

大规模复杂信息网络表示学习:概念、方法与挑战

齐金山^{1),2)} 梁 循¹⁾ 李志宇¹⁾ 陈燕方³⁾ 许 媛¹⁾

¹⁾(中国人民大学信息学院 北京 100872)

²⁾(淮阴师范学院计算机科学与技术学院 江苏 淮安 223300)

³⁾(中国人民大学信息资源管理学院 北京 100872)

摘 要 大数据时代的到来,使得当前的复杂信息网络研究领域面临着三个基础性问题,即网络的动态性、大规模性以及网络空间的高维性.传统复杂信息网络特征的表示通常以邻接矩阵、出入度、中心性等离散型方式表达,这种表达方式在现有的大规模动态信息网络的新环境下,其计算效率及准确率都受到了很大的挑战.随着机器学习算法的不断发展,复杂信息网络的特征表示学习同样也引起了越来越多的关注.与自然语言中的词向量学习的目标类似,目前较为前沿的大规模复杂网络特征表示学习方法的目的是将网络中任意顶点的结构特征映射到一个低维度的、连续的实值向量,在进行这种映射的过程中,尽量保留顶点之间的结构特征关系,使大规模网络特征学习能够有效地应用于各类网络应用中,如网络中的链接预测、顶点分类、个性化推荐、大规模社区发现等.通过对复杂信息网络特征的学习,不仅能够有效缓解网络数据稀疏性问题,而且把网络中不同类型的异质信息融合为整体,可以更好地解决某些特定问题.同时,还能够高效地实现语义相关性操作,从而显著提升在大规模,特别是超大规模的的网络中进行相似性顶点匹配的计算效率等.该文主要对近些年来关于复杂信息网络表示学习的方法和研究现状进行了总结,并提出自己的想法和意见.首先概述了表示学习的发展历史,然后分别阐述了有关大规模复杂信息网络、网络表示学习等基本概念与理论基础;接着,根据学习模型的不同,对经典的、大规模的、基于内容的、基于融合的以及异构的网络表示学习模型进行了全面的分析与比较.另外,对当前的网络表示学习方法所采用的实验数据集、评测指标以及应用场景等也进行了总结概括.最后给出了大规模复杂信息网络表示学习的研究难题以及未来的研究方向.大规模复杂网络表示学习是一个复杂的问题.当前研究中,大多数学习模型是根据复杂网络的结构或者内容来进行顶点的特征表示学习.只有融合复杂网络结构特征和内容特征的表示学习才能够更好地反映出一个网络特征的真实情况,使得学习得到的网络特征表示更具有意义与价值.

关键词 大规模复杂信息网络;网络特征;顶点嵌入;网络表示学习;深度学习;特征学习

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2018.02394

Representation Learning of Large-Scale Complex Information Network: Concepts, Methods and Challenges

QI Jin-Shan^{1),2)} LIANG Xun¹⁾ LI Zhi-Yu¹⁾ CHEN Yan-Fang³⁾ XU Yuan¹⁾

¹⁾(School of Information, Renmin University of China, Beijing 100872)

²⁾(School of Computer Science & Technology, Huaiyin Normal University, Huai'an, Jiangsu 223300)

³⁾(School of Information Resource Management, Renmin University of China, Beijing 100872)

Abstract With the arrival of the age of big data, the current study of complex information network is facing three severe challenges: dynamicity, large scale and high-dimensionality of the network. Traditionally, the characteristics of complex information networks are represented in discrete forms such as adjacency matrix, in-out degree and centrality. Such representations have

收稿日期:2017-01-10;在线出版日期:2017-12-08.本课题得到国家自然科学基金(71271211,71531012)资助.齐金山,男,1977年生,博士研究生,主要研究方向为社会计算、数据挖掘. E-mail: qijinshan@sina.com.cn.梁 循(通信作者),男,1965年生,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为神经网络、支持向量机、社会计算. E-mail: xun_liang@163.com.李志宇,男,1991年生,博士研究生,中国计算机学会(CCF)会员,主要研究方向为社会计算、自然语言处理、Web 挖掘.陈燕方,女,1992年生,博士研究生,主要研究方向为 Web 挖掘、自然语言处理.许 媛,女,1993年生,硕士研究生,主要研究方向为社会计算.

great disadvantages in computational efficiency and accuracy in the new environment of large-scale dynamic information network. Meanwhile, with the advance in machine learning algorithms, the representation learning of complex information network receives more attention. Similar to the learning of word vectors in natural language processing, the representation learning of large-scale network aims to map the structural characteristics of each vertex in the network to a low-dimensional, real-valued vector, during which the structural relationship of vertices in the network is kept to the greatest extent, so that various types of network applications can be effectively applied, such as link prediction, vertex classification, personalized recommendation, and large scale community discovery. More precisely, the advantages of representation learning in complex information networks are three-fold. First, it reduces the effect of data sparsity in networks. Second, heterogeneous information is integrated into the same vector space so that specific applications can be applied easily. Third, semantic operations can be implemented in a way which dramatically improves the efficiency of node similarity computing in large-scale networks. To this end, the paper proposes a taxonomy of both the classics and the state of the arts on representation learning of information networks. We first provide a historical overview of representation learning in graphs, followed by the elaboration of correlated concepts and theories, and then a comprehensive analysis of various learning models is proposed. We consider the classic models, for example, the spectral method and optimization based methods; and models for large scale networks, including the high-order relationship based models, semi-supervised models and models with scalability. Both two categories of models focus on the structure of network. In contrast, the probability based models, i. e., the topic models, which learn the content of documents by means of finding the static and dynamic patterns, can be associated to the learning of contents in networks, e. g., content with each vertex. The combination of these two aspects, such as the matrix factorization based models and probabilistic graphical models are also considered. Finally we discuss models for heterogeneous networks. We compare different methods from several perspectives in detail and derive some conclusions. In addition, the paper presents a summary of the experimental data sets, evaluation metrics and application scenarios of different graph representation learning methods. We also discuss the existing problems and future studies in representation learning of large-scale complex information network. In brief, most existing works focus on either the structure or content of the network in the representation learning of vertices. However, network properties should be better revealed by considering both characteristics. Room for future improvements of representation learning typically lies in the fault tolerance of network feature extraction, the adaptivity for dynamic networks, the combination of heterogeneous information in network, the universality of representation, the distributed network representation and feature learning for specific sub-graph structures.

Keywords large-scale complex information network; network features; vertex embedding; network representation learning; deep learning; feature learning

1 引言

大规模复杂信息网络可以看作由顶点(nodes)和链接(links)或边(edges)所构成的复杂抽象组合, 包含顶点之间通过有向(directed)或无向(undirected)

链接进行信息或功能的交互. 不同于普通网络, 大规模复杂信息网络顶点的数量已经发展到百万甚至数十亿, 而且包含的数据亦呈现多样化、异质化以及动态性等特征. 其中以社会网络为代表的大规模复杂信息网络在顶点与链接的数量上、包含的内容形式上以及交互关系的种类上都以极快的速度不断更

新. 以新浪微博为例, 在顶点数量上, 截至 2015 年 9 月 30 日, 微博月活跃用户为 2.22 亿, 相比 2014 年同期增长 33%. 在顶点关系上, 微博已经从原有的关注关系进一步演化为转发关系、话题关系、营销关系和行业关系等多重关系网络的固定或随机组合. 在顶点与链接的属性上, 用户所发布的内容以及用户参与的形式呈现多样化, 不仅包括传统的文本、图片、音频和视频等多媒体信息, 还衍生出微博投票、微博众筹、微博旅游等多种互动与参与形式^①. 正是由于大规模复杂信息网络原始表达的丰富性和高量级, 这使得对于网络的表示学习与分析的需求急剧增长成为必然.

表示学习 (Representation Learning, RL) 是指通过将符号化的数据编码到低维向量空间中^[1]. 学习数据的表示使得在构建机器学习分类器或其它预测时更容易提取有用的信息, 同时这种表示得到的结果也可用于作为监督分类器的输入数据. 相比传统的 One-Hot 表示方法^[2], 表示学习不仅能够获取数据之间的相似性, 而且同时可以缓解数据的稀疏性 (sparsity) 问题. Bengio 等人^[1]指出: 表示学习是通过对数据进行表达形式的变化, 使其数据所包含的信息更加容易被提取和分析, 即将原来由人工设定的特征工程转换为机器的自我学习过程. 与此相似, 知识图谱表示学习从多角度解决知识库中实体和关系的建模问题, 把实体或关系映射到低维特征向量空间中, 以实现实体、关系及其之间的复杂语义的高效表示. Bordes 等人^[3]提出了一种将实体与关系嵌入到低维向量空间中的 TransE 模型. 自 TransE 提出之后, 大部分知识表示学习模型是以它为基础进行的扩展或改进, 如 TransH^[4]、TransR^[5]、TransD^[6]等模型. 同样的, 复杂信息网络表示学习则是通过对网络的结构特征以及内容特征等进行学习, 得到网络的统一表现形式即向量化, 并以此为输入, 将其拓展到例如多标签分类、社区发现、隐私保护系统等应用中.

复杂信息网络的表示学习相比于其它数据类型的表示学习难度更大. 从数据构型上看, 复杂信息网络的表示学习涉及到对象关系学习 (如顶点离散关系学习、顶点连续关系学习) 和对象属性学习 (如顶点属性学习、边属性学习), 以及它们之间的交互学习. 从数据体量上看, 复杂信息网络的表示学习涉及到从 10^2 到 10^9 甚至更大的顶点或边集规模. 其中每一个顶点既可以被单独的划分成为一个类, 也可以和众多的顶点按照社区标签形式划分成多个类. 把

图嵌入算法 (graph embedding) 如 SVD^[7]、LLE^[8]、Isomap^[9]、Laplacian Eigenmaps^[10]、MDS^[11] 等应用于网络中, 往往需要构造亲和度矩阵, 通过特征向量求解的方式获得顶点的低维表示. Li 等人^[12]研究了特征学习技术的图结构化输入. 在图神经网络的工作基础上, 通过修改使用门控循环单元以及优化技术, 然后扩展到输出序列. Ahmed 等人^[13]提出了一种图因式分解技术, 通过使用随机梯度下降优化的矩阵分解找到大规模图的低维向量形式. 然而在复杂信息网络顶点数量急剧增加时, 图嵌入算法的计算复杂度往往是无法承受的. 最近, 基于深度学习技术在语音识别、计算机视觉和自然语言处理等领域取得了很大的成功. 在自然语言处理中, 文本表示学习模型 (如词表示学习^[14-15] 和文档的表示学习^[16]) 具有算法高效性和可计算性的优点, 能够处理大规模的文本数据. 因此, 借鉴文本表示学习模型的思想, 针对大规模复杂信息网络设计有效的网络表示模型已经逐步成为当前的研究热点. 相应地, 一些重要的研究成果频频出现在《Nature》、《Science》、TKDE、AAAI、KDD 等国际重要期刊和会议上.

大规模复杂信息网络表示学习主要的作用是用向量表示网络中顶点和链接, 具有以下优点:

(1) 可以有效缓解复杂大规模信息网络数据的稀疏性问题. 网络中各顶点的向量都是稠密的, 因此可以对任意顶点之间的语义相关性进行度量. 此外, 通过把网络中的顶点映射到相同的特征空间中, 使用高出/入度顶点的结构信息用于帮助低出/入度顶点的结构或语义特征的表示, 提供低出/入度顶点的语义表示的精准性.

(2) 可实现大规模复杂信息网络中异质顶点信息融合. 大规模复杂信息网络中包含大量的不同类型的顶点和关系, 这些不同类型的异质信息需要融合为整体, 才能得到更有效应用. 通过设计合适的网络特征学习模型, 把不同种类的顶点映射到统一的向量空间里, 从而建立统一的特征空间, 例如同时基于结构或内容, 实现异质信息的融合.

(3) 可提升大规模复杂信息网络的相关应用的计算效率. 基于特征向量的降维算法在处理大规模复杂信息网络时, 算法复杂度高、扩展性差. 网络表示学习使用向量表示, 例如在大规模网络量级的情況下, 通过简单的查询和向量相似性计算就能够高

① 2015 年微博用户发展报告. <http://www.199it.com/archives/422583.html>. 2016-2-17

效地实现结构/语义的相关性,从而显著提升计算效率.同时使得学习得到的向量能够简单而快速的被其它应用所采纳.因此,大规模复杂信息网络表示学习的研究,对帮助我们分析和理解网络,预测用户行为等具有重大意义.

本文对近年来关于大规模复杂信息网络表示学习的方法及研究现状进行归纳,在此基础上提出自己的看法和建议.首先在第2节给出大规模复杂信息网络表示学习所涉及的有关概念与理论基础;在第3节讨论当前大规模复杂信息网络表示学习的主流方法与模型,同时对现有方法之间的优劣进行对比分析;第4节则从大规模复杂信息网络学习所涉及到的应用场景、数据集与评测指标三个方面,对大规模复杂网络学习模型的评测进行概述;最后,本文在第5节和第6节分别讨论当前大规模复杂信息网络表示学习的研究难点和未来的可能研究方向.

2 基本概念

2.1 网络与大规模复杂信息网络

2.1.1 网络

网络由一组顶点和连接顶点的边组成.生活中,很多系统和组织都可以用网络来刻画,如亲朋好友之间的社交网络.关于网络的最早研究可以追溯到数学中的图论问题.1735年欧拉解决了格尼斯堡七桥问题^[17],从而开创了数学的另一分支-图论.如今图论成为了专门用来刻画和分析网络性质的一门学科.在科学界,网络得到了广泛研究,如在社会学领域中,用顶点来表示行动者,顶点和顶点之间连线则代表行动者之间的关系.同样,网络在经济学、图书情报学、心理学等领域中得到较多的应用.人们通过对网络进行研究,希望能够发现其内在机制、演化规律以及网络的传播模式等内容. Rapoport 等人^[18]提出的数学模型强调了网络中度分布的重要性. Milgram 在1969年开展了“小世界”(small-world)实验,探索现实社会中信息如何在好友网络之间传递^[19]. Burt 于1992年提出的结构洞理论,揭示了网络中一些对象之间存在直接的关联,而一些对象出现关系间断的情形,即从整体上看,网络出现结构洞穴.早期的网络研究受到诸多的限制,对数据的收集和采样有限,因此传统的网络研究会有准确性低、个人主观性强以及样本规模小等方面的问题^[20].

随着计算技术的发展和越来越强大的互联网,对于网络的研究出现了很多新的方法.以往对小型

网络的研究方法,在当前大规模复杂网络中几乎没有什么用或没有什么实质意义.例如从一个小型网络中移除某个顶点后,剩余的顶点可能成为该网络链接中的关键顶点问题.在小型网络中分析这样的问题是很有意义的.但在有数以百万计顶点的网络中,即使移除任意顶点后也不会对其它顶点产生重要影响,因此这一问题就显得不再有意义.以往的网络规模都比较小,仅包含数百个顶点,对其进行研究分析可以解决一些简单的网络结构问题.现有的大规模复杂网络往往包含数以百万计的顶点,很难依靠个人力量将其描述清楚.对于当前网络的研究方法主要是如何挖掘出大规模网络的结构、特征等有效信息^[20].

2.1.2 大规模复杂信息网络

随着互联网技术以及大数据的蓬勃发展,使得顶点带有丰富信息内容的复杂网络日益增多,同时网络的规模也在不断地扩大.根据网络顶点所承载信息的不同类型,这些网络可以包括社会网络、科学家合作网络、引文网络、通讯网络、公路网络和 Web 网络等,本文将其统称为复杂信息网络.复杂信息网络是一种顶点带有明显多媒体内容(例如:文本、图像、音频或视频)特征的现实网络.复杂信息网络在宏观上具有网络的结构关系属性,在微观上每个顶点具有内容属性.同时,随着大数据时代的来临,各类复杂信息网络的量级还在迅猛增长且呈动态形式,其顶点的个数甚至达到几十亿之多.因此,原有的网络分析方法在如此大规模的甚至超大规模的数据面前显得力不从心,所以针对大规模复杂信息网络的相关研究已经受到了学术界和工业界的广泛关注.

关于网络的定义最早来源于图论, Tang 等人^[21]则给出了有关大规模的复杂信息网络的定义: $G=(V, E, \mathbf{W})$ 表示一个复杂信息网络,其中 $V=\{v_1, v_2, \dots, v_N\}$ 代表网络中顶点的集合,每个顶点代表一个数据对象, $E=\{e_1, e_2, \dots, e_M\}$ 则是边的集合,每条边代表两个数据对象之间的关系, $\mathbf{W}$ 为一个赋权矩阵,表示关系的强度.根据以上定义并结合之前的相关研究,本文给出了大规模的复杂信息网络的另一种形式化定义.

定义 1(大规模复杂信息网络). 给定 m 种类型的数据对象集合 $\{X_1, X_2, \dots, X_m\}$, 其中 $X_1=\{x_{11}, x_{12}, \dots, x_{1n_1}\}, \dots, X_m=\{x_{m1}, x_{m2}, \dots, x_{mn_m}\}$. 用 $G=(V, E, \mathbf{W}, C)$ 表示一个大规模的复杂信息网络,

其中 $V = \bigcup_{i=1}^m X_i$ 表示所有的数据对象组成的顶点集合, $|V|$ 的大小一般在数十万以上, E 表示 V 中任意两个数据对象之间的链接构成的集合, W 与上述定义一样, 代表一个赋权矩阵, 表示数据对象之间链接的权重, C 表示所有数据对象的多媒体内容构成的集合, $C = \bigcup_{i=1}^m D_i$, $D_i = \bigcup_{j=1}^{n_i} d_{ij}$, d_{ij} 则表示与顶点 v_{ij} 相关联的数据对象内容. 不包含顶点内容的嵌入时 C 为空集, $C = \emptyset$.

若 $m=1$ 时, G 为一个同构网络, 否则 G 为一个异构网络. 若不考虑数据对象上的多媒体内容 C , 则对网络的分析是基于网络结构形式.

2.2 表示学习和网络表示学习

2.2.1 表示学习

基于数据的问题通常需要完成数据表示、目标构建以及问题求解三个方面. 而解决整个问题的关键则是数据表示, 它也决定了后续的计算空间和时间复杂度. 尤其是随着非结构化数据、半结构化数据以及多媒体数据的急剧增长, 如何统一且高效地进行数据表示就变得非常重要.

表示学习是机器学习领域中另外一个研究方向, 又称为特征学习, 其主要任务是自动对原始数据集进行学习从而得到一个新的特征. 将新的特征作为不同机器学习算法的输入, 免除了低效的人工提取特征过程. 表示学习已经广泛应用于语音识别与信号处理、多任务和迁移学习、对象识别和自然语言处理等方面. 例如, 文本表示学习通过对文本的共现情景进行学习, 得到词语在低维空间 (low dimension) 上的向量表示^[14, 22-23], 这种低维度的向量表示能够有效地显示出词语之间的语义关系^[14], 而且更易于被应用到其它的系统中. 由于数据集的不同, 可以把表示学习分为监督表示学习 (Supervised Representation Learning)、无监督表示学习 (Unsupervised Representation Learning) 两类.

监督表示学习是指从已经标注的数据中学习数据的新特征表示. 比如多层神经网络在处理分类或者回归任务时的隐含层就可以作为输入数据的新的特征. Zhuang 等人^[24]提出的基于深度自编码的迁移学习就是一类典型的有监督表示学习. 这种深度自编码由两个编码层组成: 嵌入层和标签编码层. 在嵌入层, 源域和目标域之间的嵌入实例的分布采用最小化 KL (Kullback-Leibler) 散度. 在标签编码层, 源域的标签信息使用 softmax 回归模型编码. 类似的监督表示学习算法有文献^[25-26]等.

无监督表示学习是指从未标注数据中学习数据新的特征表示. 有关算法通常用来从高维输入数据中发现有意义的低维数据特征表示, 实质上也就是一种降低维度算法. O'Shea 等人^[27]提出的无线电通讯信号的无监督表示学习就属于此类算法. 类似的代表性算法还包括无监督字典学习^[28]、独立成分分析^[29]、主成分分析和局部线性嵌入^[30]等. Bengio 等人^[1]详细地综述了表示学习的评价准则和主要方法, 但没有涉及大规模的网络数据的特征表示学习.

2.2.2 网络表示学习

网络表示学习又称为网络特征学习, 定义为给定一个复杂信息网络 $G=(V, E, W, C)$, G 对应的顶点特征矩阵 X 是一个高度稀疏的矩阵, 其维数通常为 $|V| \times m$ (m 是顶点属性的特征空间大小), 对每个顶点 $v \in V$, 低维向量表示学习 $r_v \in R^k$, k 实质上远小于 $|V|$, r_v 表示为一个稠密的实数向量表示.

在定义中, 复杂信息网络 G 的链接关系不限定方向, 即链接可以是有向链接也可以是无向链接. 例如 Enron 数据集包含了 50 多万封电子邮件. 如果把邮件地址作为顶点, 若两封邮件之间有通信往来, 则在这两个地址之间建立一条无向链接, 这样构成的网络有 36 692 个顶点以及 367 662 条链接; 又如提供检索计算机方面的重要国际期刊和会议论文的数据库 DBLP, 收录了各类刊物达到 140 万种. 该检索库存在两种类型的网络: 作者引用网络和论文引文网络. 作者引用网络记录了一位作者所写的论文数以及被其他作者所引用的次数, 两者之间就形成了一条有向连接, 同样在论文引用网络中, 一篇论文引用了另一篇论文, 两者之间也形成了一条有向连接. 除了 G 的链接关系不考虑方向性之外, 其包含的顶点类型也不做限制. 如果 G 中的顶点代表的实体是单一类型, 则 G 是一个同构网络. 大部分已有的网络表示学习模型是基于同构网络而建立的算法, 如参考文献^[21, 31-32]等. 如果 G 中的顶点含有多种类型, 则 G 是一个异构网络. 以 YouTube 社交网站为例, 在该网络中, 视频、标签和用户代表不同类型的顶点, 顶点间又通过相同或不同类型的联系而关联. 有关异构网络的表示学习模型研究开展相对较少, 已有的研究方法主要把异构网络映射到同构网络上, 采用同构网络上的学习模型来加以实现, 如参考文献^[33-35]等, 这些方法没有充分利用不同的顶点标签或特性之间的相关性, 且假定任务的本质在现实世界场景中的结果也并不理想. 如果

仅仅获取网络中的结构信息, 顶点本身没有提供其它内容信息, 那么复杂信息网络中的特征矩阵 $\mathbf{X}$ 只是一个空矩阵, 否则 $\mathbf{X}$ 是非空矩阵.

近年来, 研究人员广泛地开展了在复杂信息网络方面的研究应用, 如顶点分类^[36]、标签推荐^[37]、异常检测^[38]以及链接预测^[39]等. 这些应用所面临的共同问题就是复杂信息网络的数据稀疏性. 为了解决这一问题, 网络表示学习用统一的低维向量空间对每个顶点进行编码和表示, 这有助于更好地理解顶点之间的语义关联, 并进一步减轻因为数据稀疏性造成的不便^[31]. Li 等人^[40]进一步阐述了网络表示学习的主要任务是对复杂信息网络的内容或结构进行表示. 用低维度的连续特征表示原有的高维度离散特征, 从而有效地解决网络中存在的稀疏性难题, 还可得到网络的不同表示形式, 作为其它应用的输入从而进行下一步的分析, 如对原始网络构建拓展的研究模型.

图 1 展示了使用网络表示学习方法获得原始网络 Karate 中的社区结构和嵌入向量空间 $\mathbf{R}^2$ 中社区结构之间的对应关系. 其中图 1(a) 代表 Karate 在模块度最大化时社区划分结构 (通过顶点颜色表示), 图 1(b) 显示该原始网络映射到二维向量空间中各顶点的分布, 它可以容易地利用机器学习算法进行聚类. 以上这种对应关系使得原来网络拓扑结构中

的各种离散属性关系能够清楚地呈现在连续的向量空间中.

一个较好的网络表示学习方法应具有以下的特点:

① 适应性: 真实的复杂信息网络在不断演化, 新的网络关系不应要求一遍又一遍重复学习同样的过程.

② 社区意识: 潜在特征维度之间的距离可以用于网络中对应成员之间的社会相似性的度量, 这允许在具有同质网络中的泛化作用.

③ 低维: 当标签数据稀缺时, 低维模型可以更好地推广并加速收敛和推断.

④ 连续: 需要潜在表示学习来建模连续空间中的部分社区成员. 除了提供社区成员的一种不同视角之外, 连续的表示学习在社区之间具有平滑的决策边界, 这使得分类更具鲁棒性.

在网络中各个相互连结的顶点之间常常存在依赖关系. 以网页分类任务为例, 通常把一个网页作为一个文档, 然后对文档进行文本主题提取进而对网页进行分类, 在这个过程中往往忽视了网页之间的重要链接关系. 现实世界的网络存在同质性, 网络中相似的结点可能具有一定的联系, 这说明有可能通过网络中的链接信息获得更好的网络特征表示. 在本文中, 不加区分的话通常说的网络表示学习是指用顶点的特征表示来替代网络特征表示. Chen 等人^[41]给出了网络表示学习用顶点的特征表示来表达的优势在于: 首先, 将表示学习得到的顶点作为机器学习算法的输入特征, 避免因网络数据变化而改变相应的机器学习方法; 其次, 直接对网络数据进行分析相当困难, 因为对于数据分析很重要的一些概念如距离、内积等在网络中很难定义. 使用特征表示网络中的顶点, 能够在相应的特征空间中使用各种计算; 最后, 大规模网络数据中各顶点之间的链接关系十分复杂, 通过在低维特征空间中进行可视化分析, 能够直观地观察到各顶点之间的关系.

3 大规模复杂信息网络表示学习

大规模复杂信息网络表示学习能够有效缓解复杂信息网络的数据稀疏性, 实现复杂信息网络异质信息融合以及显著提升相关拓展应用的高效计算, 因而在理论研究以及具体应用上都具有非常重要的意义. 目前, 大规模复杂信息网络表示学习的研究工作主要分为以下几类: 包括基于概率的网络内容表

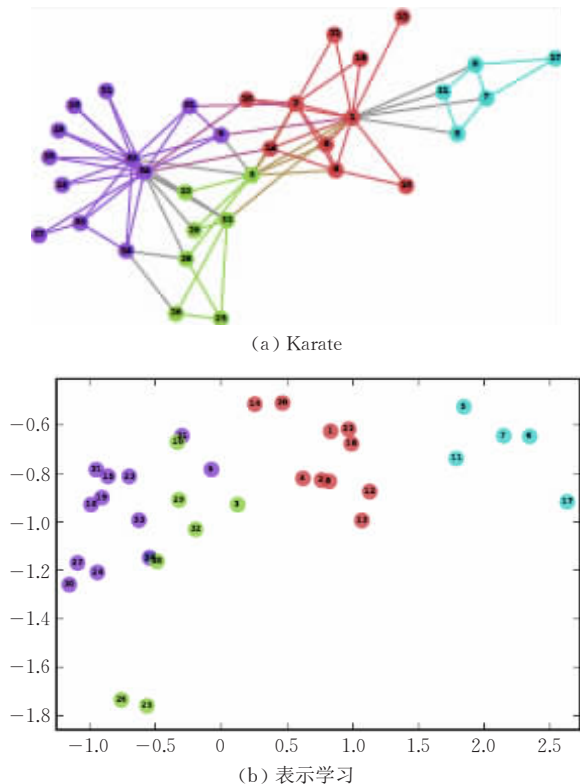


图 1 网络表示学习示例图

示学习、经典的网络表示学习,以及最近提出的大规模网络结构的表示学习、结构-内容融合的表示学习和异构网络的表示学习.下面详细介绍这些方法.

3.1 经典网络表示学习

本文把图嵌入学习算法应用在复杂信息网络上进行的特征学习的这一类算法统称为经典的网络表示学习方法.这些方法首先将网络转换成矩阵表示,然后通过求解矩阵特征向量的形式,进行降维以获取网络的低维表达.在本文中把经典网络表示学习分为基于谱方法的网络表示学习和基于最优化的网络表示学习两类.

3.1.1 基于谱方法的网络表示学习

基于谱方法的网络表示学习是一种直接从矩阵特征值等角度出发进行网络特征学习的一类算法.谱方法可用于获取数据的低维表达^[42],它们广泛应用于工程、应用数学、统计学和计算机科学等领域,以处理“离散”以及“连续”问题.针对网络表示学习,数据矩阵有可能以显式矩阵(例如邻接矩阵)或隐式矩阵(例如文档-词矩阵、超链接结构、对象特征表示、网络路径等)作为其输入.代表性的谱算法有主成分分析 PCA(Principal Components Analysis)算法^[43]和奇异值分解 SVD(Singular Value Decomposition)算法^[7].

如果一个矩阵的行被视为高维空间中的点,其中列表示坐标,则 PCA /SVD 通常用于减少这些点的维度,并且解决低维空间中的目标问题.这种映射的计算优势很明显.此外,PCA /SVD 算法通常还能够突出显示数据中的隐含结构.虽然可以把网络转化为邻接矩阵,作为 PCA /SVD 的输入以获得顶点的低维表示,但通常这种表示的质量较差^[44].本文除了介绍以上谱算法之外,对其它几个流行的谱算法也进行分析讨论.

与局部降维的聚类方法不同,Roweis 等人^[8]提出一种无监督学习的局部线性嵌入算法 LLE(Locally Linear Embedding).LLE 将其输入映射到具有较低维度的单个全局坐标系,并且优化不涉及局部最小值问题.通过利用线性重建的局部对称性,LLE 能够学习非线性流形的全局结构.LLE 算法以网络的邻接矩阵作为初始数据输入,计算得到各个顶点的局部重建权重值矩阵后,将问题最后归结为矩阵特征值求解,从而获取顶点的低维向量表示.

与此类似,Tenenbaum 等人^[9]通过分析高维流形,找到与之对应的低维嵌入,提出了 Isomap 算法.在计算高维流形上顶点之间距离时,使用测地线

距离代替传统的欧式距离,提出了一种用实际输入数据估计其测地线距离的算法. Belkin 等人^[10]根据网络中邻接的顶点在降维后的空间中距离接近,提出了拉普拉斯特征映射(Laplacian Eigenmaps)算法.例如,网络中两个顶点 i 和 j 如果很相似,那么 i 和 j 在降维后的空间中会非常靠近. Laplacian Eigenmaps 能够反映出数据内在的流形结构,通过构建邻接矩阵作为输入,来重构数据流形的局部结构特征,最终选取 Laplacian 矩阵的最小的 t 个非零特征值对应的特征向量来表示学习到的网络.为了解决局部模型的数量而不是原始数据点数量比例的问题,Teh 等人^[45]提出了局部线性协调 LLC(Locally Linear Coordination)算法,通过几个局部降维专家学习使得不同的内部表示映射到用于原始数据空间的单个且一致的全局坐标系中.该算法可以应用于任何专家组,同时,每个专家组会产生高维输入的低维局部表示.与需要修改目标函数的模型不一样,LLC 算法使用一种有效的特征分解器对训练后的模型进行后处理.这也使得它比无模型算法(例如 Isomap 或 LLE)更有效.

为了使得输入图的嵌入是低维表示并且保留图全局拓扑属性,Shaw 等人^[46]提出在欧式空间中嵌入图的结构保留嵌入方法(Structure Preserving Embedding, SPE). SPE 被归结为一个半定规划问题,其学习由一组线性不等式约束的低秩核矩阵,用于捕获输入图的链接结构.传统的图嵌入算法不能依据作者提出的要求保留图结构,因此产生的可视化可能会产生误导或信息不足. SPE 在图的可视化和无损压缩方面获得明显改善,优于拉普拉斯特征映射等方法. SPE 可以仅使用几个维度对图或网络进行合理嵌入,并且将结构保留约束引入降维算法从而产生高维数据的更准确的学习表示.

然而, SPE 方法虽然能够保留网络在低维嵌入时的相邻结构,但是 SPE 算法的复杂度较高,并且它被归结为一个半定规划问题,导致对于大规模网络的计算代价非常大.因此, Talwalkar 等人^[47]研究了基于采样的低秩近似技术,以此解决应用于大密度核矩阵时的效率问题.作者分析了两种常见的近似奇异值分解技术即 Nystrom 和 Column 采样方法.首先在理论上先对这两种方法进行比较,然后提供了这些方法对于不同任务适用性的见解,并同时在相关的数据集上进行了一定的对比实验.

3.1.2 基于最优化的网络表示学习

这类表示学习算法是指事先设定一个优化目标

函数,其参数设置为顶点在低维空间的向量形式,对目标函数进行最大化或最小化优化处理,最终得出网络中的顶点在低维空间的向量表示.例如,Chen 等人^[41]提出了有向图嵌入 DGE(Directed Graph Embedding)算法,其主要利用了转移概率与马尔可夫随机游走的思想.有向图嵌入算法定义了优化目标函数如下:

$$\sum_i T_V(i) \sum_{j, i \rightarrow j} T_E(i, j) (y_i - y_j)^2 \quad (1)$$

其中 y_i 是嵌入在一维空间中顶点 i 的坐标, $T_E(i, j)$ 代表两个顶点 i 和 j 之间的有向边的重要性, $T_V(i)$ 则用于衡量顶点在图中的重要性.

对目标函数最小化以此获得图在一维空间上的优化嵌入.嵌入过程中考虑了顶点对的局部关系和顶点的全局相对重要性.如果 DGE 算法应用于无向网络,则该算法等价于 Laplacian Eigenmaps 算法.图 2 给出了 DGE 算法把数据集 WebKB 降维映射到二维空间的效果图,3 种不同颜色的顶点由德克萨斯等大学的 2883 个网页组成.

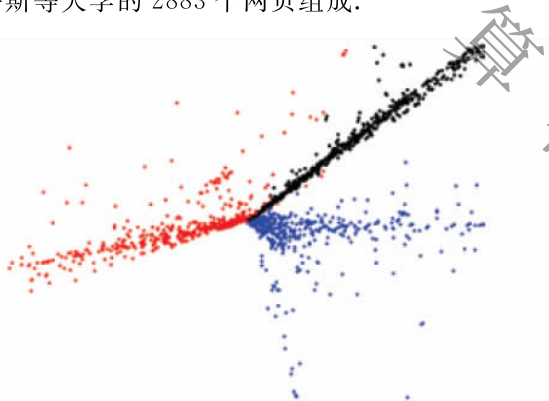


图 2 WebKB 数据降维效果图

此外, Borg 等人^[48]提出的多维量表算法 MDS (Multi-dimensional Scaling) 将网络的顶点映射到一个低维的欧氏空间,使得在新空间中可以保持网络顶点的相似性.而这个相似性可基于网络连通性计算得到^[11].在网络表示学习过程中, MDS 算法的输入数据为一个距离矩阵 $\mathbf{P} \in \mathbb{R}^{n \times n}$, 其中元素 p_{ij} 代表网络中顶点 i 和 j 之间的距离.用 $\mathbf{S} \in \mathbb{R}^{n \times l}$ 表示顶点在 l 维空间的坐标且 $\mathbf{S}$ 的列是正交的,有如下公式:

$$\mathbf{SS}^T \approx -\frac{1}{2} \left(\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^T \right) (\mathbf{P} \cdot \mathbf{P}) \left(\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^T \right) = \tilde{\mathbf{P}} \quad (2)$$

在式(2)中 $\mathbf{I}$ 表示单位矩阵, $\mathbf{1}$ 代表每个元素都是 1 的一个 n 维列向量,而 $\cdot$ 表示矩阵按元素进行相乘. $\mathbf{S}$ 可通过 $\min \|\mathbf{SS}^T - \tilde{\mathbf{P}}\|_F^2$ 获得,使得 $\tilde{\mathbf{P}}$ 与 $\mathbf{SS}^T$ 的差最小.假设 V 包含 $\tilde{\mathbf{P}}$ 中的 l 个最大特征值所对

应的特征向量, $\mathbf{A}$ 代表 l 个特征值所组成的对角矩阵,即表示为 $\mathbf{A} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_l)$, 则 $\mathbf{S}$ 的最优值为 $\mathbf{S} = \mathbf{VA}^{1/2}$. 最终 MDS 算法学习到一个矩阵的特征向量作为网络的特征表示,进一步采用传统的挖掘算法实施社区的划分.本质上这种方法是从社区发现的角度来表达网络表示学习.

具体化到特定类型的网络环境,以社会网络为例, Tang 等人^[49]提出了潜在社会维度的关系学习模型(见图 3 所示).该模型首先基于网络信息提取潜在的社会维度,然后利用它们作为判别式学习的特征.这些社会维度描述了隐藏在网络中的社会行为者的不同隶属关系,并且随后的判别式学习可以自动确定哪些关联更好地与类别标签对应.当多个不同的关系与同一网络相关联时,这是一种优选方案.不同社区的权重大小由学习到的网络特征向量的不同维度来刻画.模型中的目标函数是期望最大化模块度^[50],选取模块度矩阵的 l 个最大的特征向量作为网络特征表示^[51].

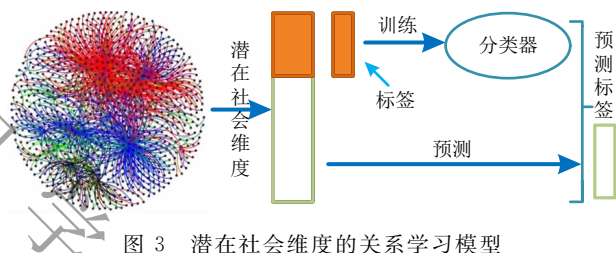


图 3 潜在社会维度的关系学习模型

在动态与静态关系建模方面, Sarkar 等人^[52]探讨了社会网络建模的两个问题,首先提出将一个静态关系模型推广为一个动态模型,考虑了随着时间的推移而变化的因素;其次展示了在网络中的顶点数目较多时如何使该动态模型易于从数据中进行表示学习.该动态模型将每个顶点与 k 维欧氏潜在空间中的数据点相关联.欧氏空间中数据点可以随着时间的推移而移动,但移动的范围有限.如果顶点在潜在空间中接近,则顶点之间更有可能存在链接.文中通过对潜在空间中的相似度使用合适的核函数、低维 kd 树以及非线性局部优化的有效共轭梯度更新规则等来展示如何使模型更易于处理.

基于最优化的网络表示学习在特定的网络分析任务里也有表达,如在网络信息传播预测中依据用户的行为模式来探索信息传播的机制.其实质就是建立一个由用户组成的网络,通过在该网络上建立传播模型模拟信息扩散的动态过程. Yang 等人^[53]指出社交媒体是生产和传播实时信息的中心领域,即使这种信息流传统上被认为是社交网络上的传播

过程,但是基本现象是众多参与者之间的复杂的互动网络的结果.同时作者还提出了一种线性影响模型,这种模型通过预测哪个顶点将影响网络中的其他顶点来建立传播过程.他们建模一个顶点对于通过网络的传播速率的全局影响,并且模拟了新感染顶点的数量作为过去已感染的其他顶点的函数.对于每个顶点,估计一个影响函数,用于量化随着时间的推移有多少被感染的顶点可能受到该顶点的影响.线性影响模型准确地模拟了顶点在传播过程中的影响,可靠地预测信息传播的时序动态性,发现各个参与者的影响模式根据顶点的类型和信息的主题差异十分显著.

考虑到社交媒体上信息的时序传播主要被视为对已知图或近似结构的传播过程,Bourigault 等人^[54]提出了一个新的解决方案,其目标是学习观察到的时间动态到连续空间上的映射.参与传播级联的顶点被投影在潜在表示空间中,使得可以使用热扩散过程来有效地对信息传播进行建模.这相当于学习扩散核,其中投影空间中的顶点的邻近度反应了它们在级联中的感染时间的邻近度.与现有方法相比,所提出的方法具有以下独特的特征:其参数直接从级联样本学习而不需要任何附加信息,因此其不依赖于任何预先存在的传播结构;传播方程的解在投影空间中以封闭形式表示,所以与离散模型相比,用于预测新信息的传播的推理时间大大减少.

以上经典网络表示学习模型往往通过网络的邻接矩阵或关联矩阵采用降维分析的学习方法.通常这类学习只适合应用在较小规模静态网络上.在缺乏网络顶点内在的信息且顶点数量较多的情况下,经典网络表示学习模型效果不尽理想,并且其复杂度通常是网络顶点数量的二次方,使得不能有效地处理大规模复杂信息网络^[21].

为了解决当前大规模复杂信息网络所带来的新特征,包括巨量性、动态性、丰富媒体性等.近年来学者基于深度学习或相关领域的研究启发,提出了一系列的基于网络结构、网络顶点内容属性或两者之间的融合方法,为当前大规模复杂信息网络的特征抽取提供了一系列的解决方案.

3.2 大规模网络结构的表示学习

大规模网络结构表示学习方法通过仅考虑网络顶点之间的链接关系来进行特征抽取,即基于 $G=(V,E,W)$ 形式,其学习的目的在于从网络数据中学习得到任一个顶点的低维向量表示,用于克服经典网络表示学习无法适应大规模网络的问题(例如,大

量顶点带来的计算维数灾难问题).这种低维度的向量表示对大规模(如顶点数以亿计算)的网络中进行相关的操作,如聚类、分类、顶点相似性匹配等等都具有良好的效率及适应性.同时,在动态性方面,通过针对性的在顶点特征的学习过程中设计相关的增量式局部搜索算法,如 DNPS 模型^[40],可以较好地解决网络动态增长所带来的重复性特征学习及其衍生的计算效率较低等问题.

近些年来,深度学习技术在图像识别、语音处理以及自然语言处理等多个领域取得了巨大的成功^[55],如在自然语言处理领域里基于神经网络的语义空间模型的文本分布式特征表达的相关模型得到了广泛的研究并取得了很大的进展^[56-57],这些模型通过将文本中词语的语法或者语义特征映射到一个固定维度的连续向量空间中以解决原来方法所存在的词语矩阵所带来的稀疏性问题以及维数灾难性问题^[58].

本质上,深度学习技术是一种特征学习方法.不同于特征工程需要人为地抽取特征,深度学习可以主动学习数据的特征表示.深度学习通过融合低层特征来实现数据的分布式特征表示.受到深度学习技术广泛应用于不同领域的启发,针对大规模网络结构特征的表示学习算法相继被提出.其中最具代表性的网络结构表示学习模型为 DeepWalk^[31],该模型借鉴了神经网络语言模型 Word2Vec^[59]的思想,将文本中词语间的关系转换为复杂信息网络中顶点之间的关系.

本文介绍有关复杂信息网络结构表示学习模型之前,先简要回顾神经网络语言模型的发展过程.Hinton 等人^[60]最早提出了学习分布式表示思想,然而利用神经网络语言模型学习词汇的向量表示则由 Bengio 等人^[61]提出.他们在文中设计了一个多层神经网络语言模型(Neural Network Language Model, NNLM),通过一个可以在文本语料中滑动的长度为 n 的窗口,用前 $n-1$ 词预测窗口内的当前词的概率. NNLM 模型由输入层、投影层、隐含层和输出层组成,其中输入层为滑动窗口内前 $n-1$ 个词的向量表示,把这些向量拼接起来则发生在投影层,隐含层则将投影层的输出结果进行非线性变换,最后每个词在下个位置的出现概率则由输出层显示.求解该模型的方法则使用随机梯度下降和反向传播算法.由于输出层 softmax 函数的计算复杂度和词汇表大小同阶,因此这种前馈神经网络语言模型的复杂度较高.

在后续的研究中,有相关文献替换或优化处理

NNLM 模型中的 Softmax 函数, 如 Morin 等人^[62]提出了层次 softmax 函数, 将模型的输出层表示为一棵哈夫曼树; Collobert 等人^[63]则直接用 hinge loss 函数替换原来的函数; Gutmann 等人^[64]提出了一种用于降低 NNLM 的时间复杂度的算法. Mikolov 等人提出的词向量训练模型 Word2Vec, 由于去掉了前馈神经网络中的中间层, 从而极大地提升了训练效果. 该模型实质上由两个子模型组成: CBOW 和 Skip-Gram 模型, 前者是利用窗口的上下文来预测中心词, 而后者刚好相反, 是用窗口的中心词来预测这个词的上下文. Word2Vec 模型可用于大规模的文本连续性表达学习, 在很多应用上取得了较好的结果.

根据模型学习方式的不同, 本文把大规模复杂信息网络结构的表示学习模型分为高阶关系建模学习、半监督表示学习, 以及具有可伸缩性的特性表示学习.

3.2.1 基于高阶关系的表示学习

高阶关系建模学习是把网络顶点关系的相似性从一阶拓展到高阶. 对各阶关系采用不一样的目标函数, 然后将各阶关系获取的分布式表示进行拼接, 进而获得顶点特征学习. 具有代表性的高阶关系建模学习模型包括 DeepWalk、LINE、GraRep 等模型.

Word2Vec 在训练词向量时, 以文本语料作为输入数据. 而网络表示学习则以复杂信息网络作为数据输入, 两者看上去没有任何关联. 然而 DeepWalk 模型的作者注意到, 在训练过程中词语出现的频次与根据原始网络结构进行随机游走时顶点被访问到的次数两者皆服从幂律分布. 因此该文作者将网络中的顶点看作是文本语料中的词汇, 把进行随机游走时获得的游走路径当成语料中的句子, 通过这种方式获得的网络数据就可以作为 Word2Vec 的数据输入来达到训练顶点的分布式表达.

DeepWalk 模型的框架示意图如图 4 所示, 其中图 4(a) 显示采用随机游走的方法产生标准的输入序列. 图 4(b) 表示使用 SkipGram 模型对序列建模得到网络顶点的向量表示, 其中 $\mathcal{W}_i$ 代表以顶点 i 为根的随机游走序列, 函数 Φ 把每个顶点映射到一个指定的空间向量. 而图 4(c) 代表分层 softmax 函数, 用于解决顶点高维度输出问题. 相比较基准方法, DeepWalk 模型克服了网络训练数据稀疏性这一问题, 在训练数据较少的情况下也能取得比使用全部数据较好的训练效果. DeepWalk 模型的提出为网络结构特征的向量表示提出了一种新的研究思路.

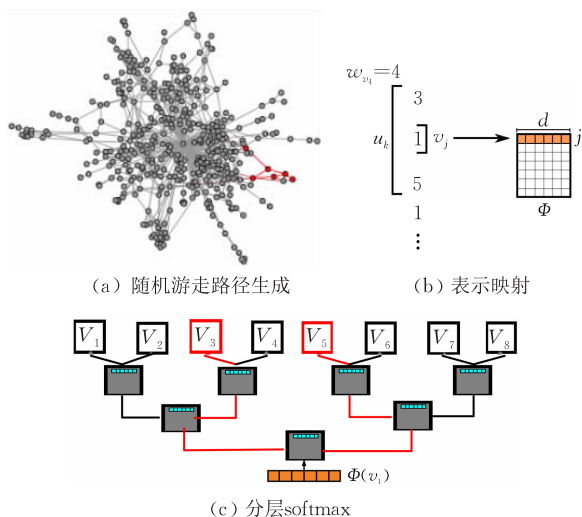


图 4 DeepWalk 模型框架

针对 DeepWalk 模型仅适用于未加权网络, 在此之后, Tang 等人^[21]提出了一种适用于不同类型网络(例如, 无向图、有向图或带权图)的表示学习模型 LINE (Large-scale Information Network Embedding). 该模型从一阶相似性和二阶相似性两个方面设置不同的目标函数. 一阶相似性定义成网络中两个顶点之间的点对相似性, 其值为顶点之间链接的权值(如果点对之间不存在链接, 一阶相似性值则为 0). 为了对一阶相似性建模, 对于每条无向链接 (i, j) , 顶点 i 和 j 之间的联合概率如下:

$$p_1(i, j) = 1 / (1 + \exp(-\mathbf{i}^T \cdot \mathbf{j})) \quad (3)$$

式中 $\mathbf{i} \in \mathbb{R}^d$, $\mathbf{j} \in \mathbb{R}^d$ 分别是顶点 i 和顶点 j 的低维向量表示, 顶点 i 和 j 的经验联合概率为 $\hat{p}_1(i, j) = w_{ij} / \Omega$, $\Omega = \sum_{(i, j) \in E} w_{ij}$, 其中 w_{ij} 为链接 (i, j) 的权值.

网络表示最后通过最小化概率分布 $p_1(\cdot, \cdot)$ 与 $\hat{p}_1(\cdot, \cdot)$ 的 KL 距离来获得. 二阶相似性适用各种类型的网络, 其假设如果顶点间共享相似的邻居顶点, 那么两者就趋于相似. 顶点之间的二阶相似性表示为两个顶点在整个网络上的一阶相似性的分布相似度. 在这种情形下, 每个顶点 i 扮演两个角色: 顶点本身和其他顶点的特定“上下文”, 分别用向量 $\mathbf{i}$ 和 $\mathbf{i}'$ 表示这两个角色. 二阶相似性模型则定义顶点 i 和 j 的条件概率如下:

$$p_2(j|i) = \exp(\mathbf{j}'^T \cdot \mathbf{i}) / \sum_{k=1}^{|V|} \exp(\mathbf{k}'^T \cdot \mathbf{i}) \quad (4)$$

顶点 i 和 j 的经验条件概率 $\hat{p}_2(j|i) = w_{ij} / d_i$, d_i 表示顶点 i 的出度. 二阶相似性模型的网络表示则由最小化概率分布 $p_1(\cdot, \cdot)$ 与 $\hat{p}_1(\cdot, \cdot)$ 的 KL 距离来获得. 基于这两类相似性, LINE 可以分别学习到一阶顶点向量和二阶顶点向量, 对两种顶点向量进行

拼接,从而获得最终的顶点表示.同时,在训练过程中 LINE 对于相似性的表示均采用了基于链接的负采样优化算法.

相比于 LINE 模型中,基于局部的一阶相似性和二阶相似性的特征提取,Cao 等人^[32]提出的 GraRep(Graph Representations)模型在 LINE 基础上进行了更高阶相似性关系的建模,以获取网络顶点的全局特征表示.该模型首先定义一阶概率转移矩阵 $\mathbf{A}=\mathbf{D}^{-1}\mathbf{S}$,式中矩阵元素 a_{ij} 的值是顶点 i 到 j 的一阶转移概率, $\mathbf{D}$ 为度对角矩阵, $\mathbf{S}$ 为邻接矩阵.为了学习到顶点对之间关系的全局表示,需要利用已知的一阶概率转移矩阵来计算 k 阶概率转移矩阵,即通过计算 $\mathbf{A}^k=\mathbf{A}\cdots\mathbf{A}$ 获得.

同样的,受到 SkipGram 模型的启发,GraRep 模型采用 Gutmann 等人^[65]提出的噪声对比估计(Noise Contrastive Estimation, NCE)方法来定义目标函数.优化目标在于最大化 (i,c) 的出现概率并最小化随机生成的 (i,c') 的出现概率,这里的 i 为当前顶点, c 是 i 的上下文顶点, c' 是随机得到的上下文顶点.同样使用负采样的方法建模 k 阶信息,并使用 SVD 矩阵分解的方法得到网络顶点的向量表示. GraRep 模型考虑了更高阶的上下文信息,在网络结构数据的表示学习中起到了很好的效果.虽然在模型中使用了复杂度较高的 SVD 矩阵分解的方法,但实际中可以使用随机梯度下降的优化方法,因而对于大规模的网络结构表示学习该模型同样适用.

与借鉴 Word2Vec 的学习模式相对比的是,Niepert 等人^[66]提出了对任意的网络结构通过卷积神经网络(CNN)进行特征抽取的学习框架.与已有方法相同的是该模型同样既可以针对有向网络,也可以学习无向网络.作者基于传统图像(图片)特征抽取中关于局部关联区域抽取方法的启发,在网络中也采用局部关联的方式进行相关操作.该方法与传统的基于 graph kernel 的模型相比,具有较好的性能.然而,作者在对比较方法的选取上却忽视了当前具有较好计算性能的前沿方法,例如 DeepWalk 以及 LINE 等方法,这也使得该方法的计算性能在大规模网络环境下的可行性需要进一步的验证并给出判定.

针对 DeepWalk、LINE 以及 GraRep 等网络表示学习的方法多基于静态网络或并没有对复杂信息网络的特有属性进行修正学习模式,Li 等人^[40]提出了基于动态阻尼正负采样的网络结构特征嵌入模型 DNPS(Damping based Negative-Positive Sampling

of Social Network Embedding). DNPS 模型通过对不同阶顶点关系构建基于有正负阻尼的顶点共现学习.引入这种学习策略能够有效地把深度特征、广度特征和社区特征等不同顶点特征环境进行统一表示学习.同时,在采样模式中使用局部搜索的增量式算法来更好地适应网络的动态性.由于 DNPS 模型融合了顶点结构特征学习以及复杂信息网络自身特性,因此它是一个具有动态以及大规模适应性的网络表示学习模型. DNPS 模型可以看成一种动态逻辑回归模型,其目标函数的优化采用快速的随机梯度上升的方法.在大规模的网络数据上体现出 DNPS 模型不仅在具有动态变化过程中的结构特征的提取变得可行,而且由此获得的顶点特征表达具备更好的动态鲁棒性.与 DNPS 模型进行动态顶点的特征学习相似,Hamilton 等人^[67]提出了一种 GraphSAGE 模型,该模型能够基于顶点的属性特征,如文本特征等,对顶点在当前时刻 T 下所存在的未知数据(unseen data)顶点进行特征学习.该方法同样基于局部的特征采样来对当前顶点的特征进行聚合.

相比于局部特征,通过基于社区属性这一重要的全局模式出发,不同于现有的网络表示学习方法主要考虑从顶点的局部上下文进行学习表示 Tu 等人^[68]提出了一种新的社区增强型网络表示学习模型 CNRL(Community-enhanced Network Representation Learning). CNRL 模型通过引入顶点的社区信息,学习更具有判别性的网络表示. CNRL 对每阶顶点关系分别建模,最终发现每个顶点的社区分布,同时还学习到顶点以及社区的嵌入. CNRL 模型克服了有些网络表示学习方法仅考虑顶点局部信息的缺陷,并且将所学习的表示应用于网络分析任务时,与 LINE 模型等相比较有着明显的改进.

3.2.2 基于半监督的表示学习

半监督表示学习首先对网络的顶点特征学习采用无监督方式,学习顶点数据完整的“语义”信息,使其在后续的监督学习过程中很好的适应训练数据.这类表示学习在分类预测任务上效果显著.

针对网络表示学习应用于机器学习任务如顶点分类时通常所缺乏的辨别能力,Tu 等人^[69]提出了一种新颖的半监督模型 MMDW(Max-Margin DeepWalk). MMDW 是基于矩阵分解的统一网络表示学习框架,通过优化最大间隔分类器以及目标矩阵因式分解模型来实现.受最大间隔分类器的影响,学习表示不仅包含网络结构,而且具有辨别分类

的特征. MMDW 的学习过程是: 首先把原始 DeepWalk 模型转换为矩阵因素分解形式 $\mathbf{M} = \mathbf{X}^T \mathbf{Y}$, 为了得到矩阵 $\mathbf{X} \in \mathbb{R}^{k \times |V|}$ 和矩阵 $\mathbf{Y} \in \mathbb{R}^{k \times |V|}$, 将式(5)

$$\min_{\mathbf{X}, \mathbf{Y}} L_{DW} = \min_{\mathbf{X}, \mathbf{Y}} \|\mathbf{M} - (\mathbf{X}^T \mathbf{Y})\|_2^2 + \frac{\lambda}{2} (\|\mathbf{X}\|_2^2 + \|\mathbf{Y}\|_2^2) \quad (5)$$

最小化, 式中因子 λ 控制正则化这部分的权重. 然后训练基于最大间隔分类器并扩大支持向量和分类边界之间的距离. MMDW 模型学习到的顶点表示不仅能够反映其网络结构, 而且还能反映顶点的标签信息.

考虑到 DeepWalk 模型中序列的生成方式包括宽度优先搜索(BFS)和深度优先搜索(DFS)的采样方式所学习到网络结构特征的差异性, Grover 等人^[70]提出了用于表示学习网络中顶点连续特征的半监督模型 Node2Vec 模型. 该模型学习顶点到低维特征空间的映射, 从而最大化保留顶点的网络邻域的可能性. Grover 等人发现, BFS 生成的顶点序列多集中于网络中的某个局部结构内, 生成的序列稳定, 学到的顶点表达倾向于表示顶点的直接邻居结构; 而 DFS 能较好地遍历整个网络, 学到的顶点表达可以反应顶点与其周围顶点的结构关系, 但生成的序列不够稳定. 作者通过定义一个关于顶点的网络邻域的概念并设计一个有偏差的随机游走程序, 有效地探索顶点的不同的邻域.

此外, Node2Vec 模型引入 search bias 函数, 用以调整、平衡这两种搜索采样方式. search bias 是一个阶跃函数 $\alpha_{pq}(t, x)$, 根据随机游走的上一遍历记录, 调整下一步的转移概率 $\pi_{ix} = \alpha_{pq}(t, x) \omega_{ix}$, 其中 ω_{ix} 是随机游走初始的转移概率值, p 和 q 是 search bias 的参数, t 是上一遍历的顶点, i 是当前顶点, x 是待遍历顶点. 在不同的任务中通过调整 p 和 q , 可以获得更好的顶点表达学习. 在多标签分类和链路预测两个应用问题上进行验证, 表明了 Node2Vec 模型的有效性.

网络嵌入是学习网络中顶点的低维表示的重要方法, 目的在于捕获和保留网络的结构. 现有的网络嵌入方法几乎都采用浅层模型. 由于隐含的网络结构非常复杂, 浅层模型并不能捕获到高度非线性网络结构从而导致产生非最优网络表示结果. 因此, 如何找到能够有效捕获高度非线性网络结构以及保留结构的全局和局部特性, 仍是一个开放且非常重要的问题. 为了解决这个问题, Wang 等人^[71]提出了一种结构深度网络嵌入方法 SDNE (Structural Deep

Network Embedding Method). 他们首先提出一种半监督深度模型, 该模型由多层非线性函数组成, 用于捕获高度非线性网络结构. 然后利用顶点之间的一阶相似性和二阶相似性共同保留网络结构, 二阶相似性由无监督组件应用来捕获全局网络结构, 而一阶相似性被用于监督组件中作为监督信息以保留局部网络结构. 通过在半监督深度模型中联合优化一阶相似性和二阶相似性, SDNE 能够将数据映射到高度非线性的潜在空间以保留局部和全局网络结构并且对于稀疏网络具有鲁棒性. 在大规模真实网络上实证了 SDNE 对于多标签分类、链接预测以及可视化等应用上效果明显.

3.2.3 基于可伸缩的表示学习

具有可伸缩性 (scalability) 的表示学习是指在网络中以不同层级采样方式对成对顶点进行操作的在线学习方法. 这类方法类似于使用采样的高阶过渡矩阵, 从而使得可扩展到具有数百万个顶点及以上的大规模网络.

为了对多尺度网络进行特征表示学习, DeepWalk 模型的提出者 Perozzi 等人^[72]提出了一种用于学习网络中顶点的多尺度表示学习的方法 Walklets. 这种方法可以清晰地在连续向量空间中对多尺度顶点关系进行编码, 以适用于多标签分类问题. 与以往的研究不同的是, 由 Walklets 生成的潜在特征是可解析导出.

Walklets 通过使用在随机游走过程中观测到的顶点之间的偏移量来学习一系列潜在特征表示, 每个潜在特征表示又可用于捕获连续的且更大的关系. 各种依赖信息使得同一表示策略在不同尺度顶点关系上均可建模. 类似于 DeepWalk 模型, Walklets 方法从每个顶点开始进行一系列截断的随机游走. 在这些截断的随机游走中的两个顶点的共现可以建模网络中的扩散速率. 然而在采样过程中一个至关重要的变化是: 选择跳过随机游走中的一些顶点, 以这种方式形成从邻接矩阵 $\mathbf{A}$ 的连续更高的幂中采样的一组关系. 在大规模网络上的实验验证了 Walklets 方法对多标签网络分类任务的效果优于矩阵分解方法等.

同样的, 考虑到现有的基于顶点相似性的网络链接预测方法采用简单的一阶、多阶邻居信息或针对特定类型的小型网络^[73-75], 导致这些方法在扩展性以及大规模网络中的可计算性都受到了严峻的挑战. 基于此问题, Li 等人^[76]提出了一个 LsNet2Vec (Large-scale Network to Vector) 模型对大规模网

络进行链接预测.

LsNet2Vec 模型首先通过随机游走的方法产生网络数据集序列从而进行大规模的无监督机器学习,使得网络中顶点的结构特征信息被映射到一个连续的、固定维度的实数向量空间上.然后把学习到的顶点结构特征向量用于迅速计算大规模网络中任意顶点之间的相似度.图 5 为 LsNet2Vec 模型的架构图,由输入层、投影层和输出层组成.其中输入层的主要功能是按照一定的规则对网络中的顶点采用随机游走的方法进行遍历得到序列化的格式表示输出,进而重构训练集.投影层则对序列化的训练集根据给定训练窗口大小 α 在嵌入矩阵 $\mathbf{M}$ 中进行搜索,最后将搜索到的向量集合以参数形式传递给聚合函数 f 进行处理.输出层由一棵哈夫曼树构成,选用哈夫曼树对网络顶点进行重构存储以此有效降低计算的复杂度.在多个大规模网络数据集上进行验证并与其它链接预测进行对比,证明了 LsNet2Vec 模型预测的效果是非常有效.

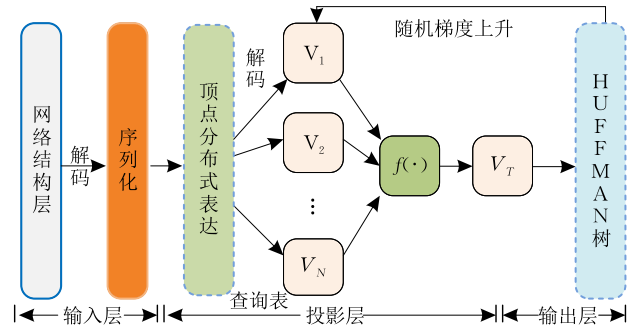


图 5 LsNet2Vec 模型架构

表 1 汇总了当前最新的大规模复杂信息网络结构表示学习模型,并比较了这些模型的核心思想与观点、实验环境以及实验中数据集的规模大小以及应用方面等.表中的模型从一定程度上都是受到 Word2Vec 模型的启发.其中 DeepWalk 模型是最先提出在大规模复杂信息网络中通过训练顶点潜在向量表示的新方法.其它大多数模型是在 DeepWalk 模型的基础上进行扩展和应用,但在模型的表示学习能力以及应用范围存在着差别.

表 1 大规模复杂信息网络结构表示学习模型

模型	核心算法与观点	对象	类型	实验规模	实验环境	评测数据集	评测应用	备注
DeepWalk ^[31] (2014)	※随机游走产生输入序列 ※基于 Skip-gram 模型	node	UD/ NW	1 138 499	单机, 24 Cores @2.0 GHz CPU, 128 GB 内存	社交网络	顶点的多标签分类	适合大规模网络
LINE ^[21] (2015)	※重构目标函数 ※带权边采样算法	node	D/UD/ W/NW	1 985 098	单机, 40 Cores @2.0 GHz CPU, 1 TB 内存	文本/社交/ 引用网络	词类比、文档分类、可视化	适合大规模网络
GraRep ^[32] (2015)	※学习网络的全局特征 ※优化部分使用 SVD	node	W/NW /UD	10 312	单机, 4 Cores @3.4 GHz CPU, 16 GB 内存	语言/社交/ 引用网络	顶点聚类、顶点分类、可视化	优化部分若使用随机梯度下降则适合大规模网络
DNPS ^[40] (2016)	※基于阻尼衰减的采样算法 ※基于局部搜索增量式学习 ※优化部分使用随机梯度上升	node	D/UD	3 223 589	单机, 6 Cores @IntelRi7-5280k CPU, 64 GB 内存	社交网络	链接预测、社区发现、用户推荐、标记分类	适合大规模网络
CNRI ^[68] (2016)	※基于社区增强型 DeepWalk 模型 ※增强型顶点表示	node	W/NW /UD	10 312	n/a	社交网络/ 引用网络/ Web 网络	链接预测、顶点分类、社区发现、可视化	统一建模局部邻域信息和全局社区结构
MMDW ^[69] (2016)	※基于最大间隔 DeepWalk 模型 ※优化部分使用矩阵因式分解的平方损失	node	D/UD	3312	单机, 4 Cores CPU, 16 GB 内存	引用网络/ 社交网络	顶点分类、可视化	半监督学习
Node2Vec ^[70] (2016)	※引入 search bias 函数进行有偏差的随机游走 ※网络邻域	node	D/UD	1 000 000	单机, 6 Cores @3.4 GHz CPU, 64 GB 内存	社交网络/ 生物网络/ 词共现网络	链接预测、顶点的多标签分类	适合大规模网络/半监督学习
SDNE ^[71] (2016)	※半监督深度模型 ※使用随机梯度下降进行参数寻优	node	D/UD	1 138 499	n/a	社交网络/ 引用网络/ 语言网络	顶点的多标签分类、链接预测、可视化	适合大规模网络/半监督学习
Walklets ^[72] (2016)	※多尺度表示学习 ※使用随机梯度下降进行参数寻优	node	D/UD/ NW	1 138 499	单机, 24 Cores @2.0 GHz CPU, 384 GB 内存	社交网络/ 引用网络	顶点的多标签分类	适合大规模网络
LsNet2Vec ^[76] (2016)	※随机游走产生输入序列 ※使用随机梯度上升对参数寻优	node	D/UD	2 390 000	单机, 4 Cores @2.5 GHz CPU, 16 GB 内存	公路/社交/ 引用/通讯/ 购买共现网络等	链接预测	适合大规模网络/无监督学习

* : D/UD; 向/无向图; W/NW; 带权/非带权图; 实验规模为作者所采用的实验数据集的最大顶点数;
实验环境: 上述表格中部分作者未在论文中提及实验时所使用机器配置情况, 笔者通过邮件联系上部分作者, 得到了相关的信息, 在上述表格中进行了相应的补充, 在此一并致谢.

3.3 基于概率的网络内容表示学习

与 DeepWalk 及其衍生方法进行对比, 基于贝叶斯公式的概率模型也成为网络内容表示学习的热点之一, 即基于 $G=(V, E, W, C)$ 形式. 由于复杂信息网络中的数据信息除了顶点间的结构关系外, 其顶点自身还可能包含大量有用的属性信息, 如互联网中网页顶点包含的文本内容、社交网络中每个用户发布的文本信息等.

基于概率的网络内容表示学习以顶点的内容信息作为输入数据, 学习得到网络中的顶点在主题空间上的向量表示. 在学习过程中主要采用基于概率生成模型方法来实现网络表示学习, 即用一个基于概率的生成过程去建模网络数据的生成过程. 求解这些模型的方法通常采用 Gibbs 采样、变分推断和期望最大化算法等. 具体的, 概率生成模型方法主要用于处理文本数据, 通常使用无监督的分级贝叶斯模型来实现, 在不考虑网络结构特征的情况下, 将文本数据与网络中的链接关联起来. 其实现主要依赖主题模型, 该模型首先假设存在一个概率生成式模型, 将每一个文本关联到多个主题后, 最终从文本中提取相应的主题. 目前的概率主题模型一般基于这样的思想: 将文本看作是由若干主题随机混合, 在不同的模型中会有不同的统计假设, 并以不同的方法获得模型参数. 本节主要介绍静态模式表示学习和动态模式表示学习两类框架下的网络内容表示学习模型.

3.3.1 静态模式表示学习

在主题模型中, 将一个给定词汇的多项式分布表示成一个主题. 主题模型用一个数量较小词汇的分布对文本进行总结, 这些分布被称为“主题”^[77]. 传统的主题生成模型包括 PLSA^[78] (Probability Latent Semantic Analysis)、LDA^[79] (Latent Dirichlet Allocation) 等. 在 PLSA 模型中, 通过用概率的方式解释了文档的生成过程. 该模型是建立在 LSA (Latent Semantic Analysis) 模型的基础上, 用于克服 LSA 无法解决一词多义从而导致表达词义的能力有限的问题, 但该模型中的文档概率值与特定文档相关联, 缺乏有效处理新文档的自然方法. 同时随着文档数量的增加, 参数估计的数量呈线性增长, 这表明模型易于过拟合^[80]. LDA 模型在 PLSA 模型的概率分布加上了先验知识, 采用分层贝叶斯推断方法求解参数, 解决了 PLSA 存在的过拟合问题. 在 LDA 模型中, 通过允许文档作者确定不同主题的混合权重, 并将其扩展到作者建模. 通过对模型

的参数学习, 得到一组出现在语料库和相关的不同文档的主题集合, 识别出作者及相关的主题. 在 LDA 模型中文档集合的生成过程可描述为: 首先从狄利克雷分布中抽取每篇文档的主题分布, 然后根据主题分布将文档中的每个词汇选择一个对应的主题, 最后从具有特定主题的词汇多项式分布中, 对每个词汇进行采样.

在主题模型 LDA 的基础上, Chang 等人^[81] 进一步提出了关系主题模型 RTM (Relational Topic Model). 这是一种结合网络链接和顶点属性的分层模型. RTM 模型在建模文档的生成过程时与 LDA 基本一致, 不同之处在于前者还建模链接关系的产生. 该模型给出假设是: 若两个文档顶点之间存在链接, 那么它们在主题上的分布应该理论上更相似. RTM 的图模型如图 6(a) 所示, 图中每个文档 d_i 与一个主题分布 θ_i 相关联, 为了在文档 d_i 中产生第 n 个词, 它们首先从 θ_i 中选择一个主题 $z_{i,n}$, 然后从 $z_{i,n}$ 的主题多项式 $\beta_{z_{i,n}}$ 选择一个词 $w_{i,n}$. θ_i 和 β_z 各自的 Dirichlet 先验为 α 和 λ . 相应地, 一对文档 d_i 和 d_j 之间的每条链接 y_{ij} 是基于在 d_i 和 d_j 中出现的主题由链接概率函数生成. 文档之间分享共同主题数量越多, 他们之间的链接就越有可能.

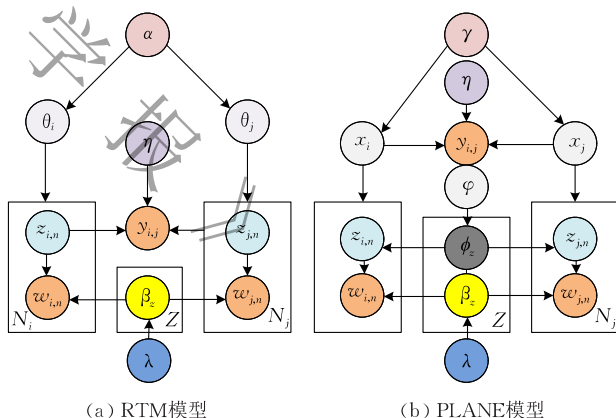


图 6 RTM 和 PLANE 的图模型比较

由于 PLSA 和 LDA 模型不能明确建模引用文档和被引用文档之间的主题关系, Nallapati 等人^[82] 提出了 Link-PLSA-LDA 模型, 该模型以学术论文引用网络为输入数据, 先后分别建模被引用论文集合和引用论文集合生成过程. 图 7 展示了这种建模的过程. 为了明确地建模从引用论文集合到被引用论文集合的信息流, 他们为被引用论文的内容定义了一个使用相同分布 Ω 的显式生成过程. 在这个新的生成过程中, 将被引用论文集合当成用词汇填充的容器, 首先关联整个被引用论文集合的主题混合

分布 π , 然后将词汇填充到这些容器中。

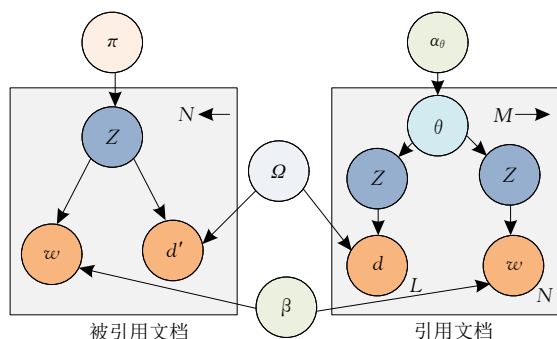


图 7 Link-PLSA-LDA 模型

可视化是理解复杂和高维数据的有用工具, 它使人们能够通过大量数据直观地进行浏览。大多数可视化方法没有考虑给定数据中的潜在结构, 例如在文档数据情况下的主题。为了克服上述局限性, Iwata 等人^[83]提出了 PLSV (Probabilistic Latent Semantic Visualization) 模型, 这是一种基于主题模型的离散数据(如文档)的可视化方法。与基于成对距离的常规可视化方法(如多维量表)不同, 文中考虑从可视化空间到文档空间的映射作为文档的生成过程。在模型中, 假设文档和主题在二维或三维欧氏空间或可视化空间中具有潜在坐标。文档的主题分布由文档和可视化空间中的主题之间的距离确定, 并且每个词汇依据主题分布从一个主题中提取。可以通过使用 EM 算法 (Expectation Maximization Algorithm) 将模型拟合到给定的一组文档来获得可视化, 即文档的潜在坐标, 使具有相似主题的文档被紧密地嵌入在一起。

文本网络的嵌入问题是把文本的高维表示嵌入到一个低维空间中, 以尽可能保留原始文本数据的属性。得到文本的低维表达具有重要的应用, 如文本在低维空间的可视化。通过将文本降维表示解释为二维或三维空间上的坐标, 可以产生一个散点图可视化。文档之间的相似性或差异性体现在空间信息上。除了可视化应用之外, 低维表示还可以用于有损压缩或者用于机器学习任务(例如聚类或分类)中的特征选择。基于以上这些应用, Le 等人^[84]提出了 PLANE (Probabilistic Latent Document Network Embedding) 模型, 通过对 RTM 模型进行扩展, 从可视化的角度学习到主题和文档顶点的低维表达。图 6(b) 为 PLANE 模型表示图, 与 RTM 模型的区别在于 PLANE 模型需要考虑低秩嵌入目标并且基于坐标而不是主题分布来建模链接的生成过程。PLANE 模型的具体生成过程主要涉及两个方面:

一是传统的主题建模; 二是表示学习, 对每个主题和每个文本顶点都学习其在低维空间上的坐标。

3.3.2 动态模式表示学习

针对以上主题模型没有考虑到网络文档的时间戳的重要性, Wang 等人^[85]提出了一个与 LDA 类似的主题模型 TOT (Topics Over Time)。该模型不仅能够捕捉网络数据的低维结构表达, 而且还能够捕获该结构随时间的变化过程。与依赖于马尔科夫假设或时间离散化的其他研究工作不同的是, 文中的每个主题与在时间戳上的连续分布相关联。对于每个生成的文档, 主题上的混合分布受到词共现和文档的时间戳两者的影响。

除此之外, 部分研究人员对信息传播主题提取也展开了相关研究。由于大多数信息传播提取工作侧重于个人层面的交互以及网络的结构拓扑^[86-88], 虽然这些提取方法获得了较好的效果, 但仍然存在如下缺点: 首先基于网络结构方法很大程度上忽略了局部的差异, 导致不能捕捉到信息模式的多样性; 高度易变的用户行为通常使得难以准确地揭示个人层面的传播模式; 最后聚合方法未能揭示详细的传播过程。

针对信息传播提取存在的以上问题, Hu 等人^[89]提出了 COLD (Community Level Diffusion) 方法用于提取社区层面的传播, 即建模不同社区的主题的传播模式, 将社区和主题作为潜在变量, 并为观察网络、文本和时间建立一个生成过程, 以准确地刻画社区层面的主题传播, 同时设计了一种基于抽样的推理算法及其并行实现方法。在学习顶点在社区空间上的分布的同时, 也学习社区在主题空间上的分布。

3.4 结构-内容融合的表达学习

表示学习在很多应用方面展示出其有效性, 如图像分类和文本挖掘中。网络表示学习旨在学习网络中每个顶点的分布式向量表示, 这也越来越多地被确认为网络分析的一个重要方面。和以往网络分析关注网络的拓扑结构解决诸如“网络的直径”^[90]、“网络如何演化”^[91-95]、“信息如何在网络上传播”^[96-97]以及“什么是网络上的社区”^[98-103]等问题的研究方式不同, 网络表示学习通过对网络顶点之间的关系(结构或内容)进行分析, 得出网络特征的低维度表示。大多数网络表示学习方法通过研究网络结构特征进行学习, 忽视了在现实中网络顶点包含的丰富信息(文本内容和其它元数据)。

以维基百科为例, 其文章相互链接形成一个大

型网络,该网络中的每个顶点即每篇文章包含丰富的文本信息,这些信息对于网络表示学习也可能很重要. 还有些网络表示学习方法从研究网络内容特征着手. 这些方法通常以文本网络作为输入数据,学习到的都是网络中的顶点在主题空间上的向量表示,但没有考虑到网络本身所具有的结构特征. 事实上这些网络结构特征对于提炼和构建主题模型是非常有用的,甚至有时是发现网络相关主题的必要条件. 例如,两个经常相互合作的研究人员可能正在处理相同的主题,因此很可能属于同一个研究团体;对于地理敏感的事件,生活在近邻地方的博主会写相似的主题;缺少网络结构的考虑也是一些其它文本挖掘技术(如文档聚类)中的缺陷. 因此,融合网络结构特征和文本内容特征的学习网络表示已经成为当前的重要研究方向.

为了解决上述问题,一种较为直观学习方法是分别对文本特征以及网络的结构特征进行表示学习,例如分别用 Word2Vec 以及 DeepWalk 方法,然后把这两种表示学习结合起来. 然而,这类方法没有考虑到网络结构和文本信息两者之间的复杂交互,因此通常学习得到的特征向量并不理想. 另一种有意义的学习方法是:在已有的网络结构表示学习框架中同时嵌入文本内容进行学习. 目前,结构-内容融合表示学习研究工作主要分为基于矩阵因子分解的方法和基于概率图模型方法.

3.4.1 基于矩阵因子分解的方法

基于矩阵因子分解方法将大规模复杂信息网络建模为矩阵形式,然后用线性代数或图论的矩阵分解法解决表示学习问题.

DeepWalk 模型仅考虑顶点的结构信息,忽视了顶点具有的丰富的内容信息,于是 Yang 等人^[104]在 DeepWalk 模型的基础上通过考虑顶点的内容信息提出了 TADW (Text-Associated DeepWalk) 方法. 该方法将 DeepWalk 中矩阵 $\mathbf{M}$ 因式分解成三个低维矩阵的乘积: $\mathbf{E} \in \mathbb{R}^{k \times |V|}$ 、 $\mathbf{H} \in \mathbb{R}^{k \times f_t}$ 和 $\mathbf{T} \in \mathbb{R}^{k \times f_t}$, 矩阵元素 m_{ij} 代表顶点 i 在固定步骤中随机地游走到顶点 j 的平均概率的对数, $\mathbf{T}$ 作为顶点的文本特征矩阵. 图 8 显示了 TADW 方法不同于 DeepWalk 的矩阵分解示意图,其中图 8(a)表明 DeepWalk 中矩阵 $\mathbf{M}$ 被分解为两个矩阵的乘积,连接这两个矩阵后作为最终的顶点向量表示,求解方法则采用正则化的低秩矩阵分解^[105],而图 8(b)则显示 TADW 方法中通过连接矩阵 $\mathbf{E}$ 、 $\mathbf{H}$ 和 $\mathbf{T}$ 后得到顶点特征矩阵,文本特征矩阵 $\mathbf{T}$ 由 TF-IDF (Term Frequency-Inverse

Document Frequency) 矩阵采用奇异值分解降维得到.

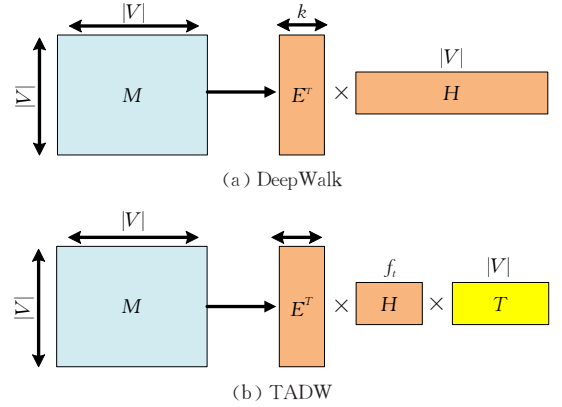


图 8 TADW 的矩阵分解示意图

相比于 DeepWalk 模型, TADW 方法在计算性能上明显占优. 在 DeepWalk 模型中因子分解矩阵 $\mathbf{M}$, 每个矩阵元素 $m_{ij} = \log([e_i(\mathbf{A} + \mathbf{A}^2 + \dots + \mathbf{A}^t)]_j / t)$, 其中 $\mathbf{A}$ 为一个转换矩阵, $[e_i(\mathbf{A} + \mathbf{A}^2 + \dots + \mathbf{A}^t)]_j$ 表示 j 出现在 i 的右侧邻居中的期望次数, 当 t 变大时, 计算准确的 $\mathbf{M}$ 的时间复杂度为 $\mathcal{O}(|V|^3)$. DeepWalk 使用基于随机游走的抽样方法以避免计算准确的矩阵 $\mathbf{M}$, 这样 DeepWalk 采样更多的步数时计算性能会更好, 但相反其效率会更低. TADW 则在速度和精度两者之间进行了折衷即设置因子分解矩阵 $\mathbf{M} = (\mathbf{A} + \mathbf{A}^2) / 2$, 这里用 $\mathbf{M}$ 替换 $\log^M$ 出于计算效率, 其原因是 $\log^M$ 具有比 $\mathbf{M}$ 更多的非零项, 并且具有平方损失的矩阵分解的复杂性与矩阵 $\mathbf{M}$ 中的非零元素的数量成比例^[106]. 由于大多数复杂信息网络是稀疏的, 即 $\mathcal{O}(\mathbf{E}) = \mathcal{O}(V)$, TADW 方法计算矩阵 $\mathbf{M}$ 的时间复杂度为 $\mathcal{O}(|V|^2)$, 如果网络是稠密的, 甚至可以直接因式分解矩阵 $\mathbf{A}$. 求解矩阵 $\mathbf{E} \in \mathbb{R}^{k \times |V|}$ 以及 $\mathbf{H} \in \mathbb{R}^{k \times f_t}$, 使其最小化公式如式(6)所示:

$$\min_{\mathbf{E}, \mathbf{H}} \|\mathbf{M} - \mathbf{E}^T \mathbf{H} \mathbf{T}\|_F^2 + \frac{\lambda}{2} (\|\mathbf{E}\|_F^2 + \|\mathbf{H}\|_F^2) \quad (6)$$

为了优化 $\mathbf{E}$ 和 $\mathbf{H}$, 通过迭代式地最小化 $\mathbf{E}$ 和 $\mathbf{H}$. 虽然 TADW 可以收敛到局部最小值而不是全局最小值, 但该方法在实验中效果较好. 不同于低秩矩阵分解, TADW 的目标是在网络结构表示学习中融入文本特征以获得更好的网络表示. 由于从 TADW 获得的 $\mathbf{E}$ 和 $\mathbf{H} \mathbf{T}$ 都可以被认为是顶点的低维表示, 通过将它们连接起来构建用于网络表示的统一的 $2k$ 维矩阵. TADW 模型能够有效地融合顶点的内容信息从而获得更好的网络表示学习.

除此之外, 针对 DeepWalk 模型的主要缺陷在于忽视现实世界网络中固有的链路稀疏问题^[107]以

及顶点包含的丰富内容, Ganesh 等人^[108]提出了一个新颖的 Author2Vec 模型用于学习低维的作者表示, 使得那些具有相似内容和共享类似网络结构的作者在向量空间中更相近. Author2Vec 模型通过协同方式融合文本信息与链接信息以用于创建作者表示. Author2Vec 模型由内容信息模型(Content-Info Model)和链接信息模型(Link-Info Model)两个子模型组成. 内容信息模型以论文的摘要作为文本内容来进行作者表示, 其目标函数定义为

$$L_C = \mathbb{P}[r_C(u, p) = l] = \text{softmax}(U_C \cdot \tanh(W_C^{(\times)} h_C^\times + W_C^{(+)} h_C^+ + b_C^{(h)}) + b_C^{(p)}) \quad (7)$$

式中 $r_C(u, p)$ 表示作者 u 和论文 p 之间的关系, $l \in [1, 2]$, $U_C \in \mathbb{R}^{2 \times n_h}$, $b_C^{(p)}$, $W_C^{(\times)} \in \mathbb{R}^{n_h \times d}$, $W_C^{(+)} \in \mathbb{R}^{n_h \times d}$, $b_C^{(h)}$ 为该模型的参数, 可通过最大化似然函数获得这些参数的最优值, $h_C^\times$ 和 h_C^+ 则采用 Zhu 等人^[109]提出的方法获取. 与内容信息模型建模过程类似, 链接信息模型以作者之间的链接来建模并进一步丰富作者表示, 其目标函数定义为

$$L_L = \mathbb{P}[r_L(u, v) = l] = \text{softmax}(U_L \cdot \tanh(W_L^{(\times)} h_L^\times + W_L^{(+)} h_L^+ + b_L^{(h)}) + b_L^{(p)}) \quad (8)$$

式中 $r_L(u, v)$ 表示作者 u 和作者 v 之间的关系, 其它参数的求解以及优化与内容信息模型方法一样. 通过共享作者嵌入权重来连接这两个子模型, 最终 Author2Vec 模型的总体目标函数, 可以写成 $L = L_C + L_L$. Author2Vec 模型学习到的作者嵌入在链接预测和聚类效果上优于 DeepWalk 模型.

3.4.2 基于概率图模型的方法

基于概率图模型方法是采用贝叶斯图模型方法对网络顶点之间的联合概率进行建模, 学习到每个顶点的低维向量表示. 例如, Cohn 等人^[110]提出了将 PLSA 和 PHITS (Probabilistic Hypertext-Induced Topic Search) 结合起来的联合概率模型, 用于建模文档集合的内容和互相链接. 该模型基于概率因子分解, 可以识别出文档集合的重要主题以及主题内的权威文档. 此外, 映射主题之间的关系以便学习构建链接内容的预测模型.

通过考虑复杂信息网络结构以及文本主题的结合方式, Mei 等人^[111]提出了一个基于网络结构嵌入的主题: TMN (Topic Modeling with Network Structure) 的问题, 并提出解决这种问题的 NetPLSA 模型, 即使用基于网络结构的调和正则化对统计主题模型进行正则化. 所提出的模型结合主题建模和社交网络结构分析, 并利用统计主题模型和离散正

则化的优势. NetPLSA 综合考虑了网络结构和文本内容来进行网络表示学习, 在学习过程中对于网络中的每个顶点的文本主题提取采用 PLSA 模型, 学者们给出其对数似然函数如下:

$$L(C) = \sum_d \sum_w c(w, d) \log \sum_{j=1}^k p(\theta_j | d) p(w | \theta_j) \quad (9)$$

其中 C 为一个聚类, $\{\{p(\theta_j | d), p(w | \theta_j)\}_{d, w, j}\}$ 为模型参数, d, w 分别代表一个文档和一个词汇集合. 自然的可使用 $\{p(\theta_j | v)\}_j$ 作为顶点 v 的权重, 由式(10)计算 $p(\theta_j | v)$ 的值:

$$p(\theta_j | v) = \sum_{d \in D_v} p(\theta_j | d) p(d | v) \quad (10)$$

为了结合网络结构进行表示学习, 文中定义了正则化似然函数 $S(C, G)$ 为式(11)所示:

$$S(C, G) = -(1 - \lambda)L(C) + \lambda R(C, G) \quad (11)$$

其中 $L(C)$ 由式(6)得到, $R(C, G)$ 是定义在网络结构 G 上的一个类似于图的调和函数的正则化^[112], 用式(12)所示:

$$R(C, G) = \frac{1}{2} \sum_{j=1}^k \mathbf{f}_j^T \mathbf{\Delta} \mathbf{f}_j \quad (12)$$

这里的 $\mathbf{f}_j$ 代表每个顶点的第 j 个主题权重的 $|V|$ 维向量, $\mathbf{\Delta}$ 是图拉普拉斯矩阵, 有 $\mathbf{\Delta} = \mathbf{D} - \mathbf{W}$, 其中 $\mathbf{D}$ 为对角矩阵, 其矩阵元素表示为 $d(u, u) = \sum_v w_{uv}$. 该模型的输出可以概括文本中的主题, 将主题映射到网络上以及发现主题社区. 通过适合的实例化主题模型和基于网络的正则化, 此模型可以广泛地应用于文本挖掘问题, 如主题分析、社区发现和空间文本挖掘等.

然而, 网络中顶点除了结构信息之外, 还包含真实存在的诸多信息, 如用户产生的文本信息以及用户本身的属性信息(性别、学校、地址、职业). Li 等人^[113]充分考虑了顶点以上多方面信息, 提出了 MFR (Multi-Faceted Representations) 学习模型. 该模型的目标在于学习得到每个用户、每个属性实体以及每个用户属性关系的低维向量表示. 每个用户 u_i 用一个列表 $\mathbf{I}_i = \{\text{text}_i, \mathbf{G}_i, \mathbf{R}_i, \mathbf{M}_i\}$ 与之相关联, 整个模型最大化 $\mathbf{I}_i$ 的出现概率表示为式(13)所示:

$$p(\mathbf{I}_i | u_i; \boldsymbol{\Theta}) = p(\text{text}_i | u_i; \boldsymbol{\Theta}_u) \times p(\mathbf{G}_i | u_i; \boldsymbol{\Theta}_u) \times p(\mathbf{R}_i, \mathbf{M}_i | u_i; \boldsymbol{\Theta}_u, \boldsymbol{\Theta}_r, \boldsymbol{\Theta}_m) \quad (13)$$

式中 $\boldsymbol{\Theta}_u, \boldsymbol{\Theta}_r, \boldsymbol{\Theta}_m$ 分别代表用户、用户属性实体关系和实体的潜在表示矩阵, $\boldsymbol{\Theta}$ 表示整个模型的参数空间即 $\{\boldsymbol{\Theta}_u, \boldsymbol{\Theta}_r, \boldsymbol{\Theta}_m\}$. $p(\text{text}_i | v_i; \boldsymbol{\Theta}_u)$ 借助于段落向量 (Paragraph Vector)^[16] 的思想使用用户层级的嵌入和邻接词来预测文本中的目标词. 计算公式如式(14)所示:

$$p(\text{text}_i | v_i; \Theta_u) = \sum_{w \in \text{text}_i} \log p(w | u_i) + \sum_{w \in \text{text}_i} \sum_{w' \in c(w)} \log p(w | w') \quad (14)$$

由于式(14)中后一项词级别的表示可以预先学习好,因此只需要使用 AUC 损失函数的方法对前一项进行优化. $p(G_i | v_i; \Theta_u)$ 定义为用户 v_i 的链接顶点的似然,类似于 DeepWalk 和 LINE 模型所使用的 SkipGram 方法通过用户顶点来预测其链接的顶点,然后采用 AUC 损失函数的方法进行优化. $p(R_i, M_i | v_i; \Theta_u, \Theta_r, \Theta_m)$ 则用于预测用户 u_i 与属性实体 m 的关系. MFR 模型使用随机梯度下降法对整个模型进行优化处理.

针对 TADW 模型通过矩阵分解将文本特征融合到网络嵌入中,这通常导致计算成本较高,并且 TADW 中的文本内容被简单地融合作为无序文本特征而不是被明确建模,因此也不能很好地捕获更深的语义等问题,Sun 等人^[114]提出了结合网络结构和文本信息的内容增强网络嵌入方法 CENE(Content-Enhanced Network Embedding). 该方法通过将文本内容作为特殊类型的顶点来处理,用一个普适性框架来融合文本建模和结构建模. 本文作者描述了一个最小化目标函数的普适性框架:

$$L_g = \sum_{(u,v) \in SP} \log p(u,v;\theta) + \sum_{(u,v) \in SN} \log(1 - p(u,v;\theta)) \quad (15)$$

式中 SP 表示积极顶点对的集合, SN 表示消极顶点对的集合,顶点 u 和 v 之间的联合概率 $p(u,v;\theta)$ 表示为顶点对 (u,v) 在 SP 出现的概率, $1 - p(u,v;\theta)$ 则与之相反. 用该框架分别对 node-node link 和 node-content link 建立损失函数 L_{nn} 和 L_{nc} . 联合目标函数定义为 $L = \alpha L_{nn} + (1 - \alpha) L_{nc}$, 该函数是上述两个损失函数的加权组合,这里的 $\alpha \in [0, 1]$ 是用于平衡 node-node link 和 node-content link 两者的重

要性的参数. 随着 α 值的增加,将更多地考虑网络中的结构信息即 node-node link. 使用随机梯度下降与学习速率衰减对联合目标函数 L 进行优化,梯度则使用反向传播计算.

现有学习方法仅仅关注顶点信息的某一个方面,并且没有利用顶点的标签信息,于是 Pan 等人^[115]提出了三方深度网络表示模型 TriDNR(Tri-party Deep Network Representation Model),通过使用来自于网络的顶点结构、顶点内容和顶点标签联合学习最佳的顶点表示. TriDNR 是建立在一种新的耦合深度自然语言模块的基础上,从三个层面进行顶点表示学习:在网络结构层面,通过最大化在随机游走中给定某一顶点的情况下,观察其周围顶点的概率, TriDNR 获取了顶点间的关系. 在顶点内容层面,通过最大化给定顶点的词序列的共现关系, TriDNR 捕获了顶点-词的相关性. 在顶点标签层面,通过最大化给定一个类标签的词序列的概率, TriDNR 建模了标签-词的相关性. 以上三方信息被一起输入到神经网络模型中以彼此相互增强方式学习网络最佳表示.

最新的基于网络结构-内容融合学习的表示模型总结如表 2 所示. 这些模型可以看作是基于网络结构表示学习的拓展,通过在网络结构建模的基础上有效地融合顶点内容信息获得更好的顶点分布式表示. 从顶点内部信息的种类来说, TADW、CENE、Author2Vec 以及 TriDNR 等模型考虑了顶点自身产生的文本信息. TADW 能够有效地融合文本信息,获得比不用文本特征的基于结构表示学习模型、只用文本特征的表示学习模型以及简单将两模型所得表示串接起来的方法更好的实验效果,但降维时采用矩阵因子分解,导致模型复杂度较高. CENE 模型把文本当成一类特殊的顶点进行处理,从而学习到更深的语义信息. Author2Vec 模型通过融合两类

表 2 结构-内容融合的表示学习模型

模型	核心算法与观点	对象	类型	实验规模	评测数据集	评测应用	备注
TADW ^[104] (2015)	※随机游走产生输入序列 ※基于 Skip-gram 模型	node	D/UD	3312	文本网络、引用网络	多标签顶点分类	在 DeepWalk 模型中融入文本特征
Author2Vec ^[108] (2016)	※内容信息模型、链接信息模型 ※优化使用随机梯度下降	node	D	1 212 171	引用网络	链接预测、聚类、网络可视化	无监督学习/适合大规模网络
MFR ^[113] (2015)	※学习网络的全局特征 ※优化部分使用并行随机梯度下降	node	UD	75 000 000	社交网络	链接预测、社区发现、用户属性预测	适合大规模网络/类似于 Skip-gram 模型
CENE ^[114] (2016)	※随机梯度下降进行参数寻优 ※通用框架融合文本、结构建模	node	D/UD	629 814	社交网络、引用网络	顶点分类	适合大规模网络
TriDNR ^[115] (2016)	※耦合神经网络模型学习 ※随机游走产生输入序列	node	D/UD	60 744	引用网络、合作网络	顶点分类 网络可视化	融合结构、内容和标签联合学习/适合大规模网络

子模型,以无监督的方式学习作者嵌入,取得了优于 DeepWalk 模型的学习效果. TriDNR 模型还考虑了顶点标签信息,从不同网络层级进行学习以获得最佳的表示效果. 模型 MFR 考虑了文本信息和顶点自身的属性信息,获得多种关系的低维向量表示. 通过对网络结构、内容以及相关属性信息建模到统一空间, MFR 模型学习得到的网络表示不仅可以预测用户之间的链接关系,同时也可以预测用户本身具有的属性信息,如性别、地址、职业等.

3.5 异构网络的表示学习

异构网络是一种由不同类型的顶点(如用户、内容和群组等)和关系(如社交或相似关系)所组成的网络. 例如, Flickr 网络顶点类型同时包含用户、视频和标签等,顶点间又通过相同或不同类型的联系而关联. 尽管目前在同构网络上的网络表示学习的研究已经取得了较大的进展,但在异构网络上,由于其复杂性,使得表示学习仍是一个非常具有挑战性的研究问题.

当前,有关异构网络表示学习的文献相对较少. Tang 等人^[116]研究了社交网络中用户交互信息的分类任务. 针对社交网络由多种关系组成的异构网络,而大多数现有方法把这种异构关系当成同构进行处理,由此无法获得令人满意的分类性能. 基于这一问题,他们提出了社会维度的概念来表示用户的潜在隶属关系,并构建了一个分类框架 SocioDim(Social Dimensions). 该框架先提取网络结构的社会维度,以准确地捕获用户之间的交互模式,然后学习一个判别分类器来选择相关的社会维度. 通过区分不同类型的网络关系, SocioDim 获得了较好的分类性能.

与 SocioDim 模型出发点不同, Jacob 等人^[117]提出了隐含空间异构模型 LSHM(Latent Space Heterogeneous Model)用于处理异构网络的表示学习. 该模型的思想是无论网络中的顶点属于哪种类型,都可对顶点在同一向量空间学习其低维表示. LSHM 模型针对顶点属性采用一种扩展方式即把属性也作为一种没有标签的顶点,同时学习各种类型的顶点向量表达和标签的线性分类函数,它的优化目标函数涉及两方面:一方面考虑了网络上的平滑性,即相邻顶点的标签尽可能相似;另一方面考虑了分类函数对已知标签的预测能力. 如果只考虑网络的平滑性,则 LSHM 模型就可以用相似度矩阵作为输入. 该模型可以看成是一种半监督的网络表示学习模型.

为了结合多种类型的关系来度量异构网络中维基百科实体之间的语义相关性, Zhao 等人^[118]提出了调和矩阵因式分解 CMF(Coordinate Matrix Factorization)模型,该模型构建了同一语义空间中的实体、类别和词的低维连续表示方法. 在 CMF 模型中,学习实体表示矩阵 $E_H \in \mathbb{R}^{|E| \times K}$ 的过程被设想为对实体-实体系数矩阵 $X \in \mathbb{R}^{|E| \times |E|}$ 的估计. 为了结合实体和词之间的关系以及实体和类别之间的关系,又各自构建了两个系数矩阵: $Y \in \mathbb{R}^{|C| \times |E|}$ 、 $Z \in \mathbb{R}^{|W| \times |E|}$, 其中 Y 由实体表示矩阵 E_H 和类别表示矩阵 E_C 的内积来估计, Z 由实体表示矩阵 E_H 和词表示矩阵 E_W 的内积来估计. 求解 CMF 模型的目标函数为

$$\begin{aligned} J = & \sum_{(i,j) \in N_X} (X_{ij} - E_{H_i} E_{L_j}^T)^2 + \\ & \sum_{(i,j) \in N_Y} (Y_{ij} - C_{H_i} E_{H_j}^T)^2 + \\ & \sum_{(i,j) \in N_Z} (Z_{ij} - W_{H_i} E_{H_j}^T)^2 + \\ & \lambda (\|E_H\|_F^2 + \|E_L\|_F^2) + \\ & \gamma (\|C_H\|_F^2 + \delta \|W_H\|_F^2) \end{aligned} \quad (16)$$

式中 λ, γ, δ 为正则化因子, N_X, N_Y, N_Z 分别代表 X, Y, Z 的非零实体. CMF 模型使用随机梯度下降法对整个模型进行优化. 通过将网络中不同的关系建模到统一空间,模型学习到的网络表示,不仅可以克服网络稀疏性问题,而且通过引入更多的关联矩阵可容易地进行扩展.

为了引入标签信息来指导文本表示的学习, Tang 等人^[119]在 LINE 模型的基础上针对文本标签预测任务提出了半监督表示学习模型 PTE(Predictive Text Embedding). PTE 通过将部分标签已知的文档集合数据转换为一个包含文档、词语和标签三类顶点的异构网络,然后学习不同类型顶点的向量表示. 相比无监督表示模型 LINE 和段落向量以及有监督的神经网络模型 CNN(Convolutional Neural Network), PTE 模型在长文本语料和短文本语料都取得了较好的效果;同时 PTE 模型的参数在不同数据集上表现也更加稳定.

此外,为了分析具有各种类型的顶点和内容的异构网络特征嵌入的情况, Chang 等人^[120]提出了一种异构网络嵌入 HNE(Heterogeneous Network Embedding)表示学习算法,将不同的异构对象映射到统一的隐含空间,以便可以直接比较来自不同空间的对象. 图 9 表示为 HNE 模型的架构图, HNE

将特征学习过程分解为深层结构的多个非线性层，高度非线性的多层嵌入函数用于捕获网络中异构数据之间的复杂交互。他们的目标是创建一个多分辨率深度嵌入函数来反映局部和全局网络结构，并使

得最终的嵌入可用于各种数据挖掘任务。尤其是他们实证了异构网络中丰富的内容和链接信息可以通过这种方法捕获，使得跨模态数据之间的相似性可以度量在同一向量空间中。

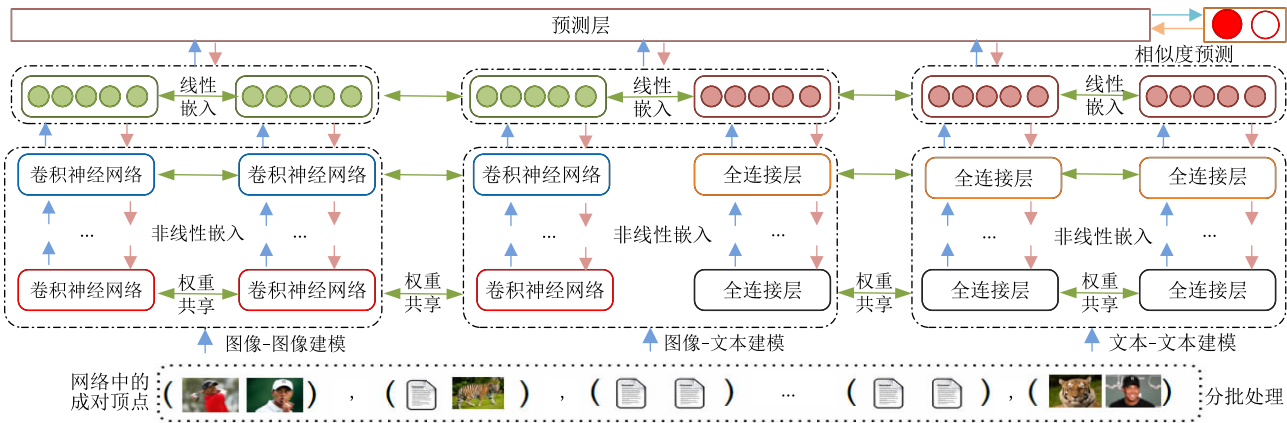


图 9 HNE 模型架构

4 大规模复杂信息网络表示学习的评测

本节将从大规模复杂信息网络学习所涉及的评测数据集、应用场景与评测指标三个方面展开讨论。

4.1 评测数据集

网络表示学习方法几乎都需要在真实的数据集上验证其方法的有效性，因此数据集的构建与选择对于完全地复现和比较不同的网络表示学习方法很

重要。构建和收集数据集是一项耗时且繁琐的工作，而且并非所有数据集都是公开和容易获取的。本文总结了目前网络表示学习模型所使用的一些大规模数据集，以及包括这些数据集的参考文献，具体信息见表 3 所示。

从表 3 可知，在线社会网络(如 YouTube、Flickr 以及 Blogcatalog)和合作网络(如 DBLP)是最受欢迎的数据集。由于这些数据集获得斯坦福大学的精心维护，所以许多研究人员更喜欢使用这些数据集。然而，同样需要看到当前数据集用在网络表示学习上验证与应用所存在的一些不足：首先，部分

表 3 大规模复杂信息网络表示学习的数据集

类型	数据名称	方向 [#]	大小		网络密度	平均度	平均集聚度	参考文献
			顶点数	边数				
社会网络	YouTube	UD	1.13E+06	2.99E+06	4.64E-06	5.265	0.081	[21,31,40,68,72,76]
	Facebook	UD	4.04E+03	8.82E+04	4.02E-04	25.640	0.606	[70]
	Digg	D/UW	2.80E+05	1.73E+06	2.21E-05	12.385	0.061	[40]
	Flickr	D	2.30E+06	3.31E+07	6.25E-06	28.781	0.089	[21,31,40,71-72,116-117,120]
	Blogcatalog	UW	1.03E+04	3.34E+05	6.3E-03	64.780	0.031	[31-32,70-71,116,120]
引文网络	Twitter	D/UW	5.26E+07	1.96E+09	7.10E-07	74.678	0.565	[105,120]
	DBLP	UD	3.17E+05	1.05E+06	2.08E-05	6.622	0.632	[21,32,69,76,114-115,117,119]
	AstroPh	UD	1.87E+04	1.98E+05	5.71E-03	17.323	0.631	[70]
通讯网络	CiteSeer	D/UW	3.84E+05	1.75E+06	1.19E-05	9.112	0.385	[68,108,115]
	Enron	UD	3.67E+04	1.84E+05	2.73E-04	10.020	0.497	[76]
	EuALL	D	2.65E+05	4.20E+05	1.04E-05	2.757	0.067	[76]
公路网络	pediaEnglish	D/UW	2.39E+06	5.02E+06	8.75E-07	4.194	0.021	[118]
	California	UD/UW	1.97E+06	2.77E+06	1.43E-06	2.815	0.046	[76]
	Pennsylvania	UD	1.09E+06	1.54E+06	2.60E-06	2.834	0.047	[76]
维基百科	Texas	UD	1.38E+06	1.92E+06	2.02E-06	2.785	0.047	[76]
	Talk network	D	2.39E+06	5.02E+06	1.63E-06	3.892	0.053	[76]
	Vote network	D	7.12E+03	1.04E+05	3.98E-03	28.323	0.141	[76]
	Wikipedia	D/W	1.99E+06	1.00E+09	1.49E-06	504.220	0.043	[21,32,70,119]

[#]: D/UD/W/UW: 有向/无向图/加权图/非加权图。

数据集包含噪音数据,必须在使用之前进行清洗.例如合作网络面临着作者名的消歧问题,这将导致大量的噪音,使得网络与真实世界网络不一致;其次,在不同数据集上比较相同的度量指标时,它们的性能等级通常不一致或者甚至相差很大,因此有必要构建和维护用于网络表示学习的基准数据集.除了基准的评测数据集之外,Niu 等人^[121]还设计了一种并行化的图嵌入(Graph Embedding)代码库,该代码库为当前图嵌入实现效率的对比提供了一个较强的基准方案.

4.2 大规模复杂信息网络表示学习的应用场景和评测指标

网络表示学习是近两年来在复杂信息网络领域中形成的新兴研究方向之一.通过学习得到良好的网络特征表示,不仅仅能够解决目前大规模复杂信息网络所存在的数据稀疏性以及动态性等问题,更能有效地应用到链接预测、顶点分类、可视化等衍生应用场景中.

以链路预测研究为例,其不仅具有广泛的实际应用价值,也具有重要的理论研究意义,特别是对相关领域方面的推动和贡献.随着网络科学的快速发展,使得链路预测的研究与网络的结构以及演化紧密联系起来,也从理论上帮助人们认识复杂网络演化的机制.虽然,针对同一类网络,有很多模型都提供了可能的演化机制^[122-123].然而,很多链接预测研究主要针对特定的网络结构进行详细的分析和设计,导致预测方法变得较为复杂,也越来越难以适应大规模的网络结构.

现在用于链接预测的网络表示学习方法中,对于网络中顶点的特征学习不再是进行人工的构建,而是利用顶点在网络中随机游走的方法,把传统网络中顶点的结构特征映射到连续的、维度固定的向量空间,从而得到大规模网络中顶点结构特征的分布式表达.然后基于顶点间相似性预测这一前提假设,计算不同顶点结构特征向量的相似性估计顶点之间存在链接的可能性.常用的链接预测精确度的评测指标主要有 AUC(Area Under the Receiver Operating Characteristic Curve)和精确度(precision)两类,其中 AUC 是最常用的一种评测指标,主要侧重于从整体上衡量算法的准确度,而精确度主要考虑在只有评价部分链接(排序前 N 位)预测的准确率.通过网络表示学习在链接预测应用场景下得到的评测指标值都有了很大的性能提升,同时具备在大规模网络上的可计算性^[40,70,71,76,113].

大规模信息网络积累了大量的多标签数据,对

这些多标签数据如何进行有效分类成为一个有待解决的课题,同时也吸引了很多研究者的关注.根据样本数据拥有标签的多少,可以将分类问题分为单标签分类和多标签分类,然而,这两种分类都面临着“维数灾难”的问题,尤其在是在大规模网络数据的背景环境下,这种问题更加明显,因此特征的低维度表示是有效的解决方案之一.

在网络标签分类应用场景中,网络部分顶点的标签是已知的,分类的目的是预测出剩余顶点的标签.该问题解决的通常流程是:首先,学习表示出每个顶点的特征向量;然后,利用已知标签的顶点训练分类模型;最后,以未知标签顶点的向量表示作为该分类模型的输入,推断出它们的标签类别.

在网络表示学习的过程中,如果利用标签信息,往往可以提高标签分类的性能^[117,119].同时,网络表示学习的半监督分类可能会优于网络表示学习的无监督分类,这是因为前者学习到的网络向量表示是和分类预测场景相关的.标签分类的评测指标通常采用汉明损失(Hamming loss)、覆盖度(coverage)、1 错误率(one-error)、排序损失(ranking loss)、分类准确率(classification accuracy)以及平均精度(classification accuracy)等作为评测指标.其中,尤其普遍使用分类准确率用于标签分类的性能评测.

网络表示学习的另一应用场景是对学习到的网络进行可视化展示^[21,32,69,71],具体的方法是提取顶点的特征向量表示进行降维处理后,将网络中的顶点都映射到低维空间上.在可视化过程中需要顶点标签信息,标签相同的顶点之间距离应该尽可能小.常用的高维向量降维算法有 t -SNE^[119]、PE^[124]等.例如,Tang 等人^[21]给出了利用 t -SNE 对 DBLP 数据集中论文合作网络进行可视化的示例,在展示的结果中每个顶点代表一个用户,每一种颜色代表一个研究领域.可视化应用的评测指标则是相同颜色的顶点在可视化空间中越集中,则表明网络表示学习算法的性能越好.信息网络表示学习的不同表现形式,不仅可以应用到以上的应用场景中,而且还可以拓展到诸如推荐系统、隐私保护系统等其它应用中.

5 研究难点和发展趋势

大规模复杂信息网络存在广泛.对这类复杂网络特征的表示学习与抽取是当前的研究热点之一.大规模复杂信息网络表示学习的研究对将其应用到网络顶点分类、推荐系统、链接预测、文本建模和可

视化处理等方面具有重要的意义,大规模复杂信息网络具有的丰富性、动态变化性、隐私保护与信息残缺性、稀疏性等特点使得可见信息的隐含网络特征学习成为可行但又具有很大的挑战的研究内容。虽然已有不少的网络表示学习模型的工作被报道,但是仍然存在许多难以解决的问题以及可进一步研究的方向,概括起来,未来的研究难点和发展趋势主要包括网络特征提取模型的可容错性研究、动态适应性研究、可融合性研究、独立架构研究、跨网络研究以及特殊子结构研究。

(1) 网络特征提取的可容错

由于现有大规模复杂信息网络来源多样化,同时其中所包含的噪音甚至错误数据也越来越多。例如很多用户会对自身的关注或者被关注数据进行隐藏甚至错误标注;不同网站之间的链接存在恶意引用或者盗用等情况。因此,针对这种大规模复杂网络中所存在的噪音数据应该如何有效地处理、过滤,甚至是有效地识别,是当前研究需要考虑的问题之一。

(2) 网络动态变化的适应性

传统的大规模网络的顶点结构特征学习主要侧重在对静态网络或增长型网络进行特征学习。例如对新增链接的信息或新增顶点的特征进行学习。然而,网络在新增信息的同时,也会伴随着信息的消失。相对于信息的新增,网络顶点或链接的消失机制更加有待探索。作为网络动态结构变化的一部分,如何学习在高速动态变化环境下,由顶点或边消失而引起的特征关系的改变,这种改变与原有的新增顶点或边之间存在何种关系,值得进一步的讨论。

(3) 网络特征的可融合性

大规模复杂信息网络所包含的特征信息不仅仅限于现有的结构特征,网络数据还包含多样化和异质化的信息。由于网络上包含不同类型的实体以及网络的链接关系多样化,针对这一问题网络表示学习的相关研究工作往往进行了简化处理,仅仅考虑顶点的异构性,很少考虑链接的异构性。最后,由于复杂信息网络数据并不孤立存在,而是与各种数据类型组合存在,这些交互可以通过它们之间的联系明确地或隐含地形成。如对在同一网页内共同出现的图像和文本提供它们之间的明确链接,而文本到文本的链接由不同网络文档之间的超链接形成。另一方面,用户的交互活动可以被视为隐式反馈,其链接不同的网络中的组件。如果用户描述具有类似标签的多个图像,则可以合理地假定在这些图像之间存在语义关系。复杂信息网络数据所呈现的以上各种特征问题导致了网络结构十分复杂,同时也对于

计算性能问题提出了更高的要求,如何根据特定的应用,充分利用丰富信息,通过有效的特征融合后,发现隐含的有用网络特征表示,对于网络分布式特征的研究具有重要意义。

(4) 网络特征学习模型架构的独立性

复杂信息网络上人类行为的复杂性以及人们在网络交流过程中的各种反应和互动的复杂性,使得网络的特征提取工作变得非常困难。复杂信息网络存在着这种千差万别,使得研究者很难找到一种适合各种类型网络的通用网络表示学习框架。不同表示学习在各种网络中的准确性以及适用范围的广泛性上也存在着一定的局限性。由于学习过程中只能依据一种或几种特征判断来进行网络表示,无法也不可能将所有的网络特征情况都考虑进去,因此如何对复杂信息网络更好地分析进而设计一个有效的网络表示学习模型是一个很重要的问题,且也是相当困难的。另外,为了支持分析网络的应用,需要通过网络嵌入来保留网络结构,然而网络的真实结构是未知的。顶点的相似性依赖于局部和全局网络结构,因此模型学习如何同时保留局部和全局结构是一个棘手的问题。许多现实世界的网络往往是非常稀疏,因此仅仅利用非常有限的、观察到的链路来建立学习模型是不足够达到令人满意的性能。

(5) 跨网络特征分布式表达模型

通常而言,针对单一网络的特征学习模型很难有效地直接迁移到多个网络特征的并行或跨领域的学习过程中来。目前,针对单一网络特征的学习还处于解决网络的动态性、大规模等基础问题上。网络分布式特征表达的跨领域迁移学习尚处于起步阶段。针对多个网络之间的关联学习(又称为跨网络匹配问题)缺少现有的探讨。针对这一问题,我们认为应该先从小领域内的跨网络学习研究入手,找出不同的领域网络之间的区别与联系,通过保存共同特征的提取模型,逐步的对现有方法进行改进和重构。

(6) 特殊子结构的网络特征学习

现有的众多网络表示学习的方法主要集中在对顶点的表示学习,即通过表示学习得到顶点的低维度向量化表示形式,但对于网络的其它组件,包括边、多边形(如三角形、四边形)、网络路径等这些特殊子结构的表示学习缺乏探讨。针对网络其它组件分布式特征表示学习的研究,可以进一步提升我们针对网络本身演化本身的认识,从而得出更多可能的应用方向。例如,可以得到包括复杂信息网络的点击流(互联网用户行为方面)的分布式特征表达等等。

6 结束语

网络表示学习从表现形式上说,是将网络中任意顶点的结构特征映射到一个低维度的、连续的实值向量.从本质上讲,又常常使用矩阵分解方法来完成,但是在实际的优化过程中,又通常伴随着顶点内容或属性等性质的融合.

一般地,由于当前的大规模复杂信息网络数据具有的各种特性,使得传统的网络分析方法显得捉襟见肘,因此针对大规模复杂信息网络表示学习研究因此应运而生.大规模学习网络的特征表示具有重要的实际意义和应用价值.

首先,学习网络中每个顶点的特征向量表达可以有效缓解网络数据稀疏性问题.其次,把网络中不同类型的异质信息融合为整体,可以更好地解决特定问题.第三,网络的分布式向量表示能够高效地实现语义相关性操作,从而显著提升在大规模,特别是超大规模的网络中进行相似性顶点匹配的计算效率.最后,学习网络的分布式表示使得在网络顶点分类、推荐系统等方面应用前景得到了更大的拓展.

本文概括了近年来关于大规模复杂信息网络表示学习的主要研究模型及其特点.当前研究中,大多数模型是根据复杂信息网络的结构以及根据内容或两者融合来进行顶点的特征表示学习.同时,我们可以看到,融合网络结构特征和内容特征的表示学习能够更好地反映出一个网络特征的真实情况,使得其学习得到的网络特征表示更具有意义与价值.除此之外,本文还对大规模复杂信息网络表示学习目前所面临的研究问题、已有的解决方案以及未来的可行研究方向进行了总结和论述.希望本文对于大规模复杂信息网络表示学习在国内的研究发展提供一些帮助.

参 考 文 献

- [1] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(8): 1798-1828
- [2] Turian J, Ratnoff L, Bengio Y. Word representations: A simple and general method for semi-supervised learning// *Proceedings of the Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, 2010: 384-394
- [3] Bordes A, Usunier N, Garcia-Duran A. Translating embeddings for modeling multi-relational data// *Proceedings of the Advances in Neural Information Processing Systems*. Cambridge, England, 2013: 2787-2795
- [4] Wang Z, Zhang J, Feng J. Knowledge graph embedding by translating on hyperplanes// *Proceedings of the 28th AAAI Conference on Artificial Intelligence*. California, USA, 2014: 1112-1119
- [5] Lin Y, Liu Z, Sun M. Learning entity and relation embeddings for knowledge graph completion// *Proceedings of the 29th AAAI Conference on Artificial Intelligence*. Austin Texas, USA, 2015: 2181-2187
- [6] Zhao J, Xu L, Liu K. Knowledge graph embedding via dynamic mapping matrix// *Proceedings of the 53rd of the Association for Computational Linguistics*. Beijing, China, 2015: 687-696
- [7] Golub G H, Van Loan C F. *Matrix Computations*. Baltimore, Maryland, USA: JHU Press, 2012
- [8] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000, 290(5500): 2323-2326
- [9] Tenenbaum J B, Silva V D, Langford J C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000, 290(5500): 2319-2323
- [10] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems*, 2002, 14(6): 585-591
- [11] Hoff P D, Handcock M S. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 2002, 97(12): 1090-1098
- [12] Li Y, Tarlow D, Brockschmidt M. Gated graph sequence neural networks// *Proceedings of the 4th International Conference on Learning Representations*. San Juan, Puerto Rico, 2015: 1-20
- [13] Ahmed A, Shervashidze N, Narayanamurthy S. Distributed large-scale natural graph factorization// *Proceedings of the 22nd International Conference on World Wide Web*. Rio de Janeiro, Brazil, 2013: 37-48
- [14] Chen X, Liu Z, Sun M. A Unified model for word sense representation and disambiguation// *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, 2014: 1025-1035
- [15] Mikolov T, Sutskever I, Chen K. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 2013, 26: 3111-3119
- [16] Le Q V, Mikolov T. Distributed representations of sentences and documents. *Computer Science*, 2014, 4: 1188-1196
- [17] Fortunato S. Community detection in graphs. *Physics Reports*, 2010, 486: 75-174.
- [18] Rapoport A. Contribution to the theory of random and biased nets. *Bulletin of Mathematical Biology*, 1957, 19(4): 257-277
- [19] Travers J, Milgram S. An experimental study of the small world problem. *Sociometry*, 1969, 32(4): 425-443
- [20] Newman M E J. The structure and function of complex networks. *Siam Review*, 2003, 45(2): 167-256

- [21] Tang J, Qu M, Wang M. Line: Large-scale information network embedding//Proceedings of the 24th International Conference on World Wide Web. New York, USA, 2015: 1067-1077
- [22] Collobert R, Weston J, Bottou L. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 2011, 12: 2493-2537
- [23] Neelakantan A, Shankar J, Passos A. Efficient non-parametric estimation of multiple embeddings per word in vector Space. *Computer Science*, arXiv.org, 2015, cs.SI: 1-12
- [24] Zhuang F, Cheng X, Luo P. Supervised representation learning: Transfer learning with deep autoencoders//Proceedings of the 24th International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 4119-4125
- [25] Mairal J, Ponce J, Sapiro G. Supervised dictionary learning//Proceedings of the 21st Advances in Neural Information Processing Systems. Vancouver, Canada, 2008: 1-8
- [26] Jiang Q, Shao F, Jiang G. Supervised dictionary learning for blind image quality assessment using quality-constraint sparse coding. *Journal of Visual Communication and Image Representation*, 2015, 33: 123-133
- [27] O'Shea T J, Corgan J, Clancy T C. Unsupervised representation learning of structured radio communication signals//Proceedings of the 1st International Workshop on Sensing, Processing and Learning for Intelligent Machines. Aalborg, Denmark, 2016: 1-5
- [28] Schölkopf B, Platt J, Hofmann T. Efficient sparse coding algorithms. *Advances in Neural Information Processing Systems*, 2006, 19: 801-808
- [29] Hyvärinen A, Oja E. Independent component analysis: Algorithms and applications. *Neural Networks*, 2000, 13(4-5): 411-430
- [30] Ma R, Wang J, Song Y. Multi-manifold learning using locally linear embedding (LLE) nonlinear dimensionality reduction. *Journal of Tsinghua University*, 2008, 48(4): 582-585
- [31] Perozzi B, Alrfou 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
- [32] Cao S, Lu W, Xu Q. Grarep: Learning graph representations with global structural information//Proceedings of the 24th ACM International Conference on Information and Knowledge Management. Melbourne, Australia, 2015: 891-900
- [33] Angelova R, Kasneci G, Weikum G. Graffiti: Graph-based classification in heterogeneous networks. *World Wide Web*, 2012, 15(2): 139-170
- [34] Ji M, Han J, Danilevsky M. Ranking-based classification of heterogeneous information networks//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. California, USA, 2011: 1298-1306
- [35] Kong X, Yu P, Ding Y. Meta path-based collective classification in heterogeneous information networks//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York, USA, 2012: 1567-1571
- [36] Sen P, Namata G, Bilgic M. Collective classification in Network data articles. *Ai Magazine*, 2008, 29(3): 93-106
- [37] Tu C, Liu Z, Sun M. Inferring correspondences from multiple sources for microblog user tags. *Communications in Computer and Information Science*, 2014, 489: 1-12
- [38] Bhuyan M H, Bhattacharyya D K, Kalita J K. Network anomaly detection: methods, systems and tools. *IEEE Communications Surveys and Tutorials*, 2014, 16(1): 303-336
- [39] Lü L, Zhou T. Link prediction in complex networks: A survey. *Physica A Statistical Mechanics and Its Applications*, 2011, 390(6): 1150-1170
- [40] Li Zhi-Yu, Liang Xun, Xu Zhi-Ming. DNPS: Learning distributed representation of dynamic social networks. *Chinese Journal of Computers*, 2016, 39(42): 1-19(in Chinese)
(李志宇, 梁循, 徐志明. DNPS: 基于阻尼采样的大规模动态社会网络结构特征表示学习. *计算机学报*, 2016, 39(42): 1-19)
- [41] Chen M, Yang Q, Tang X. Directed graph embedding//Proceedings of the International Joint Conference on Artificial Intelligence. Hyderabad, India, 2007: 2707-2712
- [42] Brand M, Huang K. A unifying theorem for spectral embedding and clustering//Proceedings of the 9th International Conference on Workshop on Artificial Intelligence and Statistics. Florida, USA, 2003: 1-8
- [43] Jolliffe I. Principal Component Analysis. Hoboken, New Jersey, USA: John Wiley and Sons, Ltd, 2002
- [44] Ivanov O U, Bartunov S O. Learning representations in directed networks//Proceedings of the 4th International Conference on Analysis of Images, Social Networks and Texts. Yekaterinburg, Russia, 2015: 196-207
- [45] Teh Y W, Roweis S T. Automatic alignment of hidden representations//Proceedings in Advances in Neural Information Processing System. Cambridge, USA, 2002, 15: 841-848
- [46] Shaw B, Jebara T. Structure preserving embedding//Proceedings of the 26th International Conference on Machine Learning. Quebec, Canada, 2009: 937-944
- [47] Talwalkar A, Kumar S, Mohri M. Large-scale svd and manifold learning. *Journal of Machine Learning Research*, 2013, 14(1): 3129-3152
- [48] Borg I, Groenen P. Modern multidimensional scaling: Theory and applications. *Journal of Educational Measurement*, 2006, 40(3): 277-280
- [49] Tang L, Liu H. Relational learning via latent social dimensions//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 817-826
- [50] Newman M E. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2006, 103(23): 8577-8582

- [51] Newman M E. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E Statistical Nonlinear and Soft Matter Physics*, 2006, 74(3): 92-100
- [52] Sarkar P, Moore A W. Dynamic social network analysis using latent space models. *ACM Sigkdd Explorations Newsletter*, 2005, 7(2): 31-40
- [53] Yang J, Leskovec J. Modeling Information diffusion in implicit networks//*Proceedings of the 10th IEEE International Conference on Data Mining*. Sydney, Australia, 2010: 599-608
- [54] Bourigault S, Lagnier C, Lamprier S. Learning social network embeddings for predicting information diffusion//*Proceedings of the 7th ACM International Conference on Web Search and Data Mining*. New York, USA, 2014: 393-402
- [55] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, 521(7553): 436-444
- [56] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. *arXiv.org*, 2014, cs.CL: 1-11
- [57] Huang E H, Socher R, Manning C D. Improving word representations via global context and multiple word prototypes//*Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, USA, 2012: 873-882
- [58] Bengio Y, Schwenk H, Senécal J S. *Innovations in Machine Learning: Neural Probabilistic Language Models*. Berlin, Germany: Springer, 2006
- [59] Mikolov T, Chen K, Corrado G. Efficient estimation of word representations in vector space. *arXiv.org*, 2013, cs.CL: 1-12
- [60] Hinton G E. Learning distributed representations of concepts//*Proceedings of the 8th Annual Conference on the Cognitive Science Society*. Mass, USA, 1986: 1-12
- [61] Bengio Y, Ducharme R. A neural probabilistic language model. *Journal of Machine Learning Research*, 2003, 3(6): 1137-1155
- [62] Morin F, Bengio Y. Hierarchical probabilistic neural network language model//*Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*. Barbados, 2005: 246-252
- [63] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning//*Proceedings of the 25th International Conference on Machine Learning*. Helsinki, Finland, 2008: 160-167
- [64] Gutmann M, Hyvärinen A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *Journal of Machine Learning Research*, 2010, 9: 297-304
- [65] Gutmann M U, Hyv, Rinen A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 2012, 13(1): 307-361
- [66] Niepert M, Ahmed M, Kutzkov K. Learning convolutional neural networks for graphs//*Proceedings of the 33rd International Conference on Machine Learning*. New York, USA, 2016: 2014-2023
- [67] Hamilton W L, Ying R, Leskovec J. Inductive representation learning on large graphs. *arXiv preprint arXiv: 1706.02216*, 2017.
- [68] Tu C, Wang H, Zeng X. Community-enhanced network representation learning for network analysis. *arXiv.org*, 2016, cs.SI: 1-8
- [69] Tu C, Zhang W, Liu Z. Max-margin deepwalk: discriminative learning of network representation//*Proceedings of the 25th International Joint Conference on Artificial Intelligence*. California, USA, 2016: 3889-3895
- [70] Grover A, Leskovec J. Node2vec: Scalable feature learning for networks//*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, USA, 2016: 1-10
- [71] Wang D X, Cui P, Zhu W. Structural deep network embedding//*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, USA, 2016: 1225-1234
- [72] Perozzi B, Kulkarni V, Skiena S. Walklets: Multiscale graph embeddings for interpretable network classification. *arXiv.org*, 2016, cs.SI: 1-16
- [73] Lei C, Ruan J. A novel link prediction algorithm for reconstructing protein-protein interaction networks by topological similarity. *Bioinformatics*, 2013, 29(3): 355-364
- [74] Fire M, Tenenboim L, Lesser O. Link prediction in social networks using computationally efficient topological features//*Proceedings of the IEEE Third International Conference on Privacy, Security, Risk and Trust*. Maryland, USA, 2011: 73-80
- [75] Brandão M A, Moro M M, Lopes G R. Using link semantics to recommend collaborations in academic social networks//*Proceedings of the 22nd International Conference on World Wide Web*. Rio de Janeiro, Brazil, 2013: 833-840
- [76] Li Zhi-Yu, Liang Xun, Zhou Xiao-Ping. A link prediction method for large-scale networks. *Chinese Journal of Computers*, 2016, 39(42): 1-18(in Chinese)
(李志宇, 梁循, 周小平. 一种大规模网络中基于顶点结构特征映射的链接预测方法. *计算机学报*, 2016, 39(42): 1-18)
- [77] Steyvers M, Griffiths T. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 2007, 427(7): 424-440
- [78] Hofmann T. Probabilistic latent semantic indexing//*Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Sheffield, UK, 2004: 56-73
- [79] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003, 3: 993-1022
- [80] Brants T, Chen F, Tsochantaridis I. Topic-based document segmentation with probabilistic latent semantic analysis//*Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*. Virginia, USA, 2002: 211-218
- [81] Chang J, Blei D M. Relational topic models for document networks//*Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*. Florida, USA, 2009: 81-88

- [82] Nallapati R M, Ahmed A, Xing E P. Joint latent topic models for text and citations//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Nevada, USA, 2008: 542-550
- [83] Iwata T, Yamada T, Ueda N. Probabilistic latent semantic visualization: Topic model for visualizing documents//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 363-371
- [84] Le T M V, Lauw H W. Probabilistic latent document network embedding//Proceedings of the 14th IEEE International Conference on Data Mining. Shenzhen, China, 2014: 270-279
- [85] Wang X, Mccallum A. Topics over time: A non-Markov continuous-time model of topical trends//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 424-433
- [86] Lin S, Wang F, Hu Q. Extracting social events for learning better information diffusion models//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2013: 365-373
- [87] Tang J, Sun J, Wang C. Social influence analysis in large-scale networks//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 807-816
- [88] Lou T, Tang J. Mining structural hole spanners through information diffusion in social networks//Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil, 2013: 825-836
- [89] Hu Z, Yao J, Cui B. Community level diffusion extraction//Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. Melbourne, Australia, 2015: 1555-1569
- [90] Leskovec J, Kleinberg J, Faloutsos C. Graphs over time: Densification laws, shrinking diameters and possible explanations//Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2005: 177-187
- [91] Backstrom L, Huttenlocher D, Kleinberg J. Group formation in large social networks: Membership, growth, and evolution //Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 44-54
- [92] Toyoda M, Kitsuregawa M. Extracting evolution of web communities from a series of web archives//Proceedings of the 14th ACM Conference on Hypertext and Hypermedia. Pennsylvania, USA, 2003: 28-37
- [93] Bergerwolf T Y, Saia J. A framework for analysis of dynamic social networks//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 523-528
- [94] Falkowski T, Bartelheimer J, Spiliopoulou M. Mining and visualizing the evolution of subgroups in social networks//Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. Hong Kong, China, 2006: 52-58
- [95] Asur S, Parthasarathy S, Ucar D. An event-based framework for characterizing the evolutionary behavior of interaction graphs. *ACM Transactions on Knowledge Discovery from Data*, 2009, 3(4): 16
- [96] Gruhl D, Liben-Nowell D, Guha R. Information diffusion through blogspace. *ACM SIGKDD Explorations Newsletter*, 2004, 6(2): 43-52
- [97] Leskovec J, Mcglohon M, Faloutsos C. Cascading behavior in large blog graphs. *SIAM on Data Mining*, 2007, 15(1): 9:3-9:56
- [98] Kleinberg J. The small-world phenomenon: An algorithmic perspective//Proceedings of the 32nd Annual ACM Symposium on Theory of Computing. New York, USA, 2000: 163-170
- [99] Palla G, Derenyi I, Farkas I, Vicsek T. Uncovering the overlapping community structures of complex networks in nature and society. *Nature*, 2005, 435 (7043): 814-818
- [100] Lancichinetti A, Radicchi F, Ramasco J J. Finding statistically significant communities in networks. *PloS One*, 2011, 6(4): 336-338
- [101] Ahn Y Y, Bagrow J P, Lehmann S. Link communities reveal multiscale complexity in networks. *Nature*, 2010, 466(7307): 761-764
- [102] White S, Smyth P. A spectral clustering approach to finding communities in graphs//Proceedings of the 2005 SIAM International Conference on Data Mining. Philadelphia, USA, 2005: 274-285
- [103] Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 2009, 11(3): 19-44
- [104] Yang C, Liu Z Y, Zhao D L. Network representation learning with rich text information//Proceedings of the 24th International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 2111-2117
- [105] Zhang Z, Zhao K, Zha H. Inducible regularization for low-rank matrix factorizations for collaborative filtering. *Neurocomputing*, 2012, 97(1): 52-62
- [106] Yu H F, Jain P, Kar P. Large-scale multi-label learning with missing labels//Proceedings of the 31st International Conference on Machine Learning. Beijing, China, 2013: 593-601
- [107] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. *Journal of the Association for Information Science and Technology*, 2007, 58(7): 1019-1031
- [108] Ganesh J, Soumyajit G, Manish G. Author2vec: Learning author representations by combining content and link information//Proceedings of the 25th International World Wide Web Conference. Montréal, Canada, 2016: 1-2
- [109] Zhu X, Sobhani P, Guo H. Improved semantic representations from tree-structured long short-term memory networks. *Computer Science*, arXiv.org, 2015, cs.SI: 1-11
- [110] Cohn D. The missing link—A probabilistic model of document content and hypertext connectivity. *Advances in Neural Information Processing Systems*, 2001, 13: 430-436

- [111] Mei Q, Cai D, Zhang D. Topic modeling with network regularization// Proceedings of the 17th International Conference on World Wide Web 2008. Beijing, China, 2008; 101-110
- [112] Zhu X, Ghahramani Z, Lafferty J D. Semi-supervised learning using gaussian fields and harmonic functions// Proceedings of the 20th International Conference on Machine Learning. Washington, USA, 2003; 912-919
- [113] Li J, Ritter A, Dan J. Learning multi-faceted representations of individuals from heterogeneous evidence using neural networks. Computer Science, arXiv. org, 2015, cs. SI; 1-13
- [114] Sun X, Guo J, Ding X. A general framework for content-enhanced network representation learning. arXiv. org, 2016, cs. SI; 1-8
- [115] Pan S, Wu J, Zhu X. Tri-party deep network representation//Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA, 2016; 1895-1901
- [116] Tang L, Liu H. Leveraging social media networks for classification. Data Mining and Knowledge Discovery, 2011, 23(3): 447-478
- [117] Jacob Y, Denoyer L, Gallinari P. Learning latent representations of nodes for classifying in heterogeneous social networks//Proceedings of the 7th ACM International Conference on Web Search and Data Mining. New York, USA, 2014; 373-382
- [118] Zhao Y, Liu Z, Sun M. Representation learning for measuring entity relatedness with rich information//Proceedings of the 24th International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015; 1412-1418
- [119] Tang J, Qu M, Mei Q. Pte: Predictive text embedding through large-scale heterogeneous text networks//Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Sydney, Australia, 2015; 1165-1174
- [120] Chang S, Han W, Tang J. Heterogeneous network embedding via deep architectures//Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, Australia, 2015; 119-128
- [121] Niu X F, Li W J. ParaGraphE: A library for parallel knowledge graph embedding. arXiv preprint arXiv: 1703.05614, 2017
- [122] Albert R, Barabási A L. Statistical mechanics of complex networks. Review of Modern Physics, 2002, 74(1): 47-97
- [123] Dorogovtsev S N, Mendes J F F. Evolution of networks. Advances in Physics, 2002, 51(4): 1079-1187
- [124] Iwata T, Saito K, Ueda N. Parametric embedding for class visualization. Neural Computation, 2007, 19(9): 2536-2556



QI Jin-Shan, born in 1977, Ph. D. candidate. His research interests include social computing, data mining.

LIANG Xun, born in 1965, Ph. D., professor, Ph. D. supervisor. His research interests include neural networks,

support vector machine and social computing.

LI Zhi-Yu, born in 1991, Ph. D. candidate. His research interests include social computing, Web mining and natural language processing.

CHEN Yan-Fang, born in 1992, Ph. D. candidate. Her research interests include Web mining and natural language processing.

XU Yuan, born in 1993, M. S. candidate. Her research interests focus on social computing.

Background

The complex information network data have characteristics of large volume, high sparsity, high information fragility, and high dynamics. Therefore, network representation learning becomes a hot topic recently in network modeling. The distributed representation of learning networks has a practical significance value in applications. For example, the eigenvector representation of each vertex in the learning network can effectively alleviate network data sparsity; the heterogeneous information of different types in the network can be applied more effectively; the distributed vector representation of the network can achieve semantic correlation operation efficiently, and improve the computational efficiency significantly. In addition, the distributed representation of learning networks is of great value in the network vertex classification, recommendation system, link prediction, text modeling and visual processing.

This paper summarizes the main models of representation learning in current information networks, most of which are based on the information network link analysis (network structure) and the content or both fusions for vertex feature representation learning. In these models, the structural features of the fusion network and text content characteristics of the learning reflect a more real network. The learning network features are more meaningful. This paper also summarizes the key challenges, existing solutions and future research directions of the network representation learning. Based on the above analysis and discussions, this paper aims to provide some useful suggestions and inspirations for the future study on network representation learning models.

This paper is supported by the National Natural Science Foundation of China under Grant Nos. 71271211 and 71531012.