

# 异质网络表征学习的研究进展

石川<sup>1</sup> 孙怡舟<sup>2</sup>

<sup>1</sup>北京邮电大学

<sup>2</sup>美国加州大学洛杉矶分校

关键词：异质网络 表征学习

## 异质网络的表征学习

信息网络在现实世界中无处不在，如航线网络、论文引用网络、社交网络以及万维网等。对这类网络的分析和研究，不管在工业界还是学术界都受到了广泛关注。通常，信息网络构成的图模型可以由邻接矩阵来表示，因此，早期的处理图结构的工作大部分采用高维稀疏向量的形式，再用矩阵分析的方法。然而，由于现实中网络的稀疏性以及其不断增长的规模，又对此类方法提出了严峻的挑战。一种更为有效的方式是将网络节点映射到一个低维向量空间中，即用一个低维稠密的向量来表示网络中的任意节点。信息网络的表征学习<sup>[1]</sup>就是将网络中的节点或者边映射到一个低维的空间里，从而可以更加灵活地应用于不同的数据挖掘任务中，如网络可视化、节点分类以及链路预测等。

根据包含的节点或边的类型是否相同，信息网络可划分为同质信息网络（也简称为同质网络）和异质信息网络（也简称为异质网络）。其中，同质网络包含单一的节点类型和边类型，如合作网、朋友圈等。现在已有大量的工作致力于同质网络的表征学习。这些工作大多利用已有的深度模型，结合网络的特征，学习网络中节点或边的特征表示。代表性模型包括：DeepWalk 模型<sup>[2]</sup>，将随机游走和 skip-gram 模型结合起来学习网络节点表示；LINE 模型<sup>[3]</sup>，在一阶邻居相似性的基础上加上二阶相似性，从而学习到对大规模稀疏网络有更好的区分能力的节点表示；SDNE 模型<sup>[4]</sup>，借助于深度自动编码器来抽取网络结构的非线性特征。除了使用网络的结构拓扑信息，也有很多方法是利用节点的内容信息或其他辅助信息（文本、图像、标签）学习更准确更有意义的节点表示。一些综述论文更加全面

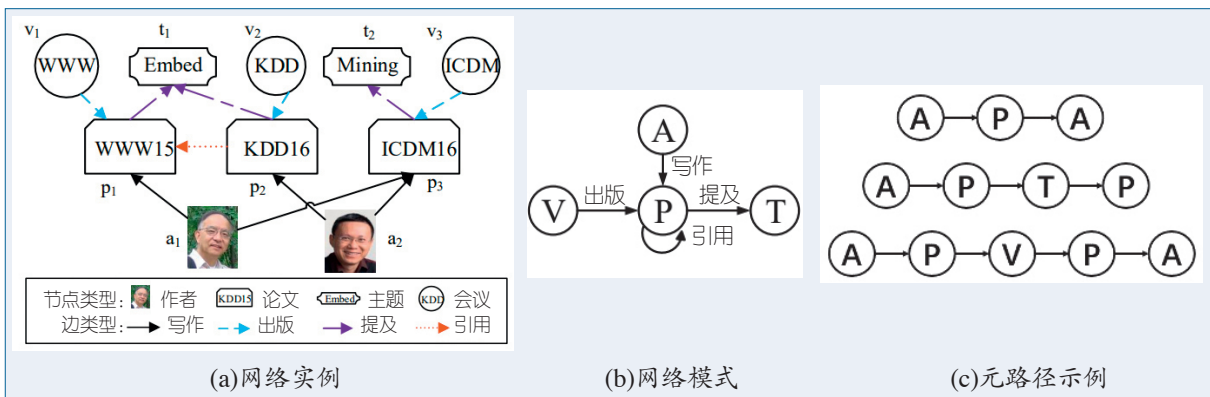


图1 由文献数据构造的异质信息网络

地总结了这方面的工作<sup>[5,6]</sup>。

同质信息网络往往是现实中的信息网络的一种简化，而包含不同类型节点和边的异质信息网络可以更加完整自然地对现实世界的网络数据建模<sup>[7,8]</sup>。图1(a)是一个由文献数据构成的异质信息网络的实例，图1(b)刻画了该网络的网络模式。该网络包含作者(A)、会议(V)、论文(P)、主题(T)等不同类型的对象以及这些对象之间不同类型的关系。在异质信息网络中，元路径表示连接两个对象之间的关系组合，它刻画了网络中包含的丰富的语义信息。例如，在图1(c)中，有多条元路径连接作者：APA表示作者之间的合作关系；APVPA表示两个作者在同一个会议上发表过文章。由于异质网络包含了更加全面的信息，蕴含了更加丰富的语义信息，针对这类网络的研究成为近些年数据挖掘的研究热点<sup>[25]</sup>。

由于异质网络的特殊性，同质网络的表征学习方法并不能直接应用于异质网络。异质信息网络的复杂性也为网络表征学习提出了新的挑战。**(1) 节点和边的异质性带来的挑战。**不同类型的节点和边代表不同的对象，因此，在异质网络的表征学习中需要考虑将不同类型的对象映射到不同的空间中。另外，如何有效地保存每个节点的异质邻居以及有效地处理异质的节点序列也是值得考虑的问题。**(2) 异质网络中丰富的信息带来的表示融合挑战。**异质网络中的各类信息蕴含着丰富的语义，它从多个维度来刻画节点的意义，因此，如何有效地抽取和利用异质网络的多维度信息，并有效地融合这些信息以便全面地学习节点的表示是一个巨大的挑战。

## 异质网络表征学习的主要进展

异质网络表征学习兴起于最近的两三年，但是发展迅猛。当前的异质网络表征学习大致可以分为三种类型：基于随机游走的方法、基于分解的方法和基于深度网络的方法。

### 1. 基于随机游走的方法

随机游走作为一种经典的图分析模型，常用于

刻画网络中节点间的可达性，因此也被广泛应用于网络表征学习中采样节点的邻居关系。在同质网络中，节点类型单一，可以沿着任意的路径游走；而在异质信息网络中，由于节点和边关系的类型约束，通常采用基于元路径的随机游走模型。图2展示了电影异质网络中基于UMU等元路径进行游走的样例。基于元路径的随机游走可以更好地抽取网络中的结构信息并蕴含丰富的语义。

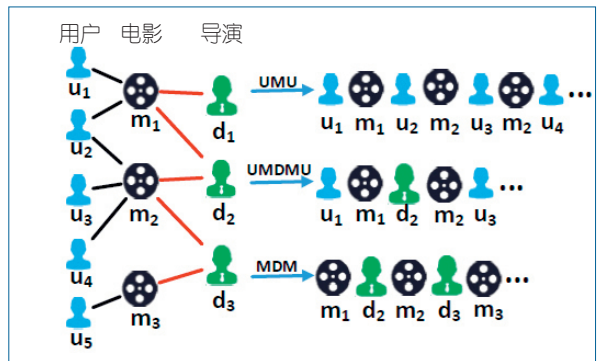


图2 电影异质网络中基于元路径随机游走实例

Metapath2vec<sup>[9]</sup>通过在异质信息网络中做基于元路径的随机游走来抽取节点结构信息，并利用skip-gram算法来学习节点表示。在Metapath2Vec基础上，作者进一步深化节点类型带来的差异提出了Metapath2Vec++。Metapath2Vec++通过将不同类型的节点映射到不同的向量空间，进一步刻画节点间的区别。HIN2Vec<sup>[10]</sup>同时考虑了不同类型节点及节点间复杂多样的关系，通过刻画节点对其之间的边关系来学习节点及元路径的向量表示。HINE<sup>[11]</sup>首先基于元路径随机游走来计算节点间的相似性，并将其作为监督信息来指导节点的向量表示。Shang等人提出的ESim<sup>[12]</sup>在异质网络上基于元路径进行随机游走，通过使元路径实例的概率最大化来学习出现在该实例中节点的向量表示。上述方法均尝试通过随机游走的方式来获得节点序列，包括普通随机游走和基于元路径的随机游走。模型最终的表示结果也多种多样，包括节点向量表示、边的向量表示和节点对相似性的向量表示等。

## 2. 基于分解的方法

为了缓解异质网络的复杂性,可以将异质网络分解为比较简单的网络,分别对这些网络进行表征学习,然后再将这些信息融合起来,达到“分而治之”的效果。图3展示了在电影异质网络中通过元路径抽取用户之间不同关联关系的同质网络,分别进行表征学习之后再行表示融合。

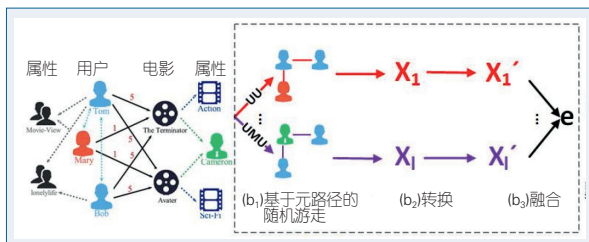


图3 异质网络中基于分解的表征学习方法

Chuan Shi 等人提出的 HERec<sup>[13]</sup> 模型利用元路径抽取异质网络中的多个同质网络,并对这些同质网络进行表征学习,然后通过融合函数对不同的表示进行融合,并结合矩阵分解模型进行评分预测。PTE<sup>[14]</sup> 将从文本中构建的异质网络分解成3个子网: word-word 网络、word-document 网络和 word-label 网络。对上述网络进行表征学习可以得到不同类型对象(如 word、document 和 label)的向量表示。EOE<sup>[15]</sup> 将复杂的学术异质网分解为单词共现网络和作者合作网络,对各个子网内节点对和子网间节点对同时进行表征学习。上述方法均尝试将复杂的异质网络分解为多个简单的网络。通过对分解后的网络分别进行表征学习,较好地应对了网络异质性带来的挑战。

## 3. 基于深度网络的方法

近年来,深度神经网络在计算机视觉和自然语言处理等领域取得了巨大成功。一些工作也开始尝试利用深度模型来对异质网络中不同类型和不同来源的数据分别进行建模。

BL-MNE<sup>[16]</sup> 采用自动编码器分别对异质网络在不同元路径下的信息进行编码,再对这些信息进行

联合编码,不同网络之间通过转移矩阵进行融合。SHINE<sup>[17]</sup> 也利用自动编码器分别对社交网络、情感网络和画像网络中的异质信息进行压缩编码得到特征表示,然后通过聚合函数对这些表示进行融合得到最终的节点表示。针对文本和图像并存的异质网络,HNE<sup>[18]</sup> 通过 CNN 和 MLP 分别对文本和图像数据进行特征抽取,然后通过转移矩阵将不同类型的数据投影到同一个空间。上述方法均采用深度模型来学习节点的向量表示。相对于浅层模型,深度模型可以更好地对非线性关系进行建模,能够抽取节点所蕴含的复杂语义信息。

## 应用

异质信息网络包含了丰富的语义信息,并对现实网络中的复杂结构进行了建模。该方法已经广泛应用于各类数据挖掘问题上。学习异质网络的嵌入表示能够较好地刻画网络中不同类型节点之间的复杂关联,以便和其他模态信息融合,应用于各类任务场景。

### 1. 节点分类

节点分类是网络数据挖掘中的常见任务之一,也是用于评估节点嵌入性能的基准任务。例如,在文献引用网络上,不同的文章可以根据文章的关键词划分为不同的领域。Guolei Sun 等人<sup>[19]</sup> 提出的 GERI 模型通过构建一个双向异质网络,为富文本信息图提出了一种新颖的表征学习框架,并将学习到的向量表示用于网络节点分类。Jian Tang 等人<sup>[14]</sup> 针对大规模异质文本网络设计了半监督表征学习模型 PTE,该模型同时利用带标签数据和无标签数据学习文本的嵌入表示,依据文本的低维向量表示对其进行分类。Metapath2vec<sup>[9]</sup> 通过在异质信息网络中做基于元路径的随机游走同时学习节点的向量表示,并将学到的节点向量用于节点分类任务,取得了较好的效果。

### 2. 链路预测

作为网络分析中的重要问题,链路预测任务的

目的在于预测网络中丢失的边，或者可能出现的边。在异质信息网络中，可以利用更多类型的节点信息，学习到含有更丰富信息的表示向量，并依此预测节点间是否存在链路。Hongwei Wang 等人<sup>[17]</sup>提出的 SHINE 模型解决异质信息网络中的情感链路预测问题，该模型利用多个深度自编码器学习节点的向量表示，并将其融合到同一空间用于情感链路预测。Taoyang Fu 等人<sup>[10]</sup>提出的 HIN2Vec 模型基于随机游走同时学习节点和边的向量表示，并将学到的节点向量用于链路预测任务，取得了较好的效果。

### 3. 推荐系统

个性化推荐是根据用户的购买历史，为用户推荐其感兴趣的信息和商品。在现代推荐系统中，用户和物品都关联着复杂的异质信息，因此，基于异质信息网络的推荐不仅能够考虑到用户和物品之间的信息，还能够考虑到其他属性节点带来的丰富的语义信息。Chuan Shi 等人<sup>[13]</sup>提出的 HERec 推荐模型，基于随机游走策略生成节点序列，学习节点的向量表示，然后利用扩展的矩阵分解模型产生推荐。Huan Zhao 等人<sup>[20]</sup>根据异质信息网络中不同的元图计算得到多个交换矩阵，然后使用矩阵分解基于不同的元图分别学习多个不同的用户和物品的隐向量，最后通过因子分解机学习用户对物品的评分，并进行推荐。Yongfeng Zhang 等人<sup>[21]</sup>提出的 JRL 推荐模型设计了一个深度表征学习框架，利用异质信息源（评分、评论和图片）学习用户和物品的特征表示，然后通过一个额外层学习到用户和物品的联合表示，最后通过 Pair-wise 学习，来推荐 top-N 的物品。Ting Chen、Yizhou Sun 等人<sup>[22]</sup>提出一种混合表征学习推荐框架，并对采样策略进行优化，提出了三种可以大大提升训练效率的采样策略。

### 4. 其他任务

异质信息网络由于蕴含丰富的语义信息，常被应用于各种任务。针对双盲评审的作者身份识别问题，Ting Chen、Yizhou Sun 等人<sup>[23]</sup>基于学术合作异质网络，提出了一个任务引导和路径增强的异质网

络表示模型。该模型根据特定任务选择元路径，学习文章的向量表示，对双盲审情况下的匿名论文作者进行识别，取得了较高的准确率。针对不同类型事件的异常检测，Ting Chen、Yizhou Sun 等人<sup>[24]</sup>提出了一种实体嵌入模型 APE。该模型利用在不同事件中观察到的实体的共生性，将待检测事件和其他实体建模为一个异质信息网络，然后将各个实体学习为一个低维向量，从而有效地检测异常事件。

## 总结和展望

网络表征学习已经成为数据挖掘的一个热点方向。表征学习可以为各类学习任务提供优秀的特征输入，并方便和其他模态的信息融合。但是，目前的工作主要集中在同质网络上，异质网络表征学习的研究还相对较少。异质网络是表示现实世界中对象交互的更加通用的建模方式，因此异质网络的表征学习还有巨大的发展空间。未来的发展方向如下。

### 1. 异质信息网络中的信息融合

异质信息网络中存在复杂的语义关系。根据不同的元路径可以从多个维度对节点的丰富信息建模。如何有效地自动筛选和融合不同元路径下的网络节点表示是一个值得关注的研究方向。目前节点的特征学习研究比较多，关系和元路径的表征学习还比较少。不同类型的节点有不同的特征表示空间，对这些表示空间之间的关系还缺乏深入探究。

### 2. 融合其他信息的表征学习

异质网络可以通过融合丰富的异质异构数据从而解决大数据的“多样性”挑战。因此异质网络天然包含丰富的多模态信息，如属性、文本和图像等。只有将更多模态的信息考虑进来，学习到的网络节点表示才能更加准确地对节点进行描述。而现有的主要工作都致力于对异质信息网络中的结构信息建模，没有充分挖掘网络中的其他模态信息。如何更好地融合这些多模态信息的表征学习将是一个研究难点。

### 3. 大规模动态网络

目前网络表征学习相关工作主要集中于小规模静态网络,但是现实世界中的网络往往规模较大并且动态变化。我们需要考虑节点和边的异质性,快速高效地获取新增节点的表示,研究面向增量计算和在线计算的表征学习方法。

### 4. 结合具体应用

目前表征学习算法主要集中在做通用的表征学习,与具体应用相结合做任务优化的节点表示的较少。在社区发现、异常检测等特定任务上,通用表征学习的效果往往不尽如人意。如何与具体应用结合学习网络表示是重要发展方向。另外,社交网络服务的大量兴起,积累了丰富的异质数据。如何将异质网络表征学习技术真正应用于实际业务也是值得关注的方向。 ■



石川

CCF 高级会员。北京邮电大学教授。主要研究方向为数据挖掘、机器学习、演化计算。  
shichuan@bupt.edu.cn



孙怡舟 (Yizhou Sun)

美国加州大学洛杉矶分校助理教授。主要研究方向为数据挖掘、机器学习、网络科学。  
yzsun@cs.ucla.edu.

### 参考文献

- [1] Cai H, Zheng V W, Chang K C C. A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications[OL]. (2017-09-22).arXiv preprint arXiv:1709.07604.
- [2] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations[C]//*Proceedings of the 20th ACM SIGKDD International Conference ON Knowledge Discovery and Data Mining*. ACM, 2014: 701-710.
- [3] Grover A, Leskovec J. node2vec: Scalable feature learning for networks[C]//*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016: 855-864.
- [4] Wang D, Cui P, Zhu W. Structural deep network embedding[C]//*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016: 1225-1234.
- [5] Goyal P, Ferrara E. Graph Embedding Techniques, Applications, and Performance: A Survey[OL].(2017-05-08). arXiv preprint arXiv:1705.02801.
- [6] Hamilton W L, Ying R, Leskovec J. Representation Learning on Graphs: Methods and Applications[OL]. (2017-09-27). arXiv preprint arXiv:1709.05584.
- [7] Shi C, Li Y, Zhang J, et al. A survey of heterogeneous information network analysis[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2017, 29(1): 17-37.
- [8] Shi C, Philip S Y. *Heterogeneous Information Network Analysis and Applications*[M]. Springer, Data Analytics, 2017.
- [9] Dong Y, Chawla N V, Swami A. metapath2vec: Scalable representation learning for heterogeneous networks[C]// *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017: 135-144.
- [10] Fu T Y, Lee W C, Lei Z. HIN2Vec: Explore Meta-paths in Heterogeneous Information Networks for Representation Learning[C]//*Proceedings of the 26th ACM International Conference on Information and Knowledge Management*. ACM Press, 2017: 1797-1806.
- [11] Huang Z, Mamoulis N. Heterogeneous Information Network Embedding for Meta Path based Proximity[OL]. arXiv preprint arXiv:1701.05291, 2017.
- [12] Shang J, Qu M, Liu J, et al. Meta-Path Guided Embedding for Similarity Search in Large-Scale Heterogeneous Information Networks[OL]. arXiv preprint arXiv:1610.09769, 2016.
- [13] Shi C, Hu B, Zhao W X, et al. Heterogeneous Information Network Embedding for Recommendation[OL]. arXiv preprint arXiv:1711.10730, 2017.
- [14] Tang J, Qu M, Mei Q. Pte: Predictive text embedding through large-scale heterogeneous text networks[C]// *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015: 1165-1174.

- [15] Xu L, Wei X, Cao J, et al. Embedding of Embedding (EOE): Joint Embedding for Coupled Heterogeneous Networks[C]//Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. ACM, 2017: 741-749.
- [16] Zhang J, Xia C, Zhang C, et al. BL-MNE: Emerging Heterogeneous Social Network Embedding through Broad Learning with Aligned Autoencoder[OL].arXiv preprint arXiv:1711.09409, 2017.
- [17] Wang H, Zhang F, Hou M, et al. SHINE: Signed Heterogeneous Information Network Embedding for Sentiment Link Prediction[OL]. arXiv preprint arXiv:1712.00732, 2017.
- [18] Chang S, Han W, Tang J, et al. Heterogeneous network embedding via deep architectures[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 119-128.
- [19] Sun G, Zhang X. Graph Embedding with Rich Information through Bipartite Heterogeneous Network[OL]. arXiv preprint arXiv:1710.06879, 2017.
- [20] Zhao H, Yao Q, Li J, et al. Meta-Graph Based Recommendation Fusion over Heterogeneous Information Networks[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017: 635-644.
- [21] Zhang Y, Ai Q, Chen X, et al. Joint representation learning for top-n recommendation with heterogeneous information sources[C]// Proceedings of the 26th ACM International Conference on Information and Knowledge Management. ACM Press, 2017.
- [22] Chen T, Sun Y, Shi Y, et al. On Sampling Strategies for Neural Network-based Collaborative Filtering[C]// Proceedings of the 23th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2017.
- [23] Chen T, Sun Y. Task-Guided and Path-Augmented Heterogeneous Network Embedding for Author Identification[C]//Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. ACM, 2017: 295-304.
- [24] Chen T, Tang L A, Sun Y, et al. Entity embedding-based anomaly detection for heterogeneous categorical events[J]. arXiv preprint arXiv:1608.07502, 2016.
- [25] 石川, 孙怡舟, 菲利普·俞. 异质信息网络的研究现状和未来发展 [J]. 中国计算机学会通讯 . 2017, 13(11): 35-40.