软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.000000]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

异质信息网络分析与应用综述*

石 川[1,2], 王睿嘉[1,2], 王 啸[1,2]

[1](北京邮电大学 计算机学院,北京 100876)

[2](智能通信软件与多媒体北京市重点实验室,北京 100876)

通讯作者: 王啸, E-mail: xiaowang@bupt.edu.cn

摘 要: 实际系统往往由大量类型各异、彼此交互的组件构成.当前大多数工作将这些交互系统建模为同质信息网络,并未考虑不同类型对象的复杂异质交互关系,因而造成大量信息损失.近年来,越来越多的研究者将这些交互数据建模为由不同类型节点和边构成的异质信息网络,从而利用网络中全面的结构信息和丰富的语义信息进行更精准的知识发现.特别是,随着大数据时代的到来,异质信息网络能够自然融合异构多源数据的优势使其成为解决大数据多样性的重要途径.因此,异质信息网络分析迅速成为数据挖掘研究和产业应用的热点.本文对异质信息网络分析与应用进行了全面综述.除了介绍异质信息网络领域的基本概念外,重点聚焦基于异质网络元路径的数据挖掘方法、异质信息网络的表示学习技术和实际应用三个方面的最新研究进展,并对未来的发展方向进行了展望.

关键词: 异质信息网络;元路径;网络表示学习;图神经网络

中图法分类号: TP138

A Survey of Heterogeneous Information Networks Analysis and Applications

SHI Chuan[1,2], WANG Rui-Jia[1,2], WANG Xiao[1,2]

[1](School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China)

[2](Beijing Key Laboratory of Intelligent Communication Software and Multimedia, Beijing 100876, China)

Abstract: The real-world systems usually contain different types of components that interact with each other. Most existing work models these interaction systems as homogeneous information networks, which does not considerthe heterogeneous interaction relationships among objects, resulting in lots of information loss. In recent years, more researchers model these interaction data as heterogeneous information networks (HINs) and conduct knowledge discovery based on the comprehensive structural information and rich semantic informati on in HINs. Especially, with the advent of the era of big data, HINs naturally merge heterogeneous data sources, which makes it an important way to solve the variety of big data. Therefore, heterogeneous information network analysis has quickly become a hot spot in data mining research and industrial applications. This article provides a comprehensive overview of heterogeneous information network analysis and applications. In addition to the basic concepts in heterogeneous information networks, the focus of this article is on the latest research progress in meta-path based data mining, heterogeneous information networks representation learning and practical applications of heterogeneous information networks. In the end, this article points out the possible directions of future development.

Key words: heterogeneous information networks; meta-path; network representation learning; graph neural networks

现实生活中形形色色的系统,通常由大量类型各异、彼此交互的组件构成[1],例如生物、社交和计算机系统等.在这些系统中相互作用的组件可以抽象为信息网络[2].信息网络无处不在,已经成为了现代信息基础设施的重要组成部分.因此,信息网络分析引起了学术界和工业界研究者的共同关注.为更好地进行分析挖掘,大多数工作将信息网络建模成同质信息网络(Homogeneous Information Network,简称同质网络),即网络中仅包含相同类型的对象和链接,例如作者合作网[3]和朋友圈[4]等.同质网络建模方法往往只抽取了实际交互系统中的部分信

* 基金项目: 国家自然科学基金(No.61772082, 61702296, 61806020); 国家重点研发计划(2018YFB1402600)

Foundation item: National Natural Science Foundation of China (No.61772082, 61702296, 61806020); National Key Research and Development Program of China (2018YFB1402600)

息,或者没有区分对象及其之间关系的异质性,从而造成不可逆的信息损失.近年来,更多的研究者将多类型且互连的网络化数据建模为异质信息网络(Heterogeneous Information Network,简称异质网络*),实现对现实世界更完整自然的抽象.例如,文献数据中包含作者、论文、会议等不同类型的对象,这些对象间存在多种类型的关系:作者和论文间的撰写/被撰写关系、会议和论文间的出版/被出版关系等.利用异质网络建模这种类型丰富且交互复杂的数据,可以保留更全面的语义及结构信息.

相较于同质网络,异质网络建模带来了两方面的好处.(1)异质网络是融合信息的有效工具,不仅可以自然融合不同类型对象及其交互,而且可以融合异构数据源的信息.特别地,随着“大数据”时代的到来,在“大数据”中许多类型不同的对象互联,将这些交互对象建模为同质网络很困难,但可以很自然地利用异质网络建模.同时,不同平台产生的异构多源“大数据”仅捕获了部分甚至是有偏差的特征,异质网络也可以自然融合这些异构数据源的信息,从而全面刻画用户特征[5].因此,异质网络建模不仅成为解决大数据多样性的有力工具[6],而且成为宽度学习的主要方法[7].(2)异质网络中多类型对象和关系共存,包含丰富的结构和语义信息,从而为发现隐含模式提供了精准可解释的新途径.例如,推荐系统的异质网络中不再只有用户和商品这两种对象,而是包含店铺、品牌等更全面的内容,关系也不再只有购买,而是含有收藏、喜爱等更精细的交互.基于这些信息,利用元路径[8]和元图[9,10]等语义挖掘方法,可以产生更精细的知识发现,如提高推荐系统的可解释性及准确率等.

基于以上信息融合优势,异质网络分析迅速成为数据挖掘、数据库和信息检索等领域的研究热点[6,11],大量论文发表在相关领域的顶级会议和期刊上,且全面涉及了各类基本任务,如分类、聚类、推荐等.随着网络表示学习的兴起,异质网络表示学习也迅速激发了广大研究者的兴趣,学得的低维向量表示在加速下游任务的同时也可以提升性能表现[12,13].近年来,异质网络建模被广泛应用到实际系统中,如电子商务[14]和网络安全[15],同样取得了显著效果.与此同时,相关研究者举办了一些异质网络研讨会和讲习报告,吸引了人工智能从业者的广泛关注.例如,异质网络分析研讨会(HINA†和 HENA‡)与 IJCAI 和 CIKM 等会议联合举办了多年.

本文全面总结了异质网络分析的工作,特别是近三年的研究新进展.目前已经有一些英文文献[2,6,16,17]介绍了该方向的发展情况.与现有工作相比,本文的不同主要体现在两方面:(1)[2,16,17]侧重介绍作者自身的工作,而本文则通过系统调研已发表的 160 多篇异质网络相关论文,总结了异质网络分析领域的总体进展. (2)[6]综述 2017 年前异质网络的研究内容,而本文全面涵盖了异质网络领域的最新发展和前沿成果,如加权元路径[18]、元图[9,10]和属性异质网络[19]等.特别地,近三年随着网络表示学习的兴起,本文着重介绍了异质网络表示学习的研究进展,且本文是第一篇系统介绍该研究方向的中文综述论文.此外，基于已有成果和发展趋势,本文还指出了该领域未来的研究方向.

本文的剩余部分如下组织:第 1 章介绍异质网络领域的基础知识;第 2 和 3 章分别从基于元路径的数据挖掘和异质网络表示学习两个方面全面介绍异质网络的发展现状;第 4 章介绍异质网络在实际问题中的应用;第 5 章总结全文并展望未来发展方向.

1 异质网络基础知识

在本节中,将介绍异质网络的主要定义和典型结构,并进一步指出异质网络与其他网络模型的区别与联系.

1.1 基本概念

信息网络是对现实世界的抽象,重点关注于对象及其之间的交互,形式化定义如下:

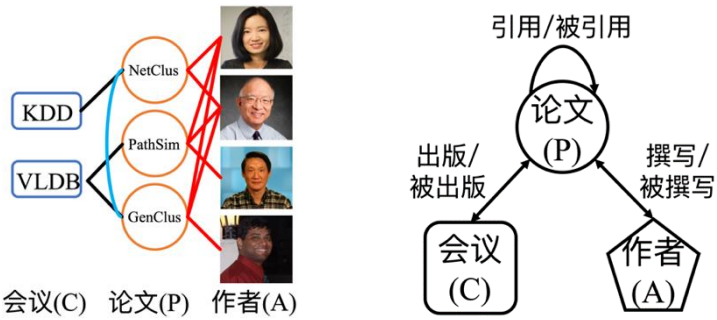
---

* 不少学者也将其称为异构信息网络或异构网络.由于这种网络中的节点和边存在不同的特质,因此称为异质网络更为贴切.另外,异构网络是通信网络里的专有名词,也容易引起不必要的误解.

† https://ijcai-17.org/tutorial-program.html

‡ http://www.shichuan.org/HENA2019.html

定义 1 信息网络[2,20].信息网络定义为一个具有对象类型映射函数$\varphi: V \to A$和关系类型映射函数$\psi: E \to R$的有向图$G = (V, E, \varphi, \psi)$.其中,每个对象$v \in V$属于对象类型集合$A$:$\varphi(v) \in A$中的一个特定对象类型,每条链接$e \in E$属于关系类型集合$R$:$\psi(e) \in R$中的一个特定关系类型.



(a)网络实例 (b)网络模式

Fig.1 Heterogeneous information networks of citation data

图 1 文献数据的异质网络

与社会网络分析不同,信息网络明确区分网络中的对象和关系类型,就产生了异质/同质网络的概念.

定义 2 异质/同质网络.若信息网络的对象类型数$|A| > 1$或者关系类型数$|R| > 1$,那么称之为**异质网络**;否则,称之为**同质网络**.

为更好地理解异质网络中复杂的对象和关系类型,网络模式的概念被提出,从而在元层次(即模式层次)上对网络进行描述.

定义 3 网络模式[2,20].网络模式记为$T_G = (A, R)$,是带有对象类型映射$\varphi: V \to A$和关系类型映射$\psi: E \to R$的信息网络$G = (V, E, \varphi, \psi)$的元模式.具体地,网络模式是定义在对象类型集合$A$上的有向图,并以$R$上的关系为边.

网络模式强调关于对象和关系集合的类型约束,这些约束使得异质网络半结构化,从而便于语义探索和模式挖掘.实际生活中,遵循某种网络模式的信息网络被称为该网络模式的**网络实例**.

示例 图 1 给出了文献数据所构建的信息网络.例如,从 DBLP§中抽取的涉及计算机科学研究者的网络,是典型的异质网络.图 1(b)说明了描述文献异质网络对象及其之间关系类型的网络模式.图 1(a)是图 1(b)的网络实例.在该实例中,包含三种类型的对象:论文(P),作者(A)和会议(C).链接连接不同类型的对象,而链接的类型由两种对象类型间的关系定义.例如,作者和论文间的链接表示撰写或被撰写的关系,而会议和论文间的链接表示出版或被出版的关系.

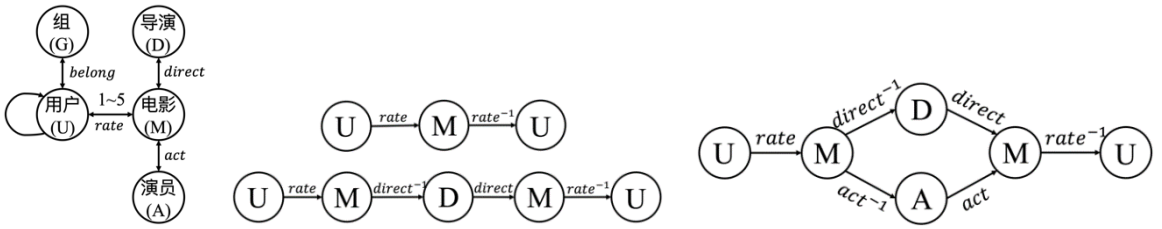
1.2 语义探索方法

与同质网络不同,异质网络中两对象可以通过不同类型定义的路径连接,而这些路径隐含不同的语义.

定义 4 元路径[8].元路径P是在网络模式$T_G = (A, R)$上定义的路径,记为$A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$.同时,定义对象$A_1, A_2, \cdots, A_{l+1}$间的复合关系$R = R_1 \circ R_2 \circ \cdots \circ R_l$,其中$\circ$表示关系上的合成运算符.简单起见,若相同对象类型间没有多种关系类型,则可以利用对象类型来表示元路径:$P = (A_1 A_2 \cdots A_{l+1})$.此外,对象$a_1$和$a_{l+1}$间的具体路径$p = (a_1 a_2 \cdots a_{l+1})$是路径$P$的**路径实例**.形式化地,如果在$p$中,对于每个$a_i$,都有$\varphi(a_i) = A_i$,且每条链接$e_i = \langle a_i, a_{i+1} \rangle$属于关系$R_i$,则记为$p \in P$.

§ http://dblp.uni-trier.de/

示例 以图 2 所示的电影推荐异质网络为例.用户可以通过元路径相连,如$U \xrightarrow{rate} M \xrightarrow{rate^{-1}} U(UMU)$路径和$U \xrightarrow{rate} M \xrightarrow{direct^{-1}} D \xrightarrow{direct} M \xrightarrow{rate^{-1}} U(UMDMU)$路径等.这些路径包含的语义不同,$UMU$路径指用户对同一电影打分(即共同评分关系),而$UMDMU$路径表示用户对同一导演的电影作品打分.



(a)网络模式 (b)元路径示例 (c)元结构/元图示例

Fig.2 Heterogeneous information networks of movie recommender system

图 2 电影推荐系统的异质网络

元路径本质上抽取了异质网络的子结构,并且体现了路径所包含的丰富语义信息,因而成为异质网络分析中的基本语义捕捉方法[8,21,22].但是由于其结构简单,在捕捉更精确或复杂的语义时往往受限.因而,作为元路径的扩展版本,许多更强大的语义探索工具先后被提出[9,18,23,24],极大提升了异质网络中的语义挖掘性能.

首先,元路径难以捕捉更加精细的语义.例如,UMU路径无法刻画精确到某些类型电影的共同评分关系.因此,受限元路径应运而生.

定义 5 受限元路径[23].受限元路径是基于某种特定约束的元路径,可以表示为$\mathcal{C}P = P|\mathcal{C}$.其中,$P = (A_1A_2 \dots A_l)$表示元路径,$\mathcal{C}$表示对元路径$P$中对象的约束.

受限元路径通过在不同约束条件下细化元路径来传达更丰富的语义.特别地,受限元路径长度为 1(即一种关系)时,退化为受约束的关系.

示例 受限元路径$UMU|M.T = "Comedy"$利用“Comedy”标签约束电影,使得该路径表示用户对喜剧电影的共同评分关系.同样,受限元路径$UMDMU|M.T = "Comedy"\&\&D = "Ang\ Lee"$表示用户对李安所导演的喜剧电影的共同评分关系.

其次,元路径并未考虑链接上的属性,如用户对于电影的评分信息,从而使得路径实例间链接的属性差异诱发较大的语义差异.因而,加权元路径的概念被提出以进一步约束链接属性信息.

定义 6 加权元路径[18].加权元路径是对关系属性值有所约束的一种扩展元路径,可以表示为$A_1 \xrightarrow{\delta_1(R_1)} A_2 \xrightarrow{\delta_2(R_2)} \cdots \xrightarrow{\delta_l(R_l)} A_{l+1}|\mathcal{C}$,也记作$A_1(\delta_1(R_1))A_2\big(\delta_2(R_2)\big)\cdots(\delta_l(R_l))A_{l+1}|\mathcal{C}$.

如果关系R在链接上具有属性值,则属性函数的函数值$\delta(R)$是关系R属性值范围内的一个取值集合;否则,$\delta(R)$为空集.$A_i \xrightarrow{\delta_i(R_i)} A_{i+1}$表示$A_i$与$A_{i+1}$之间的关系$R_i$具有属性值$\delta_i(R_i)$.另外,约束条件$\mathcal{C}$还可以用于约束属性函数之间的关系.若加权元路径中所有属性函数取值均为空集,相应的约束条件$\mathcal{C}$也为空集,则该路径退化为普通元路径,即普通元路径是加权元路径的一种特例.

示例 用户U与电影M间评分关系的属性值可以取 1 至 5 分.加权元路径$U \xrightarrow{1} M$(即$U(1)M$)表示用户对电影的评分为 1,也就意味用户并不喜欢这部电影;加权元路径$U \xrightarrow{1,2} M \xrightarrow{1,2} U$则指用户和目标用户不喜欢相同的电影.此外,

还可以在加权元路径的不同关系上灵活设置属性值函数约束.例如,路径$U(i)M(j)U|i=j$表示两用户在相同电影上的评分完全相同,而普通元路径UMU只能反映两用户对相同电影有评分,无法刻画其对电影的具体喜爱程度.

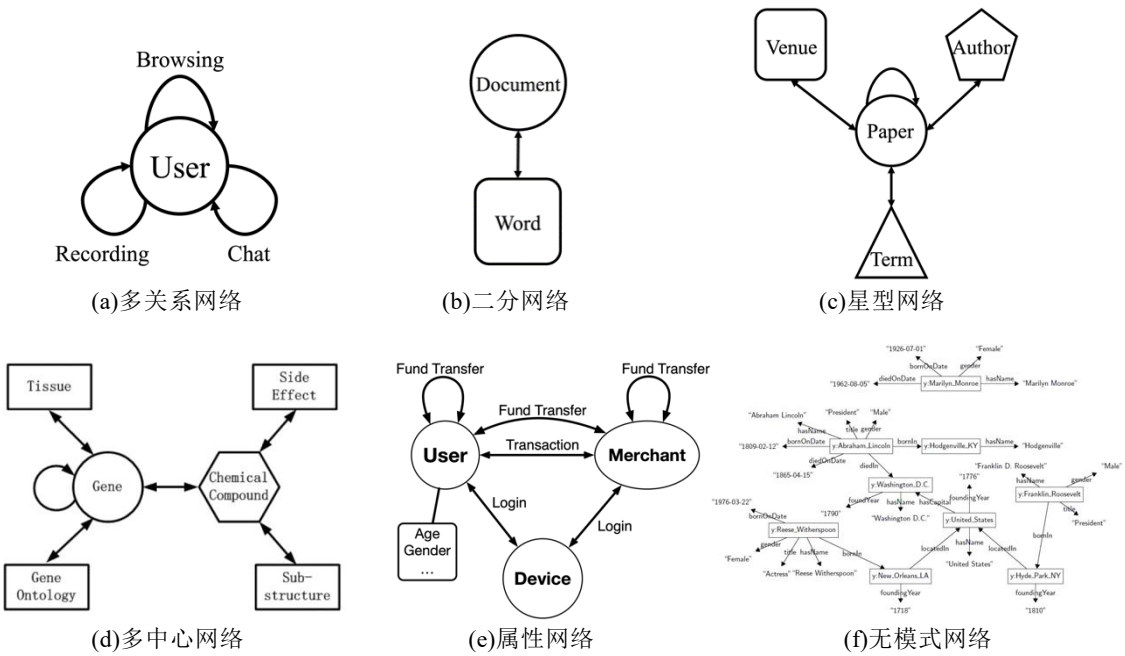
另外,元路径只能表示两对象间的简单关系,而元结构/元图可以融合多条元路径,方便地表达复杂语义.

定义 7 元结构/元图[9,10,65].元路径是定义在元模式$T_G=(A,R)$上的线性序列,而元结构/元图M可看作多条有公共节点的元路径组合而成的有向无环图.形式化地,元结构/元图M可记为$M=(V_M,E_M)$,其中V_M是M中节点集合,E_M是M中边集合.对于任意节点$v\in V_M$,v属于节点类型集合A;对于任意边$\langle u,v\rangle\in E_M$,$\langle u,v\rangle$属于链接类型集合$R$. 特别地,对于异质网络$G=(V,E,\varphi,\psi)$的子图$S=(V_S,E_S)$而言,当$S$和$M$的节点间存在双射函数$f:V_S\to V_M$时,$S$称为元结构/元图$M$的**实例**.其中,$f$需满足:对任意节点$v\in V_S$,$f(v)=\varphi(v)$;任意节点$u,v\in V_S$,边$\langle u,v\rangle\in E_S$当且仅当$\langle f(u),f(v)\rangle\in E_M$且$\psi(\langle u,v\rangle)=\langle f(u),f(v)\rangle$.显然,元路径是元结构/元图的特例.

示例 对于元路径$UMDMU$和$UMAMU$而言,只能分别描述两用户对同一导演的电影打分或已打分电影中出现相同演员,无法同时表述两条元路径蕴含的公共关系:两用户对于同一导演的电影作品进行了打分并且电影作品中出现了相同演员.而利用元结构/元图可以描述该语义,如图2(c)所示.可以看到,元结构/元图M是定义在网络模式上的有向无环图.

1.3 异质网络实例

本节列举了一些文献中广泛使用的异质网络,并按照结构特点进行了粗略分类.



(a)多关系网络 (b)二分网络 (c)星型网络

(d)多中心网络 (e)属性网络 (f)无模式网络

Fig.3 Common heterogeneous information network structures

图 3 常见异质网络结构

多关系网络 传统的多关系网络属于异质网络,该网络包含单类型对象和对象间的多类型关系.这种网络常存在于社交网站中,如 Facebook 和人人网,图 3(a)展示了这种网络的网络模式[25].可以看到,该网络只包含用户这一类对象,而用户间可以通过通话、浏览和聊天等不同类型的方式连接.

二分网络 作为典型的异质网络,二分网络被广泛用于构建两种类型对象之间的交互,例如用户-物品[24]和文档-单词[26]等.其主要特点为只含两种类型对象,且是二分图结构,即交互只存在于不同类型对象间,同类对象间没有

交互,图 3(b)展示了文档-单词二分网络的网络模式[26].作为二分图的扩展,k 分图[27]包含多种类型对象,相邻对象类型间存在关系.

星型网络 星型网络的网络模式中存在中心节点,其余非中心节点都与中心节点存在交互.例如在数据库表格中,目标对象及其属性对象自然地形成一个星型异质网络,其中目标对象作为中心节点连接不同的属性对象.如图 3(c)所示,文献信息网络是典型的星型模式异质网络[8,11],其包含不同对象(例如论文、会议、作者和术语)以及它们间的关系.许多其他的数据集也可以表示为星型网络,如电影数据[28,29]和专利数据[30]等.

多中心网络 作为星型模式的扩展,一些网络针对不同节点子集拥有不同的中心节点,其结构更为复杂.这种网络广泛存在于生物信息学数据中[31, 32],如图 3(d)所示,包括两个中心:基因和化合物.此外,豆瓣数据集[18]也是一个典型的例子.

属性网络 为充分利用异质网络融合额外信息的能力,属性网络[18,19,33]进一步整合了节点和边的属性信息,将拓扑结构和属性特征结合,从而更好地建模实际场景,显著提升语义挖掘性能.一个具体的例子是互联网金融中的套现用户检测[19],通过将支付宝的交互场景建模为异质网络来辨别有信用欺诈嫌疑的用户,其网络模式如图 3(e)所示.

无模式/丰富模式网络 除以上广泛使用的异质网络外,许多实际交互系统中的节点和链接类型繁多且复杂,从而不能被建模为简单网络模式的异质网络.以知识库建模而成的知识图谱[34]为例,它基于资源描述框架(RDF)数据[35],符合<头实体,关系,尾实体>模型.因此从拓扑结构来看,知识图谱可以被视为异质网络.如图 3(f)所示的 YAGO[36],包含超 1000 万个实体(或节点),以及超 1.2 亿个实体间的链接,图中仅展示了部分网络结构.

除了以结构特点为原则的主流分类法,还可以基于构建网络的数据类型对异质网络进行分类[6].基于结构化数据的异质网络,即通过数据库中实体-关系模型组织的结构化数据构建异质网络[2];基于半结构化数据的异质网络,例如将 XML 格式数据中的属性视为节点,属性间连接视为关系;基于非结构化数据的异质网络,如利用实体识别和关系抽取等技术,抽取文本或图像数据中的点边关系构建异质网络[157,159].

与传统同质网络相比,异质网络通过引入节点和关系类型,提供了新颖的模式发现手段.从单类型节点的多关系网络、两种类型节点的二分网络、存在中心节点的星型及多中心网络,到包含丰富属性和模式的属性及无模式网络,异质网络的研究者在逐渐探索更复杂的网络模式以模拟真实世界.然而,目前对于异质网络的分析和挖掘仍存在不少挑战.首先,某些真实数据过于复杂,无法建模为有意义的异质网络.例如,虽然可以将 RDF 数据视为异质网络,但无法简单地描述其网络模式,因此难以基于元路径等方法进行语义探索.其次,即使可以将某些网络化数据建模为异质网络,其分析挖掘仍然受限于目前的语义探索方法.这些挑战也是异质网络亟需解决的问题.

1.4 网络模型比较

随着网络分析技术的蓬勃发展,研究者提出了许多网络数据建模的概念.这些概念具有相似含义,但也有细微差异.例如,多关系网络[37]是异质网络,多视图网络[38]也可以看作异质网络.在本节中,我们将异质网络与那些最相关的网络建模方法进行比较.

同质网络仅有一种类型的对象和关系,而异质网络包括不同类型的对象或关系.因而,同质网络可以视为异质网络的特例.此外,异质网络可以通过网络映射或忽略对象异质性转换为同质网络,但会产生信息损失.传统的关系挖掘[3,4]通常基于同质网络,但由于对象和关系的异质性,许多同质网络分析技术不能直接应用于异质网络.异质网络融合了丰富信息和语义,为精细的知识发现提供了可能,但也存在一定的局限性.首先,异质网络的复杂结构使得数据处理和语义挖掘过程更为困难.其次,与丰富的同质网络分析技术相比,异质网络分析方法主要是采用元路径,因此在取得性能提升的同时也面临着元路径所带来的瓶颈.

多关系/多维/多模网络[39,40]只有一种类型的对象,但对象间关系不止一种;复合网络[25,41]中的用户具有各种关系,并在每个单独网络或子网中表现出不同的行为,实际上也属于多关系网络.而异质网络包括不同类型的节点或关系,因此以上网络均是异质网络的特例.

复杂网络是具有非平凡拓扑特征的网络,其元素间的连接模式既不是完全规则的,也不是完全随机的[42].这

种非平凡的拓扑特征包括长尾度分布、高聚类系数、社团结构和分层结构等.许多交互系统都是复杂网络,如社交网络、信息网络和生物网络[43]等.复杂网络的研究涉及众多学科的知识和理论基础,尤其是系统科学、统计物理、数学和信息科学等,重点关注结构、功能和特征.虽然许多实际中的异质网络都是复杂网络,但目前的异质网络分析集中于模式较简单、规模较小的网络,且往往聚焦关系和语义挖掘.

近年来,知识图谱引起了研究者的广泛关注,其本质是语义网络的知识库.例如,Freebase[44]、DBpedia[45]、YAGO[36]和 NELL[46]构建的知识图谱已成功应用到语义解析[47]、实体消歧[48]、信息抽取[49]和问题解答[50]等领域.具体地,知识图谱是由实体和关系组成的复杂图结构,每个边都以<头实体,关系,尾实体>的三元组表示,也称为事实.这种三元组表明两个实体通过特定关系相连,如<猫,属于,哺乳动物>.由于知识图谱包含很多类型的实体和复杂的交互关系,难以用简单的网络模式来描述,因此可以将其视为一种无模式异质网络[36].除了拓扑结构的复杂性,知识图谱与传统异质网络相比,还增加了两个主要特性:其节点为实体,往往需要考虑本体层的语义;其三元组上附加了一些限制,如个数限制、存在限制等,更适宜于逻辑推理.

知识图谱和异质网络分别是自然语言处理和社交网络分析的热门方向,但两个方向的研究相对独立,研究对象和应用目标都不同.具体地,异质网络的研究对象主要是拓扑结构,而知识图谱还需要考虑本体、逻辑和规则限制等.异质网络作为同质网络的扩展,其研究任务与传统网络挖掘类似,例如相似性度量、社区发现和节点分聚类等.而知识图谱的挖掘往往聚焦于实体层面、本体层面或逻辑层面,其应用包含关系抽取、网络构建和逻辑推理等.由于两者从本质上而言均属于异质网络范畴,目前也出现了融合的趋势,具体工作将于第 3 章简要介绍.

2 基于元路径的数据挖掘

异质网络建模的优势在于可以整合更多信息,但同时也会形成新的难点——如何有效地利用异质信息并探索丰富语义.作为语义挖掘的有效工具,元路径可以指定对象连接序列并捕捉目标语义,因此已广泛应用于异质网络分析中的各类数据挖掘问题.文献[6,51,52,160]比较全面地介绍了基于元路径的数据挖掘方法,本章将着重介绍近三年的相关研究进展.

2.1 相似性度量

相似性度量用于评估对象的相似性,是许多数据挖掘任务的基础,如 Web 搜索和聚类等.关于相似性度量的研究已有较长历史,这些研究方法可大致分为两类:基于特征和基于链接.前者利用对象特征来度量相似性,如计算余弦相似性和欧几里德距离等.后者基于图中对象的链接结构来度量相似性,如 Personalized PageRank[53].最近,许多研究者开始关注异质网络中的相似性度量问题[54].

Table 1 The top-10 authors most similar to "Christos Faloutsos" under different meta-paths[8]

表 1 不同元路径下，与“Christos Faloutsos”最相似的前十名作者[8]

排名	作者	
	APA(共同作者)	APCPA(在同一会议上发表论文)
1	Christos Faloutsos	Christos Faloutsos
2	Spiros Papadimitriou	Jiawei Han
3	Jimeng Sun	Rakesh Agrawal
4	Jia-Yu Pan	Jian Pei
5	Agma J. M. Traina	Charu C. Aggarwal

与同质网络上的相似性度量不同,在异质网络上衡量对象间的结构相似性时,需要考虑连接两对象的元路径种类.因为不同元路径包含的语义不同,基于不同语义可能产生不同的相似性结果.例如,在文献异质网络中基于不同元路径寻找与“Christos Faloutsos”最相似的作者,结果如表 1 所示:*APA*路径下会找到他的学生,如“Spiros Papadimitriou”和“Jimeng Sun”,而*APCPA*路径下找到知名的研究者,如“Jiawei Han”和“Rakesh Agrawal”.因此,异质网络的相似性度量往往受元路径约束.考虑不同元路径所包含的语义,Sun 等人首先提出 PathSim 方法[8]来评估基于对称路径的相同类型对等对象间的相似性.基于 PathSim,一些研究者[55, 56]进一步整合动态时序

信息和节点属性等提出其扩展版本.在信息检索领域,Lao 和 Cohen 等人[57,58]提出路径约束的随机游走模型PCRW,用于度量文献异质网络中的实体相似性.但是,以上模型均针对相同类型对象间的相似性评估.为评估不同类型对象间的相似性,Shi 等人[21]提出 HeteSim 用于度量任意元路径下任意对象对的相似性,PLPIHS[59]则进一步基于 HeteSim 预测 lncRNA-蛋白质的相互作用.作为 HeteSim 的改进版本,LSH-HeteSim[60]用于挖掘异质生物网络中的药物靶标相互作用.

然而,基于元路径的相似性度量方法存在三点缺陷:

- 元路径仅适用于计算两个邻近(连接)实体的相似性.因而,Wang 等人[61]提出远程元路径相似性,捕获两个远程(隔离)实体间的语义;Liu 等人[62]提出邻近嵌入的概念,将网络结构嵌入相距较远的节点间.
- 元路径的选择往往依赖于领域知识.为此,KnowSim[63]提出无监督的元路径选择方法,并基于所选元路径集合度量相似性;Yang 等人[64]将强化学习和深度学习结合至半监督联合学习框架中,用于探索相似节点对间的有用路径.
- 元路径可以较有效地捕获源对象和目标对象间的单一关系,但往往无法衡量实际问题中的复杂相似性.因此,Huang 和 Fang 等人[9,65]提出基于元结构\元图的相似性度量方法;D2AGE[66]应用有向无环图建模节点间连接,并在进行相似性度量时考虑距离影响;IPE[67]考虑路径间的丰富交互,提出交互路径概念建模路径的依赖性;SPE[68]引入子图增强路径的概念,同时利用路径的距离感知和子图的高阶结构.

进一步,很多工作整合网络结构和额外信息,用于度量异质网络中对象间的相似性.结合影响力和相似性信息,Wang 等人[69]同时度量异质网络中的社交影响力和对象相似性;Wang 等人[70]结合异质网络中在线目标的上下文情境度量相似性;Zhang 等人[71]根据属性相似性和中心间连接来计算星型网络中心间的相似性.

2.2 推荐

推荐系统帮助消费者搜寻可能感兴趣的物品,如书籍、电影和餐馆等,往往基于信息检索、统计和机器学习的各种技术计算物品和用户偏好间的相似性.传统的推荐仅利用用户-物品评分反馈信息[72].随着社交媒体的普及,越来越多的研究者利用用户的社交关系[73, 74]研究社交推荐系统.

最近,一些研究者开始意识到异质信息对于推荐的重要性——异质网络全面的信息和丰富的语义使其有望产生更好的推荐结果.以图 2 电影推荐系统[18]构建的异质网络为例,该网络不仅包含不同类型的对象(如用户和电影),而且还描述了对象间的各种关系,如观影记录、社交关系和属性信息等.基于UMU路径寻找相似用户,将倾向于推荐与目标用户具有相同观影记录的用户看过的电影,这本质上对应协同过滤模型.同样,通过UGU路径可以找到兴趣相似的用户,这对应于成员推荐,其他元路径所对应的推荐模型如表 2 所示.因而,合理设置元路径可以实现不同的推荐模型.由此,Shi 等人[28]实现基于元路径语义的推荐系统 HeteRecom 来评估电影间的相似性;考虑属性值,如链接上的评分,他们进一步将推荐系统建模为加权异质网络,并提出基于加权元路径的个性化推荐方法 SemRec[18].

Table 2 The meanings of meta-paths and their corresponding recommendation models

表 2 元路径含义及其相应推荐模型

元路径	语义含义	推荐模型
UU	目标用户的好友	社交推荐
UGU	与目标用户同一兴趣组的用户	成员推荐
UMU	与目标用户有相同观影记录的用户	协同推荐
UMAMU	与目标用户所看电影有相同演员参演的用户	内容推荐

近些年,随着网络表示学习的兴起,越来越多的异质网络推荐方法利用异质网络表示学习技术学习用户和物品的特征表示用于推荐.例如,HERec[12]基于元路径的随机游走生成节点序列以学习节点的嵌入表示,并将其集成至矩阵分解框架用于商品推荐;HueRec[75]假设用户或物品在不同元路径下有共同特征,从而利用所有元路径学习统一的用户和物品表示;LGRec[76]融合用户-物品直接交互信息和基于元路径的广义交互信息,利用共同注意力机制实现 Top-N 推荐;NeuACF[77]利用深度神经网络学习用户和物品不同方面的潜在特征,并以注意力机

制融合得到最终表示.伴随图神经网络[164,165]的兴起,异质图神经网络在推荐任务上展现出优越性能.例如,PGCN[78]在三个子图上利用池化和卷积聚合邻居特征以学习用户和物品表示;Fan 等人[14]提出基于元路径的异质图神经网络,用于学习意图推荐中的节点表征;MCCF[79]初步探索用户购买动机,提出多成分图神经网络来基于不同动机分别聚合商品信息实现更细粒度的用户偏好建模.此外,部分工作通过权重区分不同节点及关系,从而考虑推荐系统中不同节点和关系影响力的差异.Nandanwar 等人[80]使用顶点增强的随机游走,避免多个彼此相邻且影响力大的节点对推荐多样性的影响;HeteLearn[81]基于贝叶斯个性化排名技术学习链接权重,实现对用户偏好的个性化建模.针对元路径只能捕捉简单线性关系的局限性,一些工作尝试利用元图等复杂语义捕捉工具精确建模用户偏好.具体地,Zhao 等人[10]将元图的概念引入推荐来刻画复杂语义,并利用“矩阵分解+因子分解机”框架进行信息融合;MoHINRec[82]提出模体增强的元路径,进一步捕获相同类型节点间的高阶关系.可以看到,以上方法主要基于用户和物品属性及其之间的交互关系构建异质网络,并利用元路径等工具捕捉高阶关系以缓解数据稀疏,从而得到更精确的节点表征用于推荐.

此外,许多方法利用异质网络融合除用户-物品交互外的信息辅助推荐,典型代表是基于用户好友信息的社会化推荐:Zheng 等人[83]设计双重相似性正则,在预测用户行为时同时对具有高低相似度的好友施加约束;Yu 等人[84]提出自适应识别隐式朋友的算法,将相似用户作为隐式朋友来减轻不可靠社交关系对推荐的不利影响;Wen 等人[85]在社交网络上预训练嵌入模型,并将其融合入传统矩阵分解框架预测用户购买行为.此外,考虑交易的时序信息,MANN[86]利用外部存储矩阵存储和更新用户历史记录,从而挖掘用户行为受先前行为影响的直观模式.进一步考虑推荐系统中的多模态信息,Zhang 等人[87]基于深度表示学习架构,分别学习每种类型信息源(如评论文本、物品图片和数字评分等)相应的用户和物品表示.基于异质网络的推荐系统,并不局限于传统的商品推荐,Yu 等人[88]利用异质网络和贪心算法,实现兴趣点组推荐;DKN[89]将知识图谱结合至新闻推荐中,用于点击率预测等.

2.3 其他任务

分类是一种基本的数据分析任务,可以通过构建模型或分类器来预测类标签.传统机器学习的分类任务主要针对满足独立同分布的相同类型对象.与传统的分类不同,异质网络研究的分类问题具有一些新的特点:(1)异质网络中包含的对象是不同类型的,这意味着标签是通过不同类型对象间的各种链接传播的;(2)异质网络中的非目标类型对象也可能提供对目标类型对象分类有益的信息.而元路径作为对象间链接的元模式描述,可以反映给定语义并刻画高阶结构,因此被广泛用于异质网络的分类任务中.例如,HetPathMine[90]设计新颖的元路径选择模型,使得标签在传播时基于不同的路径;Wang 等人[91]提出基于元路径的核方法,以不同路径语义辅助文本分类;HeteClass[92]基于正则化加权组合元路径,从而抽取同质网络进行转导分类.最近,协同分类引起了部分研究者的关注.Kong 等人[22]利用对象间基于元路径的依赖关系进行协同分类;GraphInception[93]作为一种深度卷积协同分类方法,可以自动生成不同复杂度的关系层次结构.

聚类是将数据对象划分为一组簇的过程,簇中对象彼此相似,不同簇中对象彼此不同.基于网络化数据的聚类方法通常将数据建模为同质网络,并使用给定度量(如标准化切割[94]和模块度[95]等)将网络划分为一系列子图.最近,异质网络的聚类引起了广泛关注.与同质网络相比,(1)异质网络中共存的多类型对象使得传统的聚类方法无法直接应用,因此许多工作将传统谱聚类方法扩展到包含不同类型对象的异质网络中.例如,SClump[96]基于元路径构建相似性矩阵进行频谱聚类,Dall'Amico 等人[97]基于 Bethe-Hessian 矩阵研究异质网络中的谱聚类.(2)异质网络中包含的丰富语义使得聚类过程利用额外信息更加方便.例如,属性信息被广泛应用在异质网络的聚类分析中:Cruz 等人[98]利用结构维度和复合维度构建属性图解决社团检测问题;SCHAIN 算法[99]同时考虑对象在属性值及结构连通性方面的相似性;Chen 等人[100]将属性建模为不同类型的节点,以同时捕获结构和属性相似性.一般地,聚类是一项独立的数据挖掘任务,但它也可以与其他数据挖掘任务相结合.例如,基于排名的异质网络聚类方法表明聚类和排名可以相互增益:RankClus 方法[101]针对二分网络同时优化聚类和排名性能,为特定类型对象生成簇;ComClus 方法[32]利用具有自环的星型网络结合异质和同质信息,实现基于排名的聚类.

链接预测是链接挖掘中的基本问题,即基于观测链接和节点属性来估计两节点间存在链接的可能性.链接

预测通常被视为简单的二分类问题:对于任何两个可能连接的对象,预测链接存在或不存在.与同质网络不同,异质网络中的待预测链接可能具有不同的类型[161].因此,许多工作基于元路径刻画不同类型链接间的复杂关系,从而实现异质网络中多类型链接的预测问题.具体地,可以采用两步框架解决异质网络中的链接预测:第一步,提取基于元路径的特征;第二步,训练回归或分类模型计算链接的存在概率[102,103,104,105,106].例如,Sun 等人[102]提出PathPredict 方法,利用元路径提取特征并训练逻辑回归模型进行共同作者预测;Cao 等人[106]设计相关性度量方法来构建链接特征,并提出迭代框架同时预测多类型的链接;PME[107]结合度量学习和异质网络嵌入,基于一阶和二阶邻近度提升预测准确性.针对社交网络中的冷启动问题,Liu 等人[108]利用额外的社交网络信息,提出对齐因子图模型进行用户-用户链接预测;SHINE[109]利用多个深度自编码器将用户映射到低维特征空间,并预测社交网络中的用户情感关系.为摆脱需要领域知识预定义元路径的困境,LiPaP[110]设计自动元路径生成算法,用于模式丰富异质网络中的链接预测.除静态网络外,动态链接预测也非常重要且富有挑战性.Zhao 等人[111]提出一个通用框架,用于从异质网络数据演变中表征和预测社团成员;Aggarwal 等人[112,113]提出两级方案来有效进行宏观和微观决策,并组合拓扑和类型信息;SLIDE[114]维护并更新低秩矩阵以描述所有观察到的数据,并基于该矩阵动态推断链接.

2.4 元路径选择

异质信息网络分析中,大多数方法采用元路径进行特征和子结构抽取.这些方法往往假设存在一组给定的或可枚举的元路径,然后利用它们来计算相似性或网络嵌入.尽管这些方法都展现出了很好的性能,但它们仍然面临元路径选择困境:(1)元路径的选择很大程度上依赖于领域知识.对于不熟悉或很复杂的异质网络,难以依靠领域知识选择合适的元路径集合.并且,随着元路径长度的增加,路径数量呈指数增长,使得路径搜索过程非常昂贵.(2)简单拼接各种元路径的信息反而会引入噪音,影响性能表现.而为各元路径学习合适的权重,又常常需要监督信息.

目前解决该困境的方法主要分为两类,一类是自动生成元路径,另一类则不利用元路径进行数据挖掘:

- 自动生成元路径的算法往往基于网络模式搜寻可能连接实例对的元路径集合.具体地,针对简单模式异质网络的元路径生成问题,KnowSim[63]使用类似 Personalized PageRank 的算法搜寻子图,基于子图自动生成元路径;Yang 等人[64]将强化学习和网络嵌入集成到半监督联合学习框架中,在结构和内容信息的引导下自动发现有用路径.针对模式丰富的知识图谱中元路径的自动生成,RelSim[115]根据用户提供的简单关系实例,自动匹配元路径;Zheng 等人[116]设计元路径自动生成方法 SMPG,利用实体间的潜在关系生成并学习元路径的权重;LiPaP[110]自动提取元路径,并基于似然函数赋予元路径权重.
- 一些工作不采用元路径,而是以关系为出发点进行数据挖掘.例如,DBSCAN[98]将节点及其属性信息建模为异质网络,并提出迭代更新策略以权衡不同关系对于节点聚类的重要性;PME[107]将节点从对象空间投影到相应关系空间,并以统一方式捕获一阶和二阶邻近度用于链路预测;HeteLearn[81]基于带重启的随机游走和贝叶斯个性化排名技术,针对用户学习不同的关系权重以实现个性化推荐.此外,还有部分异质网络嵌入算法基于关系学习节点表示用于下游任务.RHINE[118]依据结构将关系分为两类,分别设计模型学习节点嵌入;NSHE[166]基于网络模式进行异质网络嵌入,从而同时保留节点对和高阶信息;基于图神经网络,HetGNN[120]利用随机游走采样固定数量的异质邻居,并依据邻居类型分组聚合信息.

3 异质网络的表示学习

早期处理网络化数据的工作大部分基于高维稀疏向量进行矩阵分析.然而,现实中网络的稀疏性及其不断增长的规模,对此类方法产生了严峻挑战.一种更有效的方式是将网络节点映射到低维向量空间中,用低维稠密向量来表示网络中的任意节点,从而更灵活地应用于不同数据挖掘任务中,即信息网络的表示学习[121].现在已有大量工作致力于同质网络的表示学习.这些工作大多基于已有的深度模型并结合网络特征,学习节点或边的特征表示.代表性模型如 DeepWalk[122],将随机游走和 skip-gram 模型结合来学习网络节点表示;LINE[123]在一阶邻居相似性的基础上加上二阶相似性,从而学习对大规模稀疏网络的强区分节点表示;SDNE[124]借助深度自动编

码器来抽取网络结构的非线性特征.除使用网络的结构信息,也有很多方法进一步利用节点的内容或其他辅助信息(如文本、图像和标签等)学习更准确更有意义的节点表示.一些综述论文全面地总结了这方面的工作[125,126].

Fig.4 Heterogeneous information network representation learning dendrogram

图 4 异质网络表示学习树状分类图

由于异质网络的特殊性,同质网络的表示学习方法并不能直接应用于异质网络,主要存在两点挑战:

- 节点和边的异质性.不同类型的节点和边代表不同的语义,因此异质网络的表示学习需要将不同类型的对象映射到不同的空间中.此外,如何保存每个节点的异质邻居及如何处理异质的节点序列也是值得探究的问题.
- 异质网络中丰富信息所带来的表示融合.异质网络从多个维度刻画节点的语义,如何有效抽取和利用多维度信息并融合得到全面的节点表示也是巨大的挑战.

异质网络表示学习兴起于最近两三年,但是发展迅猛.如图 4 所示,将已有的方法大致按照浅层模型和深层模型分类,并简要介绍研究进展.

3.1 浅层模型

为应对网络异质性带来的挑战,部分浅层模型将其分解为较简单的网络,分别对这些网络进行表示学习,然后再将信息融合起来达到“分而治之”的效果.例如,HERec[12]利用元路径抽取异质网络中的多个同质网络,对这些同质网络进行表示学习并融合;尹赢等人[162]从异质网络中抽取带权同质子图,并基于带偏置的随机游走得到同类节点序列用于节点表示学习;PTE[127]将从文本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,并分别学习节点向量表示;EOE[128]将学术异质网分解为单词共现网络和作者合作网络,对各个子网络内节点对和子网络间节点对同时进行表示学习.上述这种先拆解再融合的两步框架作为同质网络向异质网络的过渡产物,常运用于异质网络表示学习的早期工作中.后来,研究者逐渐意识到从异质网络抽取同质子图的过程中,会不可逆地损失异质邻居所带有的信息,并开始探索真正适配于异质结构的表示学习方法.

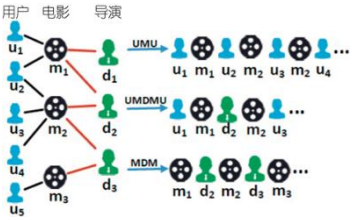


Fig.5 Meta-path based random walk in movie heterogeneous information networks

图 5 电影异质网络中基于元路径的随机游走

首先出现了许多基于随机游走的方法,以更好地刻画异质网络中的丰富语义.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,因此也被广泛用于网络表示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,从而使得采样所得节点序列不仅包含结构信息还蕴含相应语义.图 5 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和丰富语义.例如,Metapath2vec[129]基于元路径的随机游走抽取节点结构信息,并利用 skip-gram 算法学习节点表示;HIN2Vec[130]同时考虑不同类型节点及其之间复杂多样的关系,将网络嵌入转化为多分类问题学习节点及元路径的表示;HINE[131]基于元路径的随机游走计算节点间的相似性,并将其作为监督信息指导节点嵌入;ESim[132]通过使元路径实例的概率最大化学习节点的向量表示;Wang 等人[133]探索了异质网络中的幂律分布等特性,进而基于元路径进行随机游走并在双曲空间中嵌入网络.

由于元路径是一种较为简单的线性语义捕捉工具,部分工作改进了随机游走的方式来更全面地描述语义.Jiang 等人[134]提出半监督学习算法,将网络模式分解为元图后进行随机游走;He 等人[135]提出异质的个性化空间随机游走,分别学习由元路径、元图和元模式引导的节点嵌入表示;Hussein 等人[136]提出 jump 和 stay 结合的随机游走策略,打破需要预先定义元路径的约束.

此外,有些方法从异质网络中不同的关系类型及其特点出发,在规避元路径选择的同时学习节点嵌入表示.RHINE[118]将异质网络中的关系分为隶属和交互两类,并分别以不同方式建模;BHIN2vec[117]考虑网络中所有关系类型间的平衡问题,根据相对训练率生成不同训练样本;Li 等人[137]提出新颖的表示学习模型,进一步考虑三角形和平行四边形连接结构;PME[107]基于度量学习捕获一二阶邻近度,并提出自适应的损耗感知采样方法用于模型优化.

3.2 深层模型

近些年,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的数据进行建模.相对于浅层模型,深度模型可以更好地捕捉非线性关系,从而抽取节点所蕴含的复杂语义信息.我们将深层模型大致分为四类:基于自动编码器、基于生成对抗网络、基于强化学习和基于图神经网络的方法,并介绍其代表性工作.

基于自动编码器的模型,旨在利用神经网络构建编码器学习节点属性表示的同时保持网络结构特性.例如,BL-MNE[138]分别对不同元路径下的信息进行编码,然后再综合信息进行联合编码;SHINE[109]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,并通过聚合函数进行融合.基于生成对抗网络方法的核心是利用生成器和鉴别器间的博弈得到鲁棒节点表示.在同质网络中,基于对抗思想的方法往往只考虑结构信息,如 GraphGAN[170]基于广度优先搜索生成虚拟节点.而异质网络中的生成对抗模型需要生成器和鉴别器考虑关系异质性以捕获语义信息.作为第一个在异质网络中应用对抗思想的模型,HeGAN[119]训练具有关系感知能力的鉴别器和生成器,并通过学习节点的潜在分布改进负采样;MV-ACM[171]则利用生成对抗网络计算节点在不同视图中的相似性,从而综合捕获多视图信息.基于深度强化学习的深层模型则主要针对元路径选择困境,将下游任务性能作为奖励来优化整体框架,从而在规避元路径选择的同时学得节点表示.例如,Qu 等人[139]将星型网络的节点表示学习转化为马尔可夫决策过程,其动作是选择用于学习或终止训练的特定类型链接,状态是已选择的链接类型顺序;MPDRL[169]基于任务准确度发现长度不等的语义丰富元路径,并基于该元路径集合进行节点表示学习.

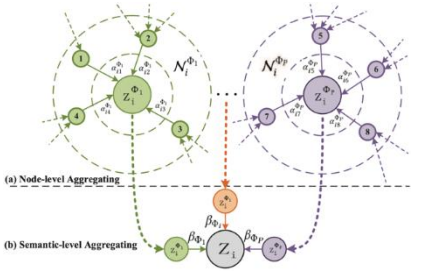


Fig.6 Hierarchical attention mechanism in HAN[13]

图 6 HAN 中的层次注意力机制[13]

近几年图神经网络兴起,其核心思想是消息传递机制,即将邻域信息聚合并作为消息传递给邻居节点.图神经网络在同质图表示学习上所展现出的强劲性能,促使越来越多的研究者探索异质图神经网络.与同质图神经网络类似,异质图神经网络的关键在于如何设计合适的聚合函数以捕获邻域所包含的语义.具体地,HAN[13]提出基于层次注意力机制的异质图神经网络.如图 6 所示,首先基于节点级别注意力机制聚合邻居信息,然后利用语义级别注意力机制聚合元路径信息,从而同时考虑基于元路径的邻居之间和元路径之间的重要性.此外,HAHE[140]基于分层注意力机制学习每个节点的个性化元路径权重;HetGNN[120]基于随机游走采样固定大小的异质邻居,并根据节点类型对其进行分组编码,最终通过组间信息融合得到节点表示;ActiveHNE[141]提出基于图卷积神经网络的半监督嵌入方法,并依据不确定性和代表性采取不同的主动选择策略以充分利用监督信息;NetCycle+[142]将图卷积神经网络扩展至动态异质网络中,利用网络节点间的深层依赖性进行推理.

3.3 浅层和深层模型比较

基于以上浅层和深层模型的代表性工作概述,可以发现浅层模型主要针对异质网络的结构,而很少利用属性等额外信息.可能的原因之一是浅层模型不易描述额外信息与结构信息的关系,从而使得同时建模两者较困难.而深层模型的学习能力更为支持这种复杂的建模方式,例如图神经网络在传播过程中自然地整合了网络的结构和属性信息,使其在复杂应用场景中取得更大的增益.

然而,浅层和深层模型各有自己的优缺点.浅层模型缺乏非线性表示能力,但高效且易于并行.深层模型有更强的表示能力,但也容易拟合噪音且时空复杂度较高,与此同时,深层模型繁琐的超参调整过程也为人所诟病.因此,浅层和深层模型并没有绝对的优劣.根据具体应用场景,灵活选择模型即可.

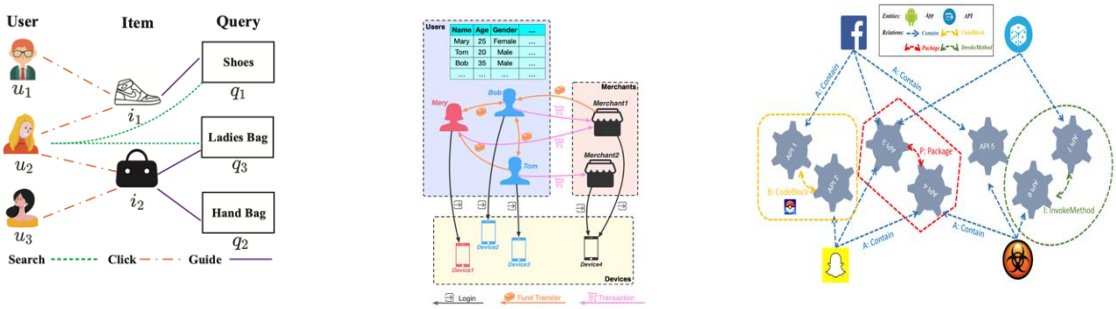
3.4 与知识图谱表示学习的区别与联系

知识图谱以图的形式表现客观世界中的常识和事实,已经成为学术界和工业界广泛使用的知识表示方式.而将知识图谱的表示学习与推理技术结合,可以给人工智能系统提供可处理的先验知识,使其具有像人类一样的解决复杂任务的能力.具体地,知识图谱表示学习为知识图谱中的实体和关系学习包含语义信息的低维向量表示,从而使得下游任务可以方便地提取和利用知识图谱中的信息.传统异质网络表示学习多以拓扑结构为出发点,而知识图谱极为丰富的节点和链接类型使其难以直接应用元路径等传统异质网络挖掘方法.因此,目前知识图谱表示学习的主流方法是基于三元组的.其中,最经典的是 Trans 系列模型.以 TransE[167]为例,其提出头实体的向量表示加上关系的向量表示应当等于尾实体的向量表示.基于这种约束,TransE 可以同时学得实体和关系代表的丰富语义.更多的知识图谱表示学习工作,建议读者参阅相关综述[168].

由于知识图谱作为无模式异质网络,本质上属于异质网络范畴,目前有一些工作尝试将二者结合.例如,MPDRL[169]在知识图谱上基于强化学习发现长度不等的语义丰富元路径,从而使得元路径这一传统异质网络语义挖掘工具迁移至知识图谱中;RHINE[118]借鉴知识图谱的表示学习技术,基于相连节点度差异将异质网络中的关系分为隶属和交互两类,并分别利用欧几里得距离约束和 Trans 模型进行学习表示.虽然已有了初步探索,但两者相融仍有相当大的挑战.首先,为了平衡效率和效果,知识图谱表示学习方法往往忽略网络本身的复杂结构.这一点与传统异质网络表示学习的关注点相违背,因此将简单模式异质网络表示学习的方法迁移至复杂模式的知识图谱是二者融合中的最大挑战.其次,知识图谱存在本体数据模型,同时基于本体和规则的推理方法经常出现.而传统异质网络表示学习如何将上述推理方法在语义层面予以刻画也是难点问题之一.

4 应用

除上面讨论的基础数据挖掘任务外,异质网络在商业、安全和医学等领域有许多实际的应用场景.



(a)电子商务异质网络[14] (b)信用支付异质网络[19] (c)HinDroid 异质网络[15]

Fig.7 Heterogeneous information networks in application scenarios

图 7 实际应用场景中的异质网络

Table 3 Results of cash-out user detection w.r.t. AUC

表 3 套现用户检测的 AUC 性能[19]

算法	十天数据集	一个月数据集
Node2vec	0.5893	0.5980
Metapath2vec	0.5914	0.6005
Node2vec+Feature	0.6455	0.6541
Metapath2vec+Feature	0.6456	0.6550
Structure2vec	0.6537	0.6641
GBDT	0.6389	0.6467
$GBDT_{Struct}$	0.6948	0.6968
HACUD	**0.7066**	**0.7132**

在商业领域,存在大量的关联数据,因而可以构建异质网络.与以往的网络分析相比,利用异质网络建模可以涵盖多类型节点及其之间的交互,整合丰富甚至异构的信息源,从而更全面地刻画节点特征.其中,最广泛的应用之一是推荐,传统的推荐算法基于协同过滤,仅考虑用户和物品的交互信息.然而,实际的推荐系统中用户与商品、店铺和朋友等存在大量的交互,利用异质网络建模这些信息,能够更精确地形成用户画像来提升推荐性能.具体地,MEIRec[14]将电子商务平台中的用户、物品和查询建模为图 7(a)所示的异质网络,并提出基于元路径的异质图神经网络学习意图推荐中的用户表示;Yu 等人[88]建模基于位置的社交网络,并利用贪心算法进行兴趣点组推荐;HIE[143]通过建模文字和响应模式等信息,预测用户的个性特征;2019 年 CIKM 最佳应用论文[144]将闲鱼 APP 中的用户、商品和评论建模为二分网络,并基于图神经网络进行垃圾评论过滤,从而减少其对用户选择的影响.另一个典型例子是套现用户检测.信用支付场景中,套现用户具有极高的贷款违约率,倘若不管检测此种用户可能引起金融风险.常规方法基于统计特征训练分类器,很少充分利用用户间的交互信息,而实际信用支付中的丰富交互关系从多方面反映了套现用户的特征.为此,HACUD[19]利用属性异质网络建模蚂蚁金服信用支付中用户与商家间的交互,如图 7(b)所示,并基于元路径和层次注意力机制学习用户表示,其性能如表 3 所示.可以看到,相较于原始版本的 Node2vec、Metapath2vec 和 GBDT,综合利用节点和属性信息后的增强版本性能均有所提升.而 HACUD 的优越性能表明基于属性异质网络建模,并利用层次注意力机制更好地融合了场景中的有效结构和属性信息.此外,像 DiDi 和 Uber 这样的在线出租车平台已经影响了数亿用户的旅行选择,那么如何提升用户乘车体验并预测用户行程成为平台所关注的问题之一.基于用户旅行数据,PHINE[145]用异质网络建模驾驶员、乘客和位置等信息,预测用户的乘车满意度;TDP[146]利用行程的起止点等信息,基于异质网络嵌入和深度神经网络预测用户可能的短期出行.

Table 4 Comparisons between HinDroid and alternative detection methods

表 4 HinDroid 与其他检测模型的性能比较[15]

模型	F1	AUC	ACC	TP	FP	TN	FN
ANN	0.9409	0.9316	0.9300	279	12	186	23
NB	0.9025	0.8891	0.8860	264	19	179	38
DT	0.9539	0.9397	0.9440	290	16	182	12
SVM	0.9590	0.9537	0.9520	281	7	191	17
HinDroid	**0.9884**	**0.9849**	**0.9860**	**299**	**4**	**194**	**3**

异质网络也被广泛用于建模网络安全问题中的复杂交互系统,目前的工作主要针对恶意软件检测、恶意账户检测和异常事件发现这三类问题.(1)对于恶意软件,传统的方法主要基于签名进行识别,而黑客可以轻松利用代码混淆或重包装等技术逃避检测.对于此种情况,2017 年 KDD 最佳应用论文不再仅利用 API 调用信息,而是提出 HinDroid[15]将 Android 应用程序、相关 API 及其丰富关系建模为异质网络,如图 7(c)所示,然后基于不同元路径度量 Android 应用的相似性,最终利用多核学习加权相似性来进行预测.其具体恶意软件检测性能如表 4 所示,相较于不利用交互关系的基线模型人工神经网络(ANN)、朴素贝叶斯(NB)、决策树(DT)和支持向量机(SVM),基于异质网络建模的 HinDroid 在各项指标上均有所提升.基于类似的异质网络构建方式,AiDroid[147]分类学习节点表示,并利用深度神经网络进行恶意软件检测;αCyber[148]提出对抗攻击模型 HG-Attack 和防御模型 Rad-HGC,两者相互增强以提升恶意软件检测的鲁棒性.(2)基于网络特有的虚拟性、超时空性、隐蔽性和信息快速流通性等,网络犯罪不断给社会平稳安定带来隐患和困扰.为检测恶意账户,Liu 等人[149]从支付宝异质子图中自适应学习嵌入表示,并利用注意力机制区分不同类型节点的重要性;为识别恐怖分子,OSNE[150]将犯罪和恐怖主义活动建模为异质网络,利用高阶关系路径进行嵌入学习;iDetector[151]利用异质网络建模地下论坛,基于不同元结构表征帖子间的相关性并进行有效融合;uStyle-uID[152]利用属性异质网络建模文本和照片等,并提出基于书写及摄影风格识别毒品贩子.(3)异常事件检测旨在发现不同模式或行为的事件.例如在电影推荐系统中,一个常评论动作电影的用户突然评论了情感类电影,挖掘此类异常的行为模式有利于深入分析用户的潜在兴趣.与以往利用属性信息和同质结构不同,Fan 等人[153]提出同时包含实体属性和二阶结构的深度异质网络嵌入方法,以提高异常检测的准确性;Ranjbar 等人[154]提出基于张量分解和聚类的异常检测方法,用户可以直接通过查询来检测异常实体.

医学领域同样也存在大量异质交互场景,可以利用异质网络建模,现有工作主要将其应用至疾病诊断和基因分析中.(1)电子健康记录(EHR)提供了患者的各种临床事件的详细记录,但 EHR 数据高维且稀疏,现有方法仅考虑不同患者记录中同时发生的临床事件以获取实体嵌入,无法捕获 EHR 包含的丰富结构和语义信息.因此,HeteroMed[155]使用异质网络对临床数据进行建模,利用元路径捕获有助于疾病诊断的重要语义.另一个典型例子是阿片类药物使用者检测.阿片类药物滥用成为世界最致命的流行病之一,许多用户愿意在 Twitter 上分享他们使用阿片类药物的经历.为检测阿片类药物使用者,HinOPU 框架[156]利用异质网络建模 Twitter 中用户和推文间的关系,并基于元图表征用户间的语义相关性进行使用者预测.(2)以往基因分析的方法仅关注于基因的本身特征,忽视了与其相关的生物网络拓扑结果中所包含的信息.因此,PLPIHS[59]使用异质网络建模 lncRNA-蛋白质网络,并基于 HeteSim 计算 lncRNA-蛋白质对的相关性评分来推断其相互作用;马毅等人[163]同样基于 HeteSim 计算致病基因间的相关性,并经生物实验证实了该方法在卵巢癌和胃癌预测分析中的有效性.

5 未来发展方向

由于异质网络能够建模复杂交互系统,融合丰富语义信息,近期出现了大量异质网络分析的工作.因而,本文对这一迅速发展的领域进行了综述.除异质网络的基础知识外,重点介绍了基于元路径的数据挖掘、异质网络的表示学习和实际应用三个方面的研究进展,并且特别关注于近三年的发展情况.虽然异质网络已经应用于许多数据挖掘任务和实际场景,但它仍然是一个年轻的、正在快速发展的研究方向,未来值得关注的方向如下.

面向多模态数据的异质网络构建与分析方法.异质网络可以通过融合丰富信息解决大数据的“多样性”挑战.现有工作主要致力于对关系数据库类的结构化数据建模,而文本、图像和多媒体等模态数据是否也可以采用异质网络建模与分析?目前已有部分工作开始将异质网络用于文本挖掘中,例如 HGAT[157]利用异质网络建模实体、文档和主题间的关系,并提出异质图注意力网络学习短文本表示用于文本分类;Hu[158]等人利用异质网络显式建模用户、新闻和潜在主题间的交互,并基于图神经网络和长短期记忆网络同时捕捉用户长短期兴趣,从而进行个性化新闻推荐.异质网络建模方式在图像领域也有了初步尝试. ReGAT[159]将图像中不同的物体及其之间的空间、语义等关系建模为异质网络,并基于图注意力机制学习关系表示用于视觉问答.由于结构关系同样广泛存在于多模态数据,因此面向多模态数据的异质网络构建与分析是值得研究的方向,但这里存在不少研究难题.首先,利用半结构化的异质网络建模多模态数据时,节点和边关系不明确且复杂.例如,视觉问答中既有图像数据又有文本数据,那么同一图像应如何抽取对象间的关系、不同图像间又如何通过关系相联系、图像特征又如何与文本特征相对应等,从多模态数据中提取合适的对象和关系是异质网络建模和分析的关键.其次,异质网络表示学习和多模态数据表示学习如何有机融合、能否探索出超越元路径的多模态数据语义挖掘及融合方法等问题也值得进一步研究.

面向复杂网络数据的异质网络分析方法.实际应用中的异质网络具有动态变化、规模巨大、模式丰富等特点,需要研究真实复杂网络数据的异质网络分析方法.具体地,(1)实际网络往往是动态异质的.例如,淘宝平台不断有新用户、新店铺、新商品等新节点产生,已有节点也在不断产生新交互,那么在对这些动态异质网络进行分析时,重训练耗时耗力,因而需要研究面向增量计算的异质网络分析方法,以求在网络结构和属性动态变化时可以动态更新预测结果.(2)实际网络是规模巨大的.淘宝包含亿级的用户和商品,响应速度会较大程度影响用户体验,因此在对大规模异质网络进行分析时往往对算法实时性有较高要求,需要研究面向快速计算的异质网络分析方法,比如合理采样子图或离线保存关键特征等.除优化上层算法,从底层架构入手结合硬件加速大图挖掘速度的图计算平台也是有前景的研究方向.(3)实际网络是模式丰富的.工业场景中构建的知识图谱是模式丰富的异质网络,难以描述其网络模式,因而利用传统语义挖掘工具——元路径时,就会涉及元路径的自动发现及约减问题.与此同时,能否探索出适宜于知识图谱的语义挖掘方法、能否将其与元路径等简单语义挖掘工具融合至统一框架中,也是未来的方向.(4)实际网络中的链接通常包含丰富信息.社交异质网络中的关注、转发等是有向的交互关系,这种有向性对于影响力分析极为重要;推荐异质网络中的打分、评论等是有属性值的交互关系,这种属性信息同样有利于推荐性能的提高.因此,扩展现有方法来充分利用链接上的信息,如加权元路径等考虑链接性质的语义挖掘方法,有望产生更准确的知识发现.

面向深度计算的异质网络表示学习.网络表示学习已成为当今热点,而图神经网络作为优美有效的表示学习算法,可以扩展至异质网络中.虽然已有部分工作提出了异质图神经网络,但与同质网络相比,仍有许多方面亟待研究.(1)异质图神经网络的内部机制.现有异质图神经网络方法的聚合方式往往分为两种:依据给定元路径聚合邻居信息;聚合直接邻居信息.那么元路径邻居与直接邻居的信息是否有内在关联,能否通过某种聚合机理统一?与此同时,不少工作探究缓解图神经网络过拟合的方法,那么在异质图神经网络中是否存在过拟合,又如何缓解?部分工作探索同质网络中局部和全局信息的关系,基于全局信息的指导使得局部信息相似的远距离节点也拥有相似的低维表示,那么在异质网络中局部信息和全局信息的关系又是怎样,它们之间又能否相互增益.(2)异质图神经网络的鲁棒性.最近研究表明,图神经网络易受人为设计扰动的攻击(即对抗攻击),因而图神经网络面对对抗攻击的脆弱性使得研究者愈发关注其在安全型关键应用中的防御问题.随着异质网络在安全领域的应用越来越广,如何找到异质网络表示学习中的薄弱环节,并设计相应防御机制以提升表示鲁棒性,从而减弱对下游任务的决策影响也是值得关注的问题之一.(3)异质网络表示学习的可解释性.由于异质网络融合了丰富信息,基于元路径等语义挖掘方法学得的节点和关系表示拥有更强的解释性,例如商品推荐可以依据元路径的注意力权重给出较明确的推荐理由,这种优越性在许多任务中还没有得到很好地分析和阐述,比如利用异质网络进行疾病诊断,倘若给出相应的诊断理由能否减小误诊的可能.(4)异质网络与知识的融合.推理能力是人工智能的核心,与知识融合进行推理可以使得基于小样本训练集得到的模型具有极强的泛化能力.因此,如何与知识融合

产生具有推理能力的异质网络,也是尚待探索的方向.

更多的实际应用 实际场景中往往存在大量交互和丰富信息,因而可以很自然地利用异质网络建模.目前异质网络研究已逐步与实际相结合,应用至电商、安全和医学等领域,并实质提升了一些挖掘任务的性能.这些工作为采用异质网络解决实际问题带来了启示,但是还有更多可利用异质网络建模的场景尚待发掘.例如,软件工程领域中需求单、问题单、测试样例等之间存在复杂的交互关系;基因工程中物种、基因序列、编码结构等也有不可分割的联系.因此,如何将异质网络分析落地,在更多具体应用中发挥作用是异质网络重要的发展方向.

References:

[1] Han J. Mining heterogeneous information networks by exploring the power of links[C]// In Proc. of ICDS, 2009: 13-30.

[2] Sun Y, Han J. Mining heterogeneous information networks: a structural analysis approach[J]. Acm Sigkdd Explorations Newsletter, 2013, 14(2): 20-28.

[3] Lichtenwalter R N, Lussier J T, Chawla N V. New perspectives and methods in link prediction[C]// In Proc. of KDD, 2010: 243-252.

[4] Leroy V, Cambazoglu B B, Bonchi F. Cold start link prediction[C]// In Proc. of KDD, 2010: 393-402.

[5] Zhang J, Philip S Y. Integrated anchor and social link predictions across social networks[C]// In Proc. of IJCAI, 2015.

[6] Shi C, Li Y, Zhang J, et al. A survey of heterogeneous information network analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 29(1): 17-37.

[7] Wang F, Qu Y, Zheng L, et al. Deep and broad learning on content-aware POI recommendation[C]// In Proc. of CIC, 2017: 369-378.

[8] Sun Y, Han J, Yan X, et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks[C]. In Proc. of VLDB, 2011, 4(11): 992-1003.

[9] Huang Z, Zheng Y, Cheng R, et al. Meta structure: Computing relevance in large heterogeneous information networks[C]// In Proc. of KDD, 2016: 1595-1604.

[10] Zhao H, Yao Q, Li J, et al. Meta-graph based recommendation fusion over heterogeneous information networks[C]// In Proc. of KDD, 2017: 635-644.

[11] Shi C, Philip S Y. Heterogeneous information network analysis and applications[M]. Springer International Publishing, 2017.

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

[13] Wang X, Ji H, Shi C, et al. Heterogeneous graph attention network[C]// In Proc. of WWW, 2019: 2022-2032.

[14] Fan S, Zhu J, Han X, et al. Metapath-guided Heterogeneous Graph Neural Network for Intent Recommendation[C]// In Proc. of KDD, 2019: 2478-2486.

[15] Hou S, Ye Y, Song Y, et al. Hindroid: An intelligent android malware detection system based on structured heterogeneous information network[C]// In Proc. of KDD, 2017: 1507-1515.

[16] Sun Y, Han J. Mining heterogeneous information networks: principles and methodologies[J]. Synthesis Lectures on Data Mining and Knowledge Discovery, 2012, 3(2): 1-159.

[17] Sun Y, Han J. Meta-path-based search and mining in heterogeneous information networks[J]. Tsinghua Science and Technology, 2013, 18(4): 329-338.

[18] Shi C, Zhang Z, Luo P, et al. Semantic path based personalized recommendation on weighted heterogeneous information networks[C]// In Proc. of CIKM, 2015: 453-462.

[19] Hu B, Zhang Z, Shi C, et al. Cash-out user detection based on attributed heterogeneous information network with a hierarchical attention mechanism[C]// In Proc. of AAAI, 2019, 33: 946-953.

[20] Sun Y, Yu Y, Han J. Ranking-based clustering of heterogeneous information networks with star network schema[C]// In Proc. of KDD, 2009: 797-806.

[21] Shi C, Kong X, Yu P S, et al. Relevance search in heterogeneous networks[C]// In Proc. of EDBT, 2012: 180-191.

[22] Kong X, Yu P S, Ding Y, et al. Meta path-based collective classification in heterogeneous information networks[C]// In Proc. of CIKM, 2012: 1567-1571.

[23] Shi C, Li Y, Philip S Y, et al. Constrained-meta-path-based ranking in heterogeneous information network[J]. Knowledge and Information Systems, 2016, 49(2): 719-747.

[24] Jamali M, Lakshmanan L. Heteromf: recommendation in heterogeneous information networks using context dependent factor models[C]// In Proc. of WWW, 2013: 643-654.

[25] Zhong E, Fan W, Zhu Y, et al. Modeling the dynamics of composite social networks[C]// In Proc. of KDD, 2013: 937-945.

[26] Long B, Zhang Z, Yu P S. Co-clustering by block value decomposition[C]// In Proc. of KDD, 2005: 635-640.

[27] Long B, Wu X, Zhang Z, et al. Unsupervised learning on k-partite graphs[C]// In Proc. of KDD, 2006: 317-326.

[28] Shi C, Zhou C, Kong X, et al. Heterecom: a semantic-based recommendation system in heterogeneous networks[C]// In Proc. of KDD, 2012: 1552-1555.

[29] Yu X, Ren X, Sun Y, et al. Recommendation in heterogeneous information networks with implicit user feedback[C]// In Proc. of RecSys, 2013: 347-350.

[30] Zhuang H, Zhang J, Brova G, et al. Mining query-based subnetwork outliers in heterogeneous information networks[C]// In Proc. of ICDM, 2014: 1127-1132.

[31] Kong X, Cao B, Yu P S. Multi-label classification by mining label and instance correlations from heterogeneous information networks[C]// In Proc. of KDD, 2013: 614-622.

[32] Wang R, Shi C, Philip S Y, et al. Integrating clustering and ranking on hybrid heterogeneous information network[C]// In Proc. of PAKDD, 2013: 583-594.

[33] Fan Y, Zhang Y, Hou S, et al. idev: Enhancing social coding security by cross-platform user identification between github and stack overflow[C]// In Proc. of IJCAI, 2019: 2272-2278.

[34] Amit Singhal. Introducing the Knowledge Graph: things, not strings. Official Google Blog, 2012.

[35] Özsu M T. A survey of RDF data management systems[J]. Frontiers of Computer Science, 2016, 10(3): 418-432.

[36] Suchanek F M, Kasneci G, Weikum G. Yago: a core of semantic knowledge[C]// In Proc. of WWW, 2007: 697-706.

[37] Long B, Zhang Z, Wu X, et al. Spectral clustering for multi-type relational data[C]// In Proc. of ICML, 2006: 585-592.

[38] Liu J, Wang C, Gao J, et al. Multi-view clustering via joint nonnegative matrix factorization[C]// In Proc. of ICDM, 2013: 252-260.

[39] Yang Y, Chawla N, Sun Y, et al. Predicting links in multi-relational and heterogeneous networks[C]// In Proc. of ICDM, 2012: 755-764.

[40] Tang L, Liu H, Zhang J, et al. Community evolution in dynamic multi-mode networks[C]// In Proc. of KDD, 2008: 677-685.

[41] Zhong E, Fan W, Wang J, et al. Comsoc: adaptive transfer of user behaviors over composite social network[C]// In Proc. of KDD, 2012: 696-704.

[42] Kim J, Wilhelm T. What is a complex graph?[J]. Physica A: Statistical Mechanics and its Applications, 2008, 387(11): 2637-2652.

[43] Newman M E J. The structure and function of complex networks[J]. SIAM review, 2003, 45(2): 167-256.

[44] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]// In Proc. of SIGMOD, 2008: 1247-1250.

[45] Lehmann J, Isele R, Jakob M, et al. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia[J]. Semantic Web, 2015, 6(2): 167-195.

[46] Carlson A, Betteridge J, Kisiel B, et al. Toward an architecture for never-ending language learning[C]// In Proc. of AAAI, 2010.

[47] Berant J, Chou A, Frostig R, et al. Semantic parsing on freebase from question-answer pairs[C]// In Proc. of EMNLP, 2013: 1533-1544.

[48] Damljanovic D, Bontcheva K. Named entity disambiguation using linked data[C]// In Proc. of ESWC, 2012: 231-240.

[49] Hoffmann R, Zhang C, Ling X, et al. Knowledge-based weak supervision for information extraction of overlapping relations[C]// In Proc. of ACL, 2011: 541-550.

[50] Bordes A, Weston J, Usunier N. Open question answering with weakly supervised embedding models[C]// In Proc. of ECML, 2014: 165-180.

[53] Jeh G, Widom J. Scaling personalized web search[C]// In Proc. Of WWW, 2003: 271-279.

[54] Patil V, Vasappanavara R, Ghorpade T. Comparative analysis of similarity measures in Heterogeneous Information Network[C]// In Proc. of ISCO, 2017: 297-301.

[55] He J, Bailey J, Zhang R. Exploiting transitive similarity and temporal dynamics for similarity search in heterogeneous information networks[C]// In Proc. of DASFAA, 2014: 141-155.

[56] Yao K, Mak H F. Pathsimext: revisiting pathsim in heterogeneous information networks[C]// In Proc. of WAIM, 2014: 38-42.

[57] Lao N, Cohen W W. Fast query execution for retrieval models based on path-constrained random walks[C]// In Proc. of KDD, 2010: 881-888.

[58] Lao N, Cohen W W. Relational retrieval using a combination of path-constrained random walks[J]. Machine learning, 2010, 81(1): 53-67.

[59] Xiao Y, Zhang J, Deng L. Prediction of lncRNA-protein interactions using HeteSim scores based on heterogeneous networks[J]. Scientific reports, 2017, 7(1): 1-12.

[60] Li C, Sun J, Xiong Y, et al. An efficient drug-target interaction mining algorithm in heterogeneous biological networks[C]// In Proc. of PAKDD, 2014: 65-76.

[61] Wang C, Song Y, Li H, et al. Distant meta-path similarities for text-based heterogeneous information networks[C]// In Proc. of CIKM, 2017: 1629-1638.

[62] Liu Z, Zheng V W, Zhao Z, et al. Semantic proximity search on heterogeneous graph by proximity embedding[C]// In Proc. of AAAI, 2017.

[63] Wang C, Song Y, Li H, et al. Unsupervised meta-path selection for text similarity measure based on heterogeneous information networks[J]. Data Mining and Knowledge Discovery, 2018, 32(6): 1735-1767.

[64] Yang C, Liu M, He F, et al. Similarity modeling on heterogeneous networks via automatic path discovery[C]/ In Proc. of ECML, 2018: 37-54.

[65] Fang Y, Lin W, Zheng V W, et al. Semantic proximity search on graphs with metagraph-based learning[C]// In Proc. of ICDE, 2016: 277-288.

[66] Liu Z, Zheng V W, Zhao Z, et al. Distance-aware dag embedding for proximity search on heterogeneous graphs[C]// In Proc. of AAAI, 2018.

[67] Liu Z, Zheng V W, Zhao Z, et al. Interactive paths embedding for semantic proximity search on heterogeneous graphs[C]// In Proc. of KDD, 2018: 1860-1869.

[68] Liu Z, Zheng V W, Zhao Z, et al. Subgraph-augmented path embedding for semantic user search on heterogeneous social network[C]// In Proc. of WWW, 2018: 1613-1622.

[69] Wang G, Hu Q, Yu P S. Influence and similarity on heterogeneous networks[C]// In Proc. of CIKM, 2012: 1462-1466.

[70] Wang C, Raina R, Fong D, et al. Learning relevance from heterogeneous social network and its application in online targeting[C]// In Proc. of SIGIR, 2011: 655-664.

[71] Zhang M, Hu H, He Z, et al. Top-k similarity search in heterogeneous information networks with x-star network schema[J]. Expert Systems with Applications, 2015, 42(2): 699-712.

[72] Srebro N, Jaakkola T. Weighted low-rank approximations[C]// In Proc. of ICML, 2003: 720-727.

[73] Ma H, King I, Lyu M R. Learning to recommend with social trust ensemble[C]// In Proc. of SIGIR, 2009: 203-210.

[74] Yang X, Steck H, Liu Y. Circle-based recommendation in online social networks[C]// In Proc. of KDD, 2012: 1267-1275.

[75] Wang Z, Liu H, Du Y, et al. Unified embedding model over heterogeneous information network for personalized recommendation[C]// In Proc. of AAAI, 2019: 3813-3819.

[76] Hu B, Shi C, Zhao W X, et al. Local and global information fusion for top-n recommendation in heterogeneous information network[C]/ In Proc. of CIKM, 2018: 1683-1686.

[77] Han X, Shi C, Wang S, et al. Aspect-Level Deep Collaborative Filtering via Heterogeneous Information Networks[C]// In Proc. of IJCAI, 2018: 3393-3399.

[78] Xu Y, Zhu Y, Shen Y, et al. Learning shared vertex representation in heterogeneous graphs with convolutional networks for recommendation[C]// In Proc. of AAAI, 2019: 4620-4626.

[79] Wang X, Wang R, Shi C, et al. Multi-Component Graph Convolutional Collaborative Filtering[J]. arXiv preprint arXiv:1911.10699, 2019.

[80] Nandanwar S, Moroney A, Murty M N. Fusing diversity in recommendations in heterogeneous information networks[C]// In Proc. of WSDM, 2018: 414-422.

[81] Jiang Z, Liu H, Fu B, et al. Recommendation in heterogeneous information networks based on generalized random walk model and bayesian personalized ranking[C]// In Proc. of WSDM, 2018: 288-296.

[82] Zhao H, Zhou Y, Song Y, et al. Motif Enhanced Recommendation over Heterogeneous Information Network[C]// In Proc. of CIKM, 2019: 2189-2192.

[83] Zheng J, Liu J, Shi C, et al. Recommendation in heterogeneous information network via dual similarity regularization[J]. International Journal of Data Science and Analytics, 2017, 3(1): 35-48.

[84] Yu J, Gao M, Li J, et al. Adaptive implicit friends identification over heterogeneous network for social recommendation[C]// In Proc. of CIKM, 2018: 357-366.

[85] Wen Y, Guo L, Chen Z, et al. Network embedding based recommendation method in social networks[C]// In Proc. of WWW, 2018: 11-12.

[86] Chen X, Xu H, Zhang Y, et al. Sequential recommendation with user memory networks[C]// In Proc. of WSDM, 2018: 108-116.

[87] Zhang Y, Ai Q, Chen X, et al. Joint representation learning for top-n recommendation with heterogeneous information sources[C]// In Proc. of CIKM, 2017: 1449-1458.

[88] Yu F, Li Z, Jiang S, et al. Personalized POI Groups Recommendation in Location-Based Social Networks[C]// In Proc. of APWeb, 2017: 114-123.

[89] Wang H, Zhang F, Xie X, et al. DKN: Deep knowledge-aware network for news recommendation[C]// In Proc. of WWW, 2018: 1835-1844.

[90] Luo C, Guan R, Wang Z, et al. Hetpathmine: A novel transductive classification algorithm on heterogeneous information networks[C]// In Proc. of ECIR, 2014: 210-221.

[91] Wang C, Song Y, Li H, et al. Text classification with heterogeneous information network kernels[C]// In Proc. of AAAI, 2016.

[92] Gupta M, Kumar P, Bhasker B. HeteClass: A Meta-path based framework for transductive classification of objects in heterogeneous information networks[J]. Expert Systems with Applications, 2017, 68: 106-122.

[93] Zhang Y, Xiong Y, Kong X, et al. Deep collective classification in heterogeneous information networks[C]// In Proc. of WWW, 2018: 399-408.

[94] Shi J, Malik J. Normalized cuts and image segmentation[J]. IEEE Transactions on pattern analysis and machine intelligence, 2000, 22(8): 888-905.

[95] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical review E, 2004, 69(2): 026113.

[96] Li X, Kao B, Ren Z, et al. Spectral Clustering in Heterogeneous Information Networks[C]// In Proc. of AAAI, 2019, 33: 4221-4228.

[97] Dall'Amico L, Couillet R, Tremblay N. Revisiting the Bethe-Hessian: improved community detection in sparse heterogeneous graphs[C]// In Proc. of NIPS, 2019: 4039-4049.

[98] Cruz J D, Bothorel C, Poulet F. Integrating heterogeneous information within a social network for detecting communities[C]// In Proc. of ASONAM, 2013: 1453-1454.

[99] Li X, Wu Y, Ester M, et al. Semi-supervised clustering in attributed heterogeneous information networks[C]// In Proc. of WWW, 2017: 1621-1629.

[100] Chen L, Gao Y, Zhang Y, et al. Efficient and Incremental Clustering Algorithms on Star-Schema Heterogeneous Graphs[C]// In Proc. of ICDE, 2019: 256-267.

[101] Sun Y, Han J, Zhao P, et al. Rankclus: integrating clustering with ranking for heterogeneous information network analysis[C]// In Proc. of ICDT, 2009: 565-576.

[102] Sun Y, Barber R, Gupta M, et al. Co-author relationship prediction in heterogeneous bibliographic networks[C]// In Proc. of ASONAM, 2011: 121-128.

[103] Sun Y, Han J, Aggarwal C C, et al. When will it happen? relationship prediction in heterogeneous information networks[C]// In Proc. of WSDM, 2012: 663-672.

[104] Yu X, Gu Q, Zhou M, et al. Citation prediction in heterogeneous bibliographic networks[C]// In Proc. of ICDM, 2012: 1119-1130.

[105] Chen J, Gao H, Wu Z, et al. Tag co-occurrence relationship prediction in heterogeneous information networks[C]// In Proc. of ICPADS, 2013: 528-533.

[106] Cao B, Kong X, Philip S Y. Collective prediction of multiple types of links in heterogeneous information networks[C]// In Proc. of ICDM, 2014: 50-59.

[107] Chen H, Yin H, Wang W, et al. PME: projected metric embedding on heterogeneous networks for link prediction[C]// In Proc. of KDD, 2018: 1177-1186.

[108] Liu F, Xia S T. Link prediction in aligned heterogeneous networks[C]// In Proc. of PAKDD, 2015: 33-44.

[109] Wang H, Zhang F, Hou M, et al. Shine: Signed heterogeneous information network embedding for sentiment link prediction[C]// In Proc. of WSDM, 2018: 592-600.

[110] Cao X, Zheng Y, Shi C, et al. Meta-path-based link prediction in schema-rich heterogeneous information network[J]. International Journal of Data Science and Analytics, 2017, 3(4): 285-296.

[111] Zhao Q, Bhowmick S S, Zheng X, et al. Characterizing and predicting community members from evolutionary and heterogeneous networks[C]// In Proc. of CIKM, 2008: 309-318.

[112] Aggarwal C, Xie Y, Yu P S. On dynamic link inference in heterogeneous networks[C]// In Proc. of ICDM, 2012: 415-426.

[113] Aggarwal C C, Xie Y, Yu P S. A framework for dynamic link prediction in heterogeneous networks[J]. Statistical Analysis and Data Mining: The ASA Data Science Journal, 2014, 7(1): 14-33.

[114] Li J, Cheng K, Wu L, et al. Streaming link prediction on dynamic attributed networks[C]// In Proc. of WSDM, 2018: 369-377.

[115] Wang C, Sun Y, Song Y, et al. Relsim: relation similarity search in schema-rich heterogeneous information networks[C]// In Proc. of ICDM, 2016: 621-629.

[116] Zheng Y, Shi C, Cao X, et al. A Meta Path based Method for Entity Set Expansion in Knowledge Graph[J]. IEEE Transactions on Big Data, 2018.

[117] Lee S, Park C, Yu H. BHIN2vec: Balancing the Type of Relation in Heterogeneous Information Network[C]// In Proc. of CIKM, 2019: 619-628.

[118] Lu Y, Shi C, Hu L, et al. Relation structure-aware heterogeneous information network embedding[C]// In Proc. of AAAI, 2019, 33: 4456-4463.

[119] Hu B, Fang Y, Shi C. Adversarial Learning on Heterogeneous Information Networks[C]// In Proc. of KDD, 2019: 120-129.

[120] Zhang C, Song D, Huang C, et al. Heterogeneous graph neural network[C]// In Proc. of KDD, 2019: 793-803.

[121] Zhang Y, Ai Q, Chen X, et al. Joint representation learning for top-n recommendation with heterogeneous information sources[C]// In Proc. of CIKM, 2017: 1449-1458.

[122] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations[C]// In Proc. of KDD, 2014: 701-710.

[123] Tang J, Qu M, Wang M, et al. Line: Large-scale information network embedding[C]// In Proc. of WWW, 2015: 1067-1077.

[124] Wang D, Cui P, Zhu W. Structural deep network embedding[C]// In Proc. of KDD, 2016: 1225-1234.

[125] Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: A survey[J]. Knowledge-Based Systems, 2018, 151: 78-94.

[126] Hamilton W L, Ying R, Leskovec J. Representation learning on graphs: Methods and applications[J]. arXiv preprint arXiv:1709.05584, 2017.

[127] Tang J, Qu M, Mei Q. Pte: Predictive text embedding through large-scale heterogeneous text networks[C]// In Proc. of KDD, 2015: 1165-1174.

[128] Xu L, Wei X, Cao J, et al. Embedding of Embedding (EOE) Joint Embedding for Coupled Heterogeneous Networks[C]// In Proc. of WSDM, 2017: 741-749.

[129] Dong Y, Chawla N V, Swami A. metapath2vec: Scalable representation learning for heterogeneous networks[C]// In Proc. of KDD, 2017: 135-144.

[130] Fu T, Lee W C, Lei Z. Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning[C]// In Proc. of CIKM, 2017: 1797-1806.

[131] Chen Y, Wang C. HINE: Heterogeneous information network embedding[C]// In Proc. of DASFAA, 2017: 180-195.

[132] Shang J, Qu M, Liu J, et al. Meta-path guided embedding for similarity search in large-scale heterogeneous information networks[J]. arXiv preprint arXiv:1610.09769, 2016.

[133] Wang X, Zhang Y, Shi C. Hyperbolic heterogeneous information network embedding[C]// In Proc. of AAAI, 2019, 33: 5337-5344.

[134] Jiang H, Song Y, Wang C, et al. Semi-supervised Learning over Heterogeneous Information Networks by Ensemble of Meta-graph Guided Random Walks[C]// In Proc. of IJCAI, 2017: 1944-1950.

[135] He Y, Song Y, Li J, et al. Hetespaceywalk: a heterogeneous spacey random walk for heterogeneous information network embedding[C]// In Proc. of CIKM, 2019: 639-648.

[136] Hussein R, Yang D, Cudré-Mauroux P. Are Meta-Paths Necessary? Revisiting Heterogeneous Graph Embeddings[C]// In Proc. of CIKM, 2018: 437-446.

[137] Li X, Hong H, Liu L, et al. A structural representation learning for multi-relational networks[J]. arXiv preprint arXiv:1805.06197, 2018.

[138] Zhang J, Xia C, Zhang C, et al. BL-MNE: emerging heterogeneous social network embedding through broad learning with aligned autoencoder[C]// In Proc. of ICDM, 2017: 605-614.

[139] Qu M, Tang J, Han J. Curriculum learning for heterogeneous star network embedding via deep reinforcement learning[C]// In Proc. of WSDM, 2018: 468-476.

[140] Zhou S, Bu J, Wang X, et al. HAHE: Hierarchical attentive heterogeneous information network embedding[J]. arXiv preprint arXiv:1902.01475, 2019.

[141] Chen X, Yu G, Wang J, et al. ActiveHNE: Active heterogeneous network embedding[J]. arXiv preprint arXiv:1905.05659, 2019.

[142] Xiong Y, Zhang Y, Kong X, et al. NetCycle+: a framework for collective evolution inference in dynamic heterogeneous networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(8): 1547-1560.

[143] Wei H, Zhang F, Yuan N J, et al. Beyond the words: Predicting user personality from heterogeneous information[C]// In Proc. of WSDM, 2017: 305-314.

[144] Li A, Qin Z, Liu R, et al. Spam Review Detection with Graph Convolutional Networks[C]// In Proc. of CIKM, 2019: 2703-2711.

[145] Yang C, Zhang C, Chen X, et al. Did you enjoy the ride? understanding passenger experience via heterogeneous network embedding[C]// In Proc. of ICDE, 2018: 1392-1403.

[146] Zhu Z, Li R, Shan M, et al. TDP: Personalized Taxi Demand Prediction Based on Heterogeneous Graph Embedding[C]// In Proc. of SIGIR, 2019: 1177-1180.

[147] Ye Y, Hou S, Chen L, et al. Out-of-sample node representation learning for heterogeneous graph in real-time android malware detection[C]// In Proc. of AAAI, 2019: 4150-4156.

[148] Hou S, Fan Y, Zhang Y, et al. αCyber: Enhancing Robustness of Android Malware Detection System against Adversarial Attacks on Heterogeneous Graph based Model[C]// In Proc. of CIKM, 2019: 609-618.

[149] Liu Z, Chen C, Yang X, et al. Heterogeneous graph neural networks for malicious account detection[C]// In Proc. of CIKM, 2018: 2077-2085.

[150] Wang P C, Li C T. Spotting Terrorists by Learning Behavior-aware Heterogeneous Network Embedding[C]// In Proc. of CIKM, 2019: 2097-2100.

[151] Zhang Y, Fan Y, Hou S, et al. iDetector: Automate underground forum analysis based on heterogeneous information network[C]// In Proc. of ASONAM, 2018: 1071-1078.

[152] Zhang Y, Fan Y, Song W, et al. Your style your identity: Leveraging writing and photography styles for drug trafficker identification in darknet markets over attributed heterogeneous information network[C]// In Proc. of WWW, 2019: 3448-3454.

[153] Fan S, Shi C, Wang X. Abnormal event detection via heterogeneous information network embedding[C]// In Proc. of CIKM, 2018: 1483-1486.

[154] Ranjbar V, Salehi M, Jandaghi P, et al. QANet: Tensor Decomposition Approach for Query-based Anomaly Detection in Heterogeneous Information Networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(11): 2178-2189.

[155] Hosseini A, Chen T, Wu W, et al. Heteromed: Heterogeneous information network for medical diagnosis[C]// In Proc. of CIKM, 2018: 763-772.

[156] Fan Y, Zhang Y, Ye Y, et al. Automatic Opioid User Detection from Twitter: Transductive Ensemble Built on Different Meta-graph Based Similarities over Heterogeneous Information Network[C]// In Proc. of IJCAI, 2018: 3357-3363.

[157] Linmei H, Yang T, Shi C, et al. Heterogeneous graph attention networks for semi-supervised short text classification[C]// In Proc. of EMNLP, 2019: 4823-4832.

[158] Hu L, Li C, Shi C, et al. Graph neural news recommendation with long-term and short-term interest modeling[J]. Information Processing & Management, 2020, 57(2): 102142.

[159] Li L, Gan Z, Cheng Y, et al. Relation-aware graph attention network for visual question answering[C]// In Proc. of ICCV, 2019: 10313-10322.

[164] Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020.

[165] Zhou J, Cui G, Zhang Z, et al. Graph neural networks: A review of methods and applications[J]. arXiv preprint arXiv:1812.08434, 2018.

[166] Zhao J, Wang X, Shi C, et al. Network Schema Preserving Heterogeneous Information Network Embedding[C]// In Proc. of IJCAI, 2020.

[167] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[C]// In Proc. of NIPS, 2013: 2787-2795.

[168] Wang Q, Mao Z, Wang B, et al. Knowledge graph embedding: A survey of approaches and applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(12): 2724-2743.

[169] Wan G, Du B, Pan S, et al. Reinforcement Learning based Meta-path Discovery in Large-scale Heterogeneous Information Networks[C]// In Proc. of AAAI, 2020

[170] Wang H, Wang J, Wang J, et al. Graphgan: Graph representation learning with generative adversarial nets[C]// In Proc. of AAAI, 2018

[171] Zhao K, Bai T, Wu B, et al. Deep Adversarial Completion for Sparse Heterogeneous Information Network Embedding[C]// In Proc. of WWW, 2020: 508-518.

附中文参考文献:

[51] 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.中国计算机学会通讯,2017,13(11):35-40.

[52] 石川,孙怡舟.异质网络表征学习的研究进展.中国计算机学会通讯,2018,14(3):16-21.

[160] 周慧,赵中英,李超.面向异质信息网络的表示学习方法研究综述[J].计算机科学与探索,2019,13(7):1082-1094.

[161] 张悦,王晓丹,卜霄菲等.异质信息网络链接预测研究与新发展[J].沈阳师范大学学报(自然科学版),2019,37(5):453-460.

[162] 尹赢,吉立新,程晓涛等.基于同质子图变换的异质网络表示学习[J].计算机工程,2019,45(11):204-212.

[163] 马毅,郭杏莉,孙宇彤等.基于 HeteSim 的疾病关联长非编码 RNA 预测[J].计算机研究与发展,2019,56(9):1889-1896.