容高动态、高并发需求的核心,本质是一个由内容、网络 and 用户 3 个域相互耦合所形成的复杂系统。传统理论与方法受限于网络域独立优化的假设,无法解决大规模用户对海量多媒体内容需求指数增长与网络资源有限的根本矛盾。为了突破网络域独立优化的局限,跨越网络域与用户域和媒体域之间的鸿沟,本文从大规模网络多媒体的微观机制和宏观规律入手,揭示了媒体内容、用户行为和网络拓扑的内在特性与相互作用机理,建立了微观动态适配、宏观跨域协同的网络资源计算理论模型。

本文观点基于作者近 20 年来在多媒体智能计算方向的研究经历与科研成果。第 2 节介绍跨媒体大数据表征学习理论方法,并以若干篇表征学习相关的代表性工作进行支撑。第 3 节介绍大规模多媒体网络资源计算理论方法,并以若干篇网络资源配置相关论文作为支撑。第 4 节展望未来研究方向,对大模型时代下的新一代多媒体智能计算发展路径进行讨论,涵盖生成式人工智能、具身智能等内容。总体上,希望对推动多媒体智能计算从语义驱动到关联驱动、从孤立优化到全局协同的跨越式发展提供一些观点和思考。

## 2 跨媒体大数据表征学习理论方法

跨媒体大数据表征学习是多媒体智能计算的基石,是从海量异构数据中挖掘深层次、多尺度关联规律的核心技术。随着图像、视频、音频、文本等不同模态数据的涌现,传统的数据表征方法已经无法有效应对这些异构数据之间复杂的内在联系。传统方法大多将数据实体和关联结构分开处理<sup>[5]</sup>,这种处理方式往往忽略了实体语义和关联结构之间的相互作用和耦合效应,从而导致所构建的表征空间无

表 1 四种经典图拓扑度量的统一数学表达形式.

**Table 1** The unified mathematical expressions of four classic graph topological metrics.  $A$  represents the adjacency matrix,  $I$  represents the identity matrix, and  $D$  represents the degree matrix.

|                       | $M_g$               | $M_l$                  |
|-----------------------|---------------------|------------------------|
| Katz                  | $I - \beta \cdot A$ | $\beta \cdot A$        |
| Personalized pagerank | $I - \alpha P$      | $(1 - \alpha) \cdot I$ |
| Common neighbors      | $I$                 | $A^2$                  |
| Adamic-adar           | $I$                 | $A \cdot D \cdot A$    |

法准确反映跨模态数据中的复杂语义关系和内在规律. 迫切需要突破传统方法的局限, 构建一个统一的跨媒体关联表征.

为了实现这一目标, 本文提出基于图结构建模的跨媒体关联表征方法. 图作为实体及其关系的数学抽象, 能够更加自然地表达不同模态数据之间的关联关系. 通过构建图的拓扑结构和向量表征空间之间的映射关系, 我们不仅能够有效刻画跨模态数据的复杂关联, 还能够在大规模数据环境中保持计算的高效性, 实现了跨媒体大数据表征从语义驱动到关联驱动的跨越. 本文提出的跨媒体大数据表征学习理论方法包含三部分: (1) 拓扑和向量表征空间的映射机理; (2) 拓扑在向量表征空间的嵌入机制; (3) 语义与拓扑的向量表征融合关系建模.

## 2.1 拓扑和向量表征空间的映射机理

传统的图拓扑分析方法通常只关注图的结构特性, 而忽视了拓扑结构与向量表征空间之间的关系. 图的拓扑结构通常具有多层次、多尺度的复杂性, 而传统的表征方法多采用浅层的线性模型进行处理, 这导致了跨媒体大数据图关联规律难以被准确刻画. 因此, 探索图的拓扑结构和向量表征空间之间的嵌入机理, 成为了解决这一问题的关键.

通过对多种图拓扑的研究, 我们发现, 图的传递性是图关联表征的重要驱动因素. 尤其是拓扑的非对称传递性, 在向量空间中往往难以得到有效保留. 因此, 我们提出了图的非对称传递性与向量空间双向度量的映射机制. 通过这一机制, 我们能够将经典的图度量 (如 Katz, PageRank, Common Neighbors 等) 有效地嵌入到低维向量空间中, 从而克服了传统方法在处理有向图时的局限性. 这一方法不仅能够保持图的非对称传递性, 还能够实现高效的关联推理, 为跨媒体数据的智能处理提供了坚实的理论基础.

具体来说, 文献 [6] 揭示了 4 种度量的统一数学表达  $S = M_g^{-1} M_l$  (如表 1 所示). 基于广义特征值分解理论探明了该数学表达在向量表征空间的嵌入规律:

$$\min \|S - U^s \cdot U^{t^\top}\|_F^2, \quad (1)$$

其中  $U^s$  代表有向边中的源节点,  $U^t$  代表有向边中的目标节点. 该规律可广泛适用于经典图拓扑度量在低秩表征空间中的嵌入和保持.

为了进一步提高计算效率, 我们基于特征值分解理论, 提出了一种高效的近似计算方法. 这一方法通过 Jacobi-Davidson 迭代算法, 能够将计算复杂度从传统的立方级降低到线性级, 从而大大提高了图关联表征的计算效率. 该方法在保持计算精度的同时, 有效减少了计算资源的消耗, 特别适用于大规模图数据的处理.

## 2.2 拓扑在向量表征空间的嵌入机制

图拓扑结构的复杂性要求我们采用深度学习方法来进行高阶结构的学习. 传统的浅层模型通常只关注图的低阶相近性, 而忽视了高阶相近性的影响. 然而, 真实的图数据通常表现出高阶相近性比低阶相近性更加密集的统计特性, 这使得高阶相近性在图结构的学习中扮演着重要的角色.

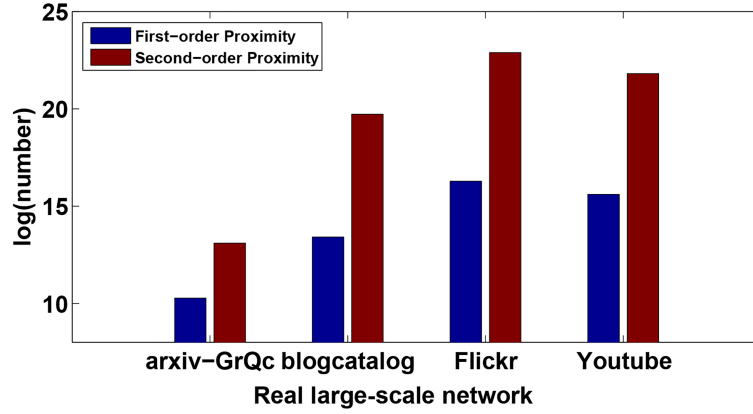


图 1 (网络版彩图) 真实图相近性统计规律.

Figure 1 (Color online) Statistical law of similarity between real graphs.

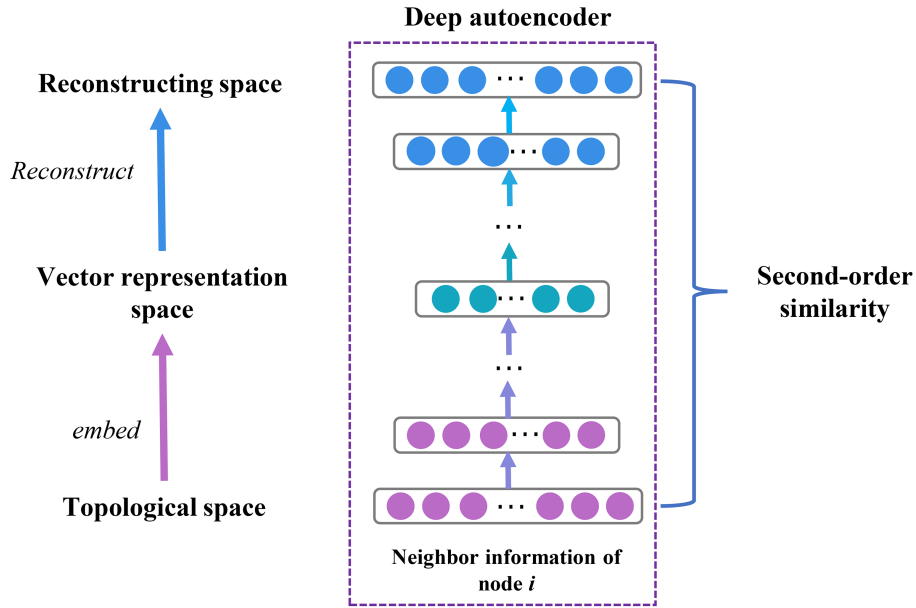


图 2 (网络版彩图) 重构优化范式对高阶相近性表达能力.

Figure 2 (Color online) The ability of the reconstruction optimization paradigm to express high-order similarity.

文献 [7] 通过对多个真实图拓扑结构进行定量分析, 发现真实图中节点高阶相近性稠密度远大于低阶相近性的统计规律 (如图 1 所示), 且高阶相近性对图结构的关联推理具有决定性影响. 因此, 我们进一步提出了一种基于深度非线性自编码器的优化范式 (如图 2 所示), 通过重构优化算法有效表达图的高阶相近性, 实现了复杂图结构非线性深层表征. 与传统方法相比, 深度非线性模型能够更好地捕捉图中复杂的非线性关系, 为图的高阶结构提供了更为精确的表征.

此外, 我们还进一步研究社区结构在图关联表征中的作用. 社区结构不仅是图拓扑的宏观特征, 而且对图嵌入学习具有重要的约束效应. 通过分析社区结构的模块度指标, 文献 [8] 揭示了其与图嵌入空间的内在关系, 提出了微观结构与宏观结构共同作用的图关联表征学习机制:

$$\min_{M, U, H, C} \|S - MU^T\|_F^2 + \alpha \|H - UC^T\|_F^2 - \beta \text{tr}(H^T B H), \quad (2)$$

其中  $S$  表示微观结构相似度矩阵,  $H$  表示节点的社区分布矩阵,  $C$  表示社区表征矩阵,  $U$  表示节点表征矩阵,  $B$  表示社区结构模块度矩阵. 这一机制不仅能够准确地捕捉图的局部和全局结构, 还能够

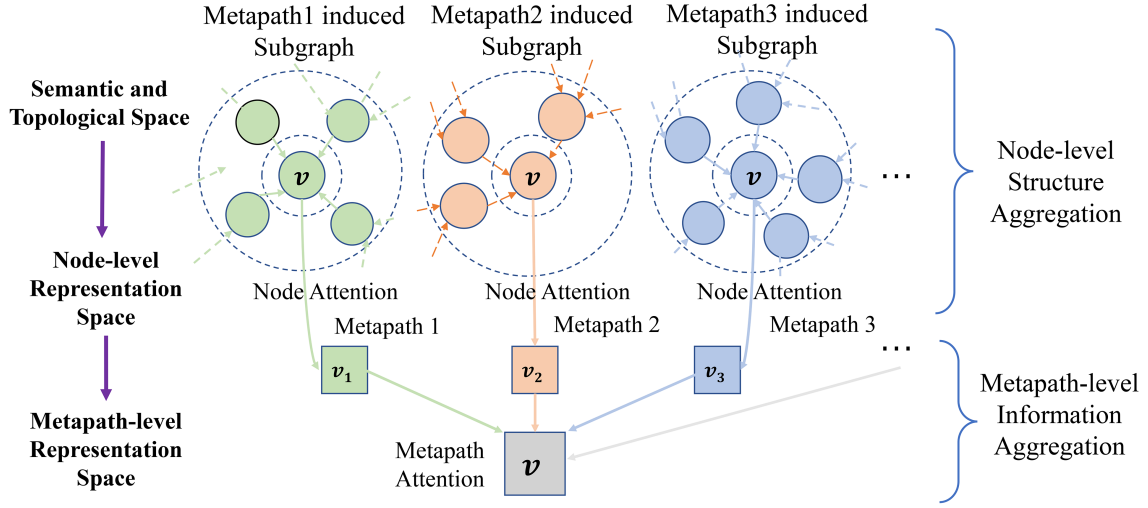


图 3 (网络版彩图) 节点级表征与元路径级表征层次关系.

Figure 3 (Color online) Hierarchical relationship between node-level representation and metapath-level representation.

多模态数据中提供更加精确的关联推理.

### 2.3 语义与拓扑的向量表征融合关系建模

传统的图关联表征方法通常只处理同构图的情况, 忽视了异构实体语义与拓扑结构之间的关系. 随着跨模态数据的广泛应用, 如何将异构数据的语义与拓扑结构有效融合, 成为了一个亟待解决的挑战. 为了解决这一问题, 文献 [9] 发现了节点级表征与元路径级表征的层次关系特性 (如图 3 所示), 进一步提出了一种基于元路径的表征融合方法:

$$\alpha_{ij}^{\phi} = \text{softmax}(e_{ij}^{\phi}) = \frac{\exp(\sigma(a_{\phi}^{\top} \cdot [h'_i || h'_j]))}{\sum_{k \in \mathcal{N}_i^{\phi}} \exp(\sigma(a_{\phi}^{\top} \cdot [h'_i || h'_k]))}, w_{\phi_i} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} q^{\top} \cdot \tanh(W \cdot z_i^{\phi} + b), \quad (3)$$

其中,  $\alpha_{ij}^{\phi}$  表示节点级注意力系数,  $w_{\phi_i}$  表示元路径级注意力系数,  $h_i$  表示节点  $i$  的表征,  $W$  和  $b$  表示可学习参数. 元路径能够有效捕捉异构数据中的复杂关系, 双层局部注意力机制能够完善跨模态表征的层次结构, 而节点级表征与元路径级表征相结合, 使得模型能够在多模态数据中捕捉到更加细粒度的关联关系, 最终实现语义与拓扑的深度融合.

与此同时, 传统图关联表征中多种实体语义混淆、多尺度拓扑结构混杂, 导致表征空间缺乏可解释机制. 为解决此问题, 文献 [10] 揭示了实体语义和拓扑结构在表征空间中的聚合规律, 即具有相似语义及稠密关联结构的实体簇对应于独立的表征子空间, 表征子空间相似性是实体之间产生关联的决定性因素. 基于上述规律, 本文提出了能够捕捉图关联内在因子的动态邻域路由机制:

$$c_k^{(t)} = \frac{z_{u,k} + \sum_{v:(u,v) \in G} p_{v,k}^{(t-1)} z_{v,k}}{\|z_{u,k} + \sum_{v:(u,v) \in G} p_{v,k}^{(t-1)} z_{v,k}\|_2}, p_{v,k}^{(t)} = \frac{\exp(z_{v,k}^{\top} c_k^{(t)} / \tau)}{\sum_{k'=1}^K \exp(z_{v,k'}^{\top} c_k^{(t)} / \tau)}, \quad (4)$$

其中  $u, v$  分别代表不同实体,  $c_k^{(t)}$  是因子  $k$  对应的邻域路由中心,  $p_{v,k}^{(t)}$  是实体  $u$  和实体  $v$  通过因子  $k$  产生关联的概率,  $z_{v,k}$  是  $v$  在因子  $k$  对应表征空间的投影特征.

动态邻域路由机制使得关联表征变得更灵活和高效. 如图 4 所示, 传统方法 (图 4(a)) 表征维度之间呈现全局弥散性关联, 无法体现产生关联的特定因子, 表征可解释性差; 而文献 [10] 提出的动态邻域路由机制可将混杂的多种因子进行有效区分, 表征维度之间形成局部性独立化效应, 有效提升了表征空间的可解释性.

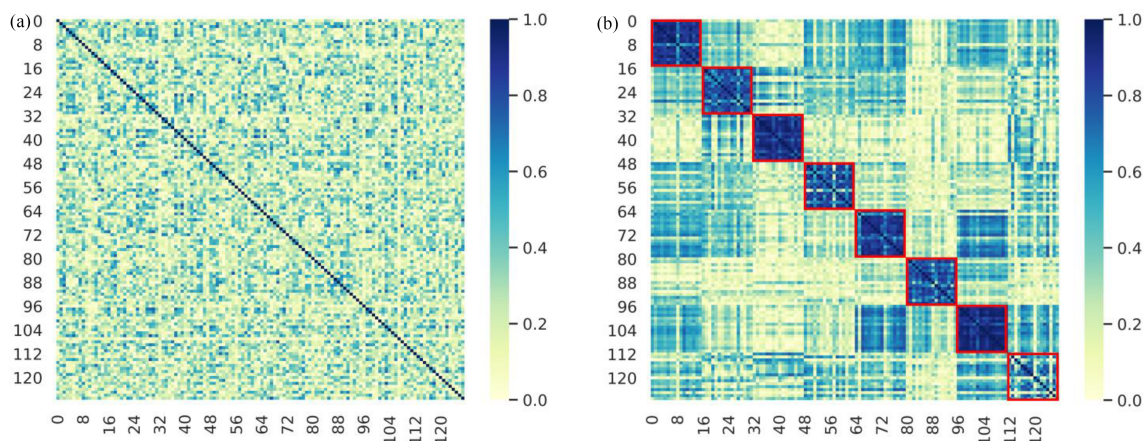


图 4 (网络版彩图) 传统图表征 (a) 与所提出独立化图表征 (b) 的表征相关度.

**Figure 4** (Color online) Representation correlation between traditional graph representation (a) and the proposed independent graph representation (b).

### 3 多媒体资源跨域协同计算理论方法

在多媒体智能计算中, 大规模多媒体网络资源计算是确保智能服务能够有效落地的关键. 大规模多媒体网络资源计算面临的主要挑战是如何在异构网络环境中, 满足海量用户对多媒体内容的高动态、高并发需求. 传统的理论方法大多假设网络域是独立优化的, 这种假设忽略了网络、用户和多媒体内容之间的复杂交互关系<sup>[11]</sup>. 随着大规模多媒体内容的增长、用户需求的动态变化, 传统的单一优化方法无法有效应对网络资源的供需矛盾. 要解决这一问题, 我们需要从内容、网络 and 用户 3 个领域的相互作用出发, 提出一种新的跨域资源协同计算理论.

本文提出了一种数据驱动的解决思路, 建立了多媒体资源跨域协同计算理论, 通过大数据分析, 揭示媒体内容、用户行为和网络拓扑之间的相互作用机制, 构建了基于微观动态适配和宏观跨域协同的资源配置理论模型, 实现了从网络域孤立优化到用户域-网络域-内容域全局协同的跨越. 这种数据驱动的思路, 不仅能有效地将用户行为、网络拓扑和媒体内容的特性相结合, 实现多媒体资源的跨域协同计算, 同时还能够提高资源利用率, 在异构网络环境下动态地调整资源配置, 以更好地满足用户需求. 本文提出的多媒体资源跨域协同计算理论方法包含三部分: (1) 用户与媒体内容交互行为量化机理; (2) 用户对网络的拓扑影响机制; (3) 媒体与网络联合率失真规律.

#### 3.1 用户与媒体内容交互行为量化机理

用户和多媒体内容之间的交互行为是影响网络资源需求的关键因素. 用户与多媒体内容的交互不仅仅依赖于内容本身的特性, 还与用户的个性化需求、行为习惯以及网络环境的变化密切相关. 传统的网络资源配置方法通常假设用户行为是与网络环境时空无关的, 这一假设显然与现实情况不符. 实际中, 用户的行为往往是有时空规律性的, 且会受到网络拓扑的影响. 因此, 如何对用户与多媒体内容之间的交互行为进行量化建模, 成为解决大规模多媒体资源优化配置的基础.

文献 [12, 13] 提出了多因子融合的用户行为量化理论. 通过对网络节点邻接关系、用户偏好以及内容语义分布等多个因素的综合分析, 揭示了复杂网络环境下语义偏好相关性和近邻节点趋同性两大主导要素对于用户与媒体信息交互行为的并发影响 (如图 5 所示), 进一步构建了用户与媒体内容交互行为的量化模型, 其量化关系如下:

$$R \sim SG^T \odot U^T V, \quad (5)$$

其中  $R$  是用户与媒体的交互行为矩阵,  $S$  是网络节点的邻接关系矩阵,  $G$  是网络节点的趋同性强度

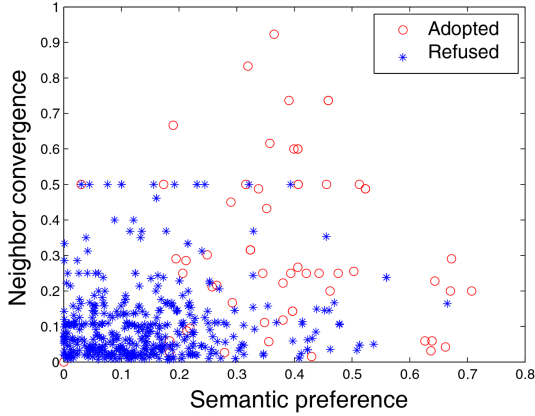


图 5 (网络版彩图) 用户 - 媒体交互行为影响因素统计分析.

Figure 5 (Color online) Statistical analysis of factors influencing user-media interaction behavior.

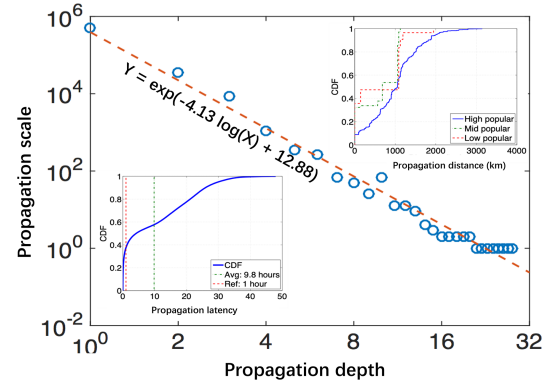


图 6 (网络版彩图) 传播局部性的统计分析.

Figure 6 (Color online) Statistical analysis of propagation locality.

矩阵,  $U$  是用户的偏好矩阵,  $V$  是媒体内容的语义分布矩阵,  $\odot$  表示 Hardamard 积, 用于描述  $SG$  和  $UV$  两个因子的融合关系. 我们提出的模型能够有效地量化用户与媒体之间的交互行为, 并揭示其背后的时空规律, 不仅能够为网络资源的优化配置提供依据, 还能够为多媒体内容的推荐和传播提供重要参考.

在实际应用中, 我们还发现了媒体内容传播的宏观统计特性. 通过对多媒体内容分发数据的统计分析, 我们发现了用户与媒体信息交互的时间局部性、空间局部性和网络拓扑局部性等规律 (如图 6 所示), 并在文献 [14] 中揭示了用户 - 媒体交互行为与全局用户、区域传播及局部用户信息的关联机理:

$$e_v^{(T)} = \frac{s_v^{(T)}}{h_v^{(T)} t^\alpha \sum_{k=1}^N \frac{1}{k^\alpha}}, \quad (6)$$

其中  $s_v^{(T)}$ ,  $h_v^{(T)}$  和  $T$  分别反映了传播的网络、空间和时间属性,  $e_v^{(T)}$  为用户访问预测. 上述规律与机理反映了用户行为和网络环境之间的深层次联系, 为多媒体内容的传播和优化提供了新的思路.

### 3.2 用户对网络的拓扑影响机制

网络拓扑是大规模多媒体内容分发的基础, 它不仅影响数据传输路径, 还决定了多媒体服务的质量. 用户行为和网络拓扑之间的关系是复杂且动态变化的, 用户的位置、需求和行为方式直接影响到网络拓扑的变化. 如何发现用户节点位置与拓扑结构之间的时空特性和动态演化机制, 是解决大规模、多动态网络服务质量提升的关键.

在传统的多媒体内容分发中, 网络拓扑往往假设是随机的, 带宽分布是均匀的, 内容请求是同质的. 然而, 这些假设忽视了现实中网络拓扑的复杂性和动态性. 为了解决这一问题, 我们基于对网络节点位置和拓扑结构的统计分析, 提出了路径度量准则, 并揭示了大规模多媒体自主网络中路径长度的高斯 (Gauss) 分布特性和路径长度与网络规模之间的类线性关系 (如图 7). 这些规律为我们理解网络拓扑的自主演化和用户行为的影响提供了新的视角.

基于以上规律, 我们在文献 [15] 中提出对等网络宏观统计 (网络规模) 和微观特征 (路径长度) 的关联理论, 其量化模型如下:

$$d < \log_M N + 3, \quad (7)$$

其中  $M$ ,  $N$  ( $N$  为覆盖网络中的主机总数目,  $M$  为常量) 定义了网络的宏观测量属性,  $d$  定义了拓扑路

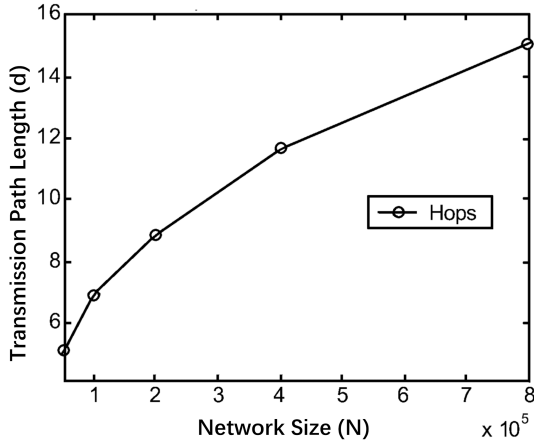


图7 路径长度的用户终端统计特性。

Figure 7 User terminal statistical characteristics of path length.

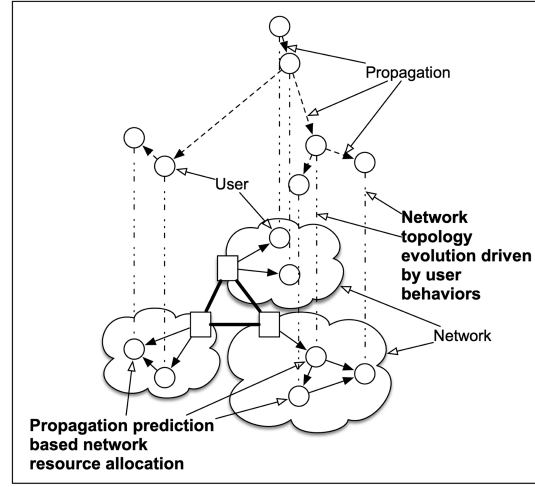


图8 行为驱动的多媒体网络资源配置。

Figure 8 Behavior-driven multimedia network resource configuration.

径特征。从而,揭示了用户行为驱动的网络拓扑自主演化规律,提出了基于内容传播预测的多媒体网络资源配置方法(如图8),建立了用户与媒体协同感知的大规模网络拓扑优化模型,通过对网络拓扑的优化,我们能够将组网计算复杂度从指数级别降低到亚线性级别  $O(\log N)$ ,从而显著提高网络资源的配置效率。该方法不仅能优化大规模网络的拓扑结构,还能提高多媒体内容的分发效率,确保用户能够在不同的网络环境中快速获取所需的内容。

### 3.3 媒体与网络联合率失真规律建模

在多媒体网络资源分配中,媒体内容的编码处理与网络资源的配置往往是相互割裂的。传统的编码方法通常忽略了网络环境和媒体内容之间的紧密联系,这导致了网络资源配置的低效性。如何建立媒体-网络跨域联合率失真优化理论模型以及实现基于信源特性的编码性能与复杂度的均衡是亟待解决的重要问题。为了解决这一问题,文献[16]发现了网络拓扑与用户行为协同感知的媒体-网络联合率失真规律,提出了网络动态适配的视频编码方法,建立了结合跨域失真度量和信源信道速率约束条件下的信源信道联合复杂度率失真优化模型:

$$\min P_s(R_s) + P_c(R_s, R_c) + \epsilon_t(P_t(g(\gamma), R_s, R_c)), \quad (8)$$

$$\text{s.t. } D_t \leq D_0, R_s + R_c \leq R_0.$$

这一模型通过综合考虑媒体内容的信源特性和网络信道的约束条件,优化了信源和信道的联合复杂度与失真度。通过动态适配视频编码方法,我们能够根据网络环境的变化对媒体内容进行实时编码优化,确保在不同的网络环境下实现最佳的传输效果。

同时,为了适配多媒体终端在不同的网络动态特性,文献[17]提出了一种基于端到端 QoS(服务质量)跨层映射的优化方法(图9),它是包括媒体可伸缩编码、网络感知端系统、网络 QoS(服务质量)控制等在内的一个全局优化模型框架,旨在将媒体编码、网络感知和 QoS 控制进行统一优化,进一步提升多媒体网络通信的效率和可靠性。该方法为跨域资源协同计算提供了新的理论支持,并为多媒体内容在动态网络环境中的分发提供了更加高效的解决方案。

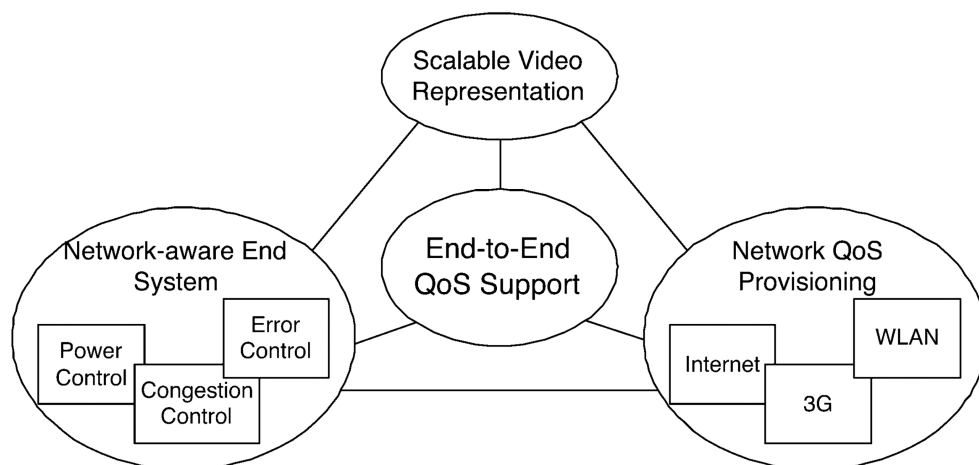


图9 端到端 QoS 跨层动态映射模型.

Figure 9 End-to-end QoS cross-layer dynamic mapping model.

## 4 未来研究方向

大模型正在对多媒体智能计算产生深远影响,多模态的预训练大模型正在蓬勃发展,多模态融合正迅速成为研究的热点.此外,如何使多媒体具身智能在动态开放环境下具备更精准的感知、认知和决策能力,实现动态开放环境下的场景自适应与内容生成已形成研究趋势.我们在总结过去20年研究成果、分享研究见解的同时,对多媒体智能的未来研究方向进行展望.

### 4.1 多模态大模型

作为多媒体智能的重要发展方向,多模态大模型在图像、文本、音频、视频等多种模态的数据理解与生成任务中展现出了卓越的性能和广阔的发展潜力.当前主流的多模态大模型(如GPT-4V, Qwen-VL等)主要依赖于大规模数据预训练和跨模态对齐技术,在跨模态信息的融合能力上取得了显著进展.然而,在跨模态推理方面,这些模型仍然存在诸多不足,尤其在抽象推理、多层次语义理解和长时依赖问题上,表现出较大的局限性.例如,在复杂场景下,现有模型难以对长视频进行跨模态的因果关系建模,从而实现精准推理.需要指出的是,这些不足与前文所讨论的理论困境高度契合:现有多模态大模型大多仍停留在语义驱动层面,缺乏对跨模态数据中复杂关联结构的系统建模;同时,其优化目标往往针对指定任务或特定模态,呈现出孤立优化的局限.未来的发展应紧密结合跨媒体表征学习与多媒体资源协同计算的理论成果,实现从语义驱动到关联驱动、从孤立优化到全局协同的跨越.具体而言,在训练数据层面,高质量的多模态数据和更高效的数据利用策略,对于提升模型在真实世界复杂分布下的学习质量至关重要;同时,结合数据增强和知识蒸馏技术,可进一步延长模型的学习过程,帮助其在跨模态推理中更好地捕捉关联规律.在模型架构层面,应在传统语义对齐机制之上,显式引入跨模态的结构化表示与因果推理建模,使模型能够理解和利用模态间的拓扑关联与演化关系.在训练策略层面,可以探索跨任务的全局协同优化,例如通过跨模态共享目标函数与强化学习奖励机制,促使模型在模态间和模态内建立合理的推理路径.此外,借助外部工具进行跨模态推理增强,以及引入自我反馈和自我反思机制,不仅能够优化推理过程,还能够在全局层面实现语义、关联和资源的动态协同,从而进一步提升模型的稳定性与可解释性,也是未来重要的研究方向.

### 4.2 感知与推理一体化

当前多媒体智能系统中,感知与推理往往相对割裂,即低层次特征提取与高层次逻辑推理各自独立,导致在面对复杂多变的环境时难以充分利用多模态信息的互补优势.事实上,感知与推理之间存

在紧密的联系:一方面,感知模块能够从原始数据中提取丰富的特征,为推理过程提供全面且细致的环境信息;另一方面,通过推理过程产生的抽象概念和预测结果又能反向指导感知模块,促使其聚焦于关键数据并优化特征提取。然而,现有的感知与推理一体化研究多集中于语义层面的特征互补,仍然缺乏对跨模态关联结构的系统建模。这意味着其本质上依然是“语义驱动”的浅层融合,难以支撑复杂场景下的深度推理。同时,大多数方法在优化过程中仅针对感知或推理单模块进行孤立改进,未能实现跨模态、跨任务的全局协同优化。未来的研究需要在感知-推理链条中引入关联驱动机制和全局协同机制,使模型能够同时捕捉多模态数据的结构性关联,并在不同任务间实现知识迁移与资源共享。一个可能的实现途径是模块化网络,它通过将整体任务划分为多个专注于不同子任务的子模块,使得每个模块均可独立负责视觉、语音、文本等多模态数据的感知,并通过设计模块间信息的共享和融合机制,形成完整的逻辑推理链路,从而实现从局部感知到全局推理的闭环协同。在这一过程中,不仅要保持语义层面的特征对齐,还应建立跨模态的关联表征,使不同模态之间的耦合关系能够被显式建模,从而推动感知与推理从语义驱动走向关联驱动。未来可能的研究方向包括以下几个方面:从模型架构的角度,设计基于一体化理念的自适应模块化网络架构,使系统能够根据复杂的动态开放环境调整各模块的功能分配和协同方式,实现感知与推理过程的无缝衔接;从模型训练的角度,通过联合训练和共享目标的方式促进各子模块联合优化,同时提升各子模块的信息提取能力和整体架构的逻辑推理能力。特别是在训练层面,应探索跨任务的全局协同优化策略,使感知与推理模块不再是孤立的功能单元,而是通过全局共享约束实现一致性的优化过程。此外,结合强化学习等技术构建具有自反馈和自优化能力的感知推理一体化系统,使系统在自我反思的过程中逐步强化感知与推理之间的互补效应,也是未来重要的研究方向。

#### 4.3 生成与理解一体化

当前多模态内容生成研究工作主要围绕扩散模型开展,多模态内容理解工作主要围绕基于自回归模型的多模态大语言模型开展,多模态生成与理解的研究相对割裂。然而,两个任务之间应当存在紧密的联系,高质量的多模态内容生成依赖强大的理解能力来捕捉用户的需求,而生成的内容也可以为理解模型提供更多的数据。因此,如何实现多模态内容生成与理解一体化是一个重要的未来研究方向。现有方法大多停留在语义层面的简单耦合,例如将生成模型的输出直接作为理解模型的输入,或利用理解结果指导生成,但这种模式依然属于“语义驱动”,缺乏对跨模态数据中深层次关联规律的刻画。与此同时,生成与理解往往在各自独立的优化框架下进行,缺乏跨任务的全局协同,难以在动态复杂场景下实现一致性推理与服务优化。未来的研究亟需在生成与理解一体化框架中引入关联驱动机制和全局协同优化策略,使系统能够同时捕捉生成与理解之间的结构性依赖关系,并实现跨任务、跨模态的知识迁移与优化协同。从数学建模上,是选择统一的自回归建模方法,还是选择扩散加自回归的建模方法;从模型架构设计上,是采用单一的稠密模型还是采用混合专家网络;从训练策略上,如何组织生成与理解的数据,都是值得探索的重要问题。在这一过程中,不仅要实现语义层面的共享,更要显式建模生成与理解之间的关联结构,使生成内容与理解机制能够在统一的表征空间中相互约束,从而推动多模态智能从语义驱动到关联驱动。另外,理解与生成任务之间的协同机制也值得探索,何时需要生成促进理解、何时需要理解促进生成,实现边理解边生成,提升模型在不同场景下的表现,也是未来的重要研究课题。

#### 4.4 多媒体具身智能

当前多媒体具身智能研究已初步实现多模态感知与物理动作的浅层耦合,但仍面临动态开放环境下的多模态感知数据异构性、多模态认知碎片化与多模态决策不确定性等挑战。现有研究多停留在模态间的语义对齐与动作映射,缺乏对感知、认知与决策复杂关联关系的系统建模,且不同环节往往局部优化,难以实现跨模态与跨任务的全局协同。未来的多媒体具身智能需要引入关联驱动与全局协同

机制,使感知-认知-决策不再割裂,而是在统一表征与跨域协同下实现闭环演化.未来研究方向包括:多模态具身感知,实现视觉-触觉-听觉的时空对齐与语义贯通;跨模态认知融合,通过神经网络模型与真实物理世界持续交互,将跨模态交互经验融合为可迁移知识;多模态人机协同,开发支持手势、语音、眼动混合多模态输入的人机协同机制,实现人类意图与机器动作的无缝衔接;以及跨模态博弈决策,训练实体在信息不对称条件下达成多模态共识,优化复杂场景中的多模态协作效率.在这些方向中,具身感知应从语义驱动走向关联驱动,显式建模多模态之间的拓扑与因果关系;具身认知与决策则需避免孤立优化,通过跨任务共享和跨域资源协同实现从局部改进到全局协同的演化.总体而言,多媒体具身智能的目标是实现具有自主演化能力的持续自适应智能,是未来多媒体智能的一个重要研究方向.

#### 4.5 多媒体空天智能

当前多媒体空天智能研究已初步实现卫星遥感影像的可视化呈现及与飞行器空天信息交互,但在广域实时感知、多源异构数据融合及沉浸式空间自主决策层面仍存在显著瓶颈.核心挑战在于空天信息的多尺度动态性,从厘米级地表监测到千米级气象云图;多模态数据的时空异构性,遥感光谱、电磁频谱、用户需求等语义割裂;以及飞行器的决策预设性,所有飞行路径均由地面站精准操控.其核心科学问题在于如何构建空天地一体化的多模态时空自演化飞行系统,实现从预设规则可视化到动态时空自演化的范式跃迁.未来研究方向包括以下4方面.空天多模态大模型:融合高分辨率遥感影像、卫星通信流、导航信号及社交媒体地理标记数据,构建具备空天地自适应交互的空天多模态大模型;多模态空间推理:借鉴具身智能的空间认知机制,开发空间关系语义网,使系统能理解“环月轨道空间站与陨石坑的拓扑关系”等复杂空间语义,实现类人空间情境认知;多模态时空演化:建立空天世界模型,模拟如婴儿学习般的自演化机制,通过持续吸收多源卫星数据流自主更新空天认知;多模态星地集群协同:利用强化学习优化低轨卫星群自主轨道计算与避障策略,同步驱动地面无人系统形成“卫星感知-边缘计算-地面执行”的闭环响应链,构建星地集群协同系统.综上,该方向旨在构建具备自演化能力的空天飞行系统,突破传统人工预设规则,最终形成“星地集群协同、时空持续演化”的空天智能,是未来多媒体智能的一个重要研究方向.

#### 4.6 自主机器学习

传统多媒体智能计算往往假设训练和测试分布,在动态开放真实世界中,多媒体数据所遇到的环境总是不断动态变化的,需要其能够具备自适应不同动态环境的能力,面临着训练和测试分布不一致的重大挑战.为了适应动态开放环境,自主机器学习是一个研究方向.自主机器学习通过自我驱动来进行具有高度自主性的自主学习过程,其学习结果反过来又可以进一步提升自我.自主机器学习主动探索感知环境,对数据选择、模型适配、任务切换三大核心过程进行动态自我优化,同时根据环境反馈和自我状态,通过动态自主演化进一步提升学习能力.

**致谢** 感谢崔鹏、王鑫等对本文的贡献.

#### 参考文献

- 1 Zhu W W, Wang X, Gao W. Multimedia intelligence: when multimedia meets artificial intelligence. *IEEE Trans Multimedia*, 2020, 22: 1823-1835
- 2 Zhu W W, Wang X, Li H. Multi-modal deep analysis for multimedia. *IEEE Trans Circuits Syst Video Technol*, 2020, 30: 3740-3764
- 3 Zhu W W, Cui P, Wang Z, et al. Multimedia big data computing. *IEEE Multimedia*, 2015, 22: 96-c3
- 4 Zhu W W, Luo C, Wang J, et al. Multimedia cloud computing. *IEEE Signal Process Mag*, 2011, 28: 59-69
- 5 Zhai X, Peng Y, Xiao J. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Trans Circuits Syst Video Technol*, 2013, 24: 965-978

- 6 Ou M D, Cui P, Pei J, et al. Asymmetric transitivity preserving graph embedding. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. 1105–1114
- 7 Wang D X, Cui P, Zhu W W. Structural deep network embedding. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. 1225–1234
- 8 Wang X, Cui P, Wang J, et al. Community preserving network embedding. In: Proceedings of AAAI, 2017. 203–209
- 9 Wang X, Ji H, Shi C, et al. Heterogeneous graph attention network. In: Proceedings of The World Wide Web Conference, 2019. 2022–2032
- 10 Ma J X, Cui P, Kuang K, et al. Disentangled graph convolutional networks. In: Proceedings of ICML, 2019. 4212–4221
- 11 Georgiadis L, Neely M, Tassiulas L. Resource Allocation and Cross-layer Control in Wireless Networks. Hanover: Now Publishers Inc., 2006. 1–144
- 12 Jiang M, Cui P, Chen X, et al. Social recommendation with cross-domain transferable knowledge. IEEE Trans Knowl Data Eng, 2015, 27: 3084–3097
- 13 Jiang M, Cui P, Liu R, et al. Social contextual recommendation. In: Proceedings of CIKM, 2012. 45–54
- 14 Wang Z, Zhu W W, Chen X, et al. Propagation-based social-aware multimedia content distribution. ACM Trans Multimedia Comput Commun Appl, 2013, 9: 1–20
- 15 Zhang X Y, Zhang Q, Zhang Z, et al. A construction of locality-aware overlay network: mOverlay and its performance. IEEE J Sel Areas Commun, 2004, 22: 18–28
- 16 Zhang Q, Ji Z, Zhu W W, et al. Power-minimized bit allocation for video communication over wireless channels. IEEE Trans Circuits Syst Video Technol, 2002, 12: 398–410
- 17 Zhang Q, Zhu W W, Zhang Y-Q. End-to-end QoS for video delivery over wireless Internet. Proc IEEE, 2005, 93: 123–134

## Research progress on multimedia intelligence computing

Wenwu ZHU\*

*Department of Computer Science and Technology, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China*

\* Corresponding author. E-mail: wwzhu@tsinghua.edu.cn

**Abstract** Multimedia intelligent computing is an interdisciplinary field that integrates artificial intelligence and multimedia computing. Its core objective is to achieve deep understanding and efficient service of multimodal data through intelligent representation, analysis, and collaborative computation of large-scale multimedia data. This field faces two major challenges. First, the scale of multimedia data entities grows exponentially over time, and their associative combinations explode exponentially, resulting in a dual exponential increase in the scale of data entity relationships. Second, due to the exponential growth of multimedia data volume, the heterogeneous and complex nature of network structures, and dynamic concurrent user demands, the network resources required for user access to media also increase exponentially over time. To address the challenge of discovering relationships and patterns under the dual exponential growth of associations in large-scale multimedia data, this paper introduces recent advances in the theory and methods of associative representation for multimedia big data. By uncovering the mapping principles between graph topological space and vector representation space, we discovered a unified mathematical expression for vectorized measurement of graph asymmetric transitivity. We proposed a structure-preserving deep associative representation method and developed a graph representation learning model that integrates micro-level local structures with macro-level global structures. This achieves a transition in graph associative representation learning from shallow to deep layers and from micro to macro structures. To tackle the challenge posed by the conflict between the exponential growth of large-scale multimedia content and limited network resources in heterogeneous network environments, this paper explores a new paradigm of multimedia network resource computing using big data analytics theory. We revealed the interaction mechanisms among media content, user behavior, and network topology, proposed a media content dissemination prediction model based on user behavior and a large-scale network topology model, and established a new theory for cross-domain collaborative computation of resources for large-scale multimedia. This enables a shift in network resource allocation from passive and static configurations to active and dynamic adaptation. The research outcomes advance multimedia intelligence from semantics-driven to relation-driven approaches and from isolated optimization to global collaboration, offering significant application value in scenarios such as public safety, smart cities, and the internet economy.

**Keywords** multimedia big data, cross-media intelligence, multimedia computing, big data analytics