智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

异质信息网络中基于网络嵌入的影响力最大化

杨宇迪，周丽华，杜国王，邹星竹，丁海燕

引用本文:
杨宇迪, 周丽华, 杜国王, 等. 异质信息网络中基于网络嵌入的影响力最大化[J]. 智能系统学报, 2021, 16(4): 757–765.
YANG Yudi, ZHOU Lihua, DU Guowang, et al. Influence maximization based on network embedding in heterogeneous information networks[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(4): 757–765.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202009047

DOI: 10.11992/tis.202009047
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20210401.1300.004.html

异质信息网络中基于网络嵌入的影响力最大化

杨宇迪[1], 周丽华[1,2], 杜国王[1], 邹星竹[1], 丁海燕[1]
（1. 云南大学 信息学院, 云南 昆明 650504; 2. 云南大学 滇池学院, 云南 昆明 650228）

摘　要: 针对当前大部分影响力最大化算法忽略了异质信息网络包含多种节点类型和多种关系类型, 且不同类型节点在原始空间无法直接度量的问题, 提出了一种异质信息网络中基于网络嵌入的影响力最大化模型（influence maximization based on network embedding, IMNE）, 用于选择初始扩散节点实现影响力最大化。该模型不仅可以在对异质信息网络进行编码的同时表征异质信息网络中潜在的信息, 还可以捕获不同类型节点间影响力的不确定和复杂性。在 3 个真实数据集上的实验验证了 IMNE 算法的有效性。

关键词: 异质信息网络; 同质信息网络; 影响力最大化; 信息扩散; 网络嵌入; 直接影响力; 间接影响力; 全局影响力
中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1673−4785(2021)04−0757−09

中文引用格式: 杨宇迪, 周丽华, 杜国王, 等. 异质信息网络中基于网络嵌入的影响力最大化 [J]. 智能系统学报, 2021, 16(4): 757–765.
英文引用格式: YANG Yudi, ZHOU Lihua, DU Guowang, et al. Influence maximization based on network embedding in heterogeneous information networks[J]. CAAI transactions on intelligent systems, 2021, 16(4): 757–765.

Influence maximization based on network embedding in heterogeneous information networks

YANG Yudi[1], ZHOU Lihua[1,2], DU Guowang[1], ZOU Xingzhu[1], DING Haiyan[1]
(1. School of Information, Yunnan University, Kunming 650504, China; 2. Dianchi College, Yunnan University, Kunming 650228, China)

Abstract: Most current influence maximization algorithms ignore the problem that heterogeneous information networks contain multiple node types and relationship types, and different types of nodes cannot be measured in the original workspace. Accordingly, to solve these issues, this paper proposes a novel model for influence maximization based on network embedding in heterogeneous information networks, which helps to realize influence maximization by choosing initial diffusion nodes. The model can not only manifest the potential information in heterogeneous information networks while encoding it but also capture the uncertainty and complexity of influence among different types of nodes. Experimental results on three real datasets demonstrate the effectiveness of the proposed model.

Keywords: heterogeneous information network; homogeneous information network; influence maximization; information diffusion; network embedding; direct influence; indirect influence; global influence

影响力最大化问题是指在特定的扩散模型下, 寻找一组初始扩散节点使影响扩散范围最大的优化问题。目前, 影响力最大化算法主要分为两类: 贪心算法[1-3]和启发式算法[4-6], 其中贪心算法主要用于提高算法的精确度; 启发式算法主要用于解决实际问题以提高算法效率。Kempe 等[7]首次形式化定义了影响力最大化问题, 并提出了一个贪心算法, 其近似值约为 $1-/e^{-\varepsilon}$, 但是该算法的效率较低。Leskovec 等[8]提出了 CELF 算法, Goyal 等[9]提出了 CELF++算法, 通过实验发现该

收稿日期: 2020−09−30.　**网络出版日期**: 2021−04−01.
基金项目: 国家自然科学基金项目 (61762090, 62062066, 61966036); 国家社会科学基金项目 (18XZZ005); 云南省高等学校科技创新团队项目 (IRTSTYN); 云南省教育厅科学研究基金项目 (2021Y026).
通信作者: 周丽华. E-mail: lhzhou@ynu.edu.cn.

算法比 CELF 快 35%~55%。接着，为了解决贪心算法的效率问题，Tang 等[10] 提出一种基于跳的算法，该算法可以在常用的扩散模型 (独立级联模型和线性阈值模型) 下轻松应用于十亿规模的网络。Peng 等[11] 认为个体之间的影响有直接影响和间接影响，社会影响力的强弱取决于个体之间的关系、网络距离、时间、网络和个体的复杂性和不确定性。但这些算法通常仅将社会网络看作同质网络，忽略了社会网络的异质性，即网络中包含不同类型的对象和链接。在异质信息网络 (heterogeneous information network, HIN) 中，影响力可以通过不同的对象和链接进行扩散，从而获得更广泛的影响范围。

尽管异质信息网络丰富的结构和语义信息有助于实现影响力扩散范围最大化，但也给影响力的分析带来了挑战。目前，HIN 中的数据挖掘任务主要集中于相似性搜索[12-13]、聚类[14-15]、分类[16-17] 等，很少有研究者关注 HIN 中的社会影响力分析。为了建模 HIN 中的社会影响力，Yang 等[18] 提出了一种基于元路径的信息熵模型 MPIE，利用多条元路径从异质信息网络中提取多个同质信息网络，在每个同质信息网络中基于熵度量直接影响力和间接影响力，然后融合多个同质信息网络中的影响力度量 HIN 中的社会影响力。尽管 MPIE 取得了一定的效果，但该算法需选择特定类型的节点设定元路径，不能灵活地度量 HIN 中不同类型节点间的影响。

HIN 嵌入[19-21] 旨在通过基于节点的结构特性保留不同类型节点之间的邻近性来学习各个不同类型的节点的低维表示，而这些低维表示可直接应用于网络分析任务，比如节点分类、聚类以及链路预测等。由于 HIN 嵌入将不同类型节点映射于同一向量空间，用同一空间的低维向量来描述不同类型的节点，不仅概括了网络的重要结构特征，而且学习到的嵌入具有同质性，易于使用和集成。最近几年，异质信息网络的表征学习受到了研究者的广泛关注。

因此，本文提出了一种异质网络中基于网络嵌入的影响力最大化模型 IMNE。IMNE 首先利用网络嵌入学习 HIN 中所有节点在同一向量空间的低维表征，保持不同类型节点在同一度量空间，然后基于 HIN 原始的网络拓扑结构，扩展传统的信息熵模型，考虑多种影响因素，度量 HIN 中不同类型节点的社会影响力并选择最具影响力的节点作为初始扩散节点，最后，选择特定的信息扩散模型实现 HIN 中的影响力最大化。本文工作一方面扩展了 HIN 嵌入的应用，另一方面将 Peng 等[1] 所提的模型扩展到了 HIN。

1　相关符号和问题定义

定义 1　异质信息网络[22]。异质信息网络通常被定义为一个带有对象类型映射函数 $\phi: V \to A$ 和边类型映射函数 $\varphi: E \to R$ 的无向图 $G=(V,E)$，其中每个对象 $v \in V$ 属于一个特定的对象类型 $\phi(v) \in A$，每条边 $e \in E$ 属于一个特定的关系类型 $\varphi(e) \in R$，且 $|A|+|R|>2$。

定义 2　HIN 中的影响力最大化。给定一个异质信息网络 $G=\{V,E\}$，$\delta(V_s)$ 是将种子集 V_s 映射到受种子集影响的对象数量的影响函数，异质信息网络中影响力最大化的目的是选择一组最具影响力的种子集 $V_s^*(|V_s^*|=k)$，且该种子集可以最大化影响力的扩散范围，即

$$V_s^* = \arg\max_{V_s \subset V, |V_s|=k} \delta(V_s) \tag{1}$$

式中：种子集 V_s^* 可包含 HIN 中各类型的节点。

2　IMNE 模型

本节将详细介绍异质网络中基于网络嵌入的影响力最大化模型 IMNE。

首先以图 1 为例说明本文动机。如果将学术网络视为同质网络，该网络中仅 Lily 和 Bob、Ada 和 Tom 间存在边，此时，若将 Ada 作为初始扩散节点，其将仅影响 Tom 一个节点，即影响范围为 1。然而，若将学术网络视为 HIN，如图 2 所示，该网络中包含 3 种类型的节点，论文和会议之间存在发表/被发表关系，论文和论文之间存在引用/被引用关系，作者和论文之间存在撰写/被撰写关系，考虑不同类型节点通过不同关系互相影响，令 Ada 作为初始扩散节点，则 P2 和 P6 两篇论文会受到 Ada 的直接影响，其次，由于 Ada 和 Tom 之间通过路径“Tom-P2-Ada(合作)”相关联，Ada 和 Mary 通过“Ada-P6-C1-P5-Mary(共同参加会议)”相关联，所以 Ada 也可间接影响 Tom 和 Mary。相比同质网络，Ada 的影响范围更广。

IMNE 模型的整体框架如图 3 所示。首先，IMNE 从 HIN(图 3(a)) 中学习各种类型节点的嵌入 (图 3(b))，然后，扩展信息熵模型度量社会影响力，并选取最具影响力的节点作为种子集 (图 3(c))；最后在特定扩散模型下实现影响力最大化 (图 3(d))。

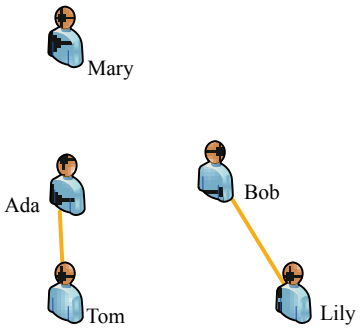


图 1　同质网络示例

Fig. 1　Example of a homogeneous network

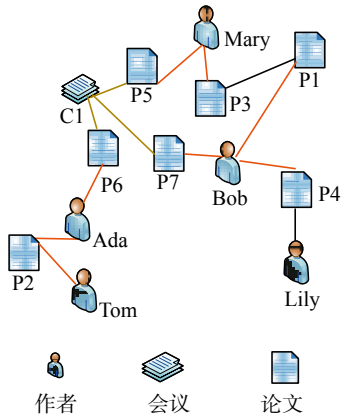


图 2　异质网络示例

Fig. 2　Example of a HIN

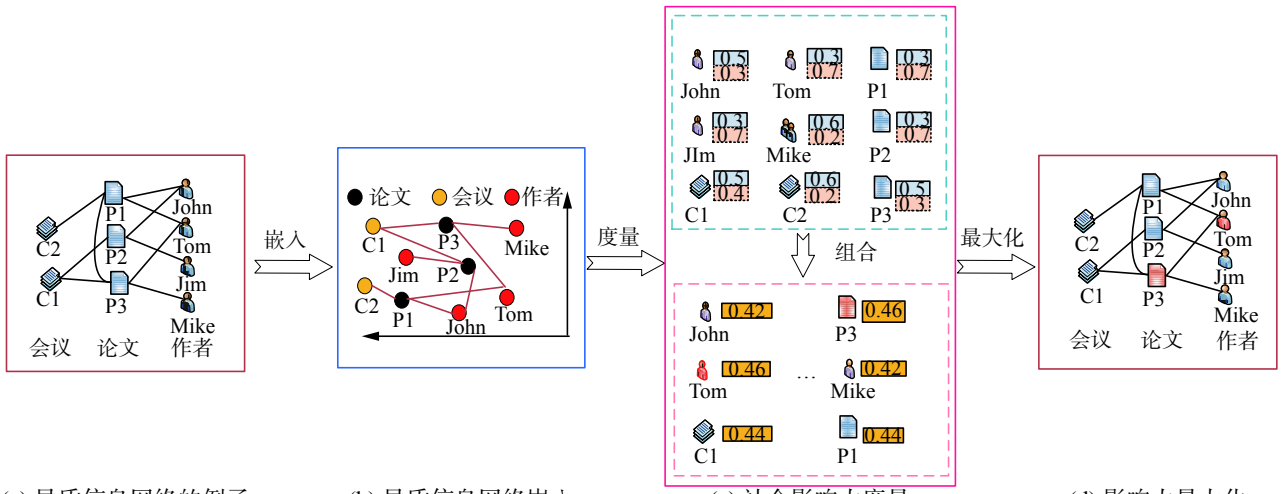
2.1　HIN 嵌入

本文使用 HIN2Vec 模型[23] 实现 HIN 嵌入。HIN2Vec 模型旨在通过最大程度地联合预测节点之间关系的可能性来学习节点向量和关系向量。模型采用一对节点 x 和 y，以及某种关系 $r \in R$ 的 one-hot 向量 $\boldsymbol{x}$、$\boldsymbol{y}$ 和 $\boldsymbol{r}$ 作为输入，通过神经网络将 $\boldsymbol{x}$、$\boldsymbol{y}$ 和 $\boldsymbol{r}$ 转化为隐藏层中的潜向量 $\boldsymbol{W}_X^{\mathrm{T}}\boldsymbol{x}$、$\boldsymbol{W}_Y^{\mathrm{T}}\boldsymbol{y}$ 和 $f_{01}(\boldsymbol{W}_R^{\mathrm{T}}\boldsymbol{r})$(HIN 中关系和节点的语义含义不同，所以关系向量 $\boldsymbol{r}$ 通过正则化转化为 $f_{01}(\boldsymbol{W}_R^{\mathrm{T}}\boldsymbol{r})$，限制 $\boldsymbol{r}$ 的值在 0 到 1 之间)，在输出层通过 Sigmoid $(\sum \boldsymbol{W}_X^{\mathrm{T}}\boldsymbol{x} \odot \boldsymbol{W}_Y^{\mathrm{T}}\boldsymbol{y} \odot f_{01}(\boldsymbol{W}_R^{\mathrm{T}}\boldsymbol{r}))$ 实现逻辑分类 ($\odot$ 表示逐元素相乘)。通过 HIN2Vec，HIN 中的每个节点转换为同一向量空间的低维度潜在表示，在捕获和表示 HIN 中的丰富信息的同时，有效避免了 HIN 中不同类型节点和关系的不兼容性，便于度量不同类型节点的社会影响力以选取初始扩散种子集。

2.2　影响力度量

HIN 中，一个对象的社会影响力通常不仅体现于紧密的直接联系，还体现在节点的间接联系。HIN 中社会影响的相关定义如下：

定义 3　直接/间接影响力。给定异质信息网络 G 中的对象 u、v，若对象 u 和 v 之间有边相连，即 $e_{uv}=1$，则 $De_u(v)$ 表示对象 u 和对象 v 间的直接影响力；若对象 u 和 v 之间没有边直接相连，即 $e_{uv}=0$，则 $II_u(v)$ 表示对象 u 和对象 v 间的间接影响力。



(a) 异质信息网络的例子　(b) 异质信息网络嵌入　(c) 社会影响力度量　(d) 影响力最大化

图 3　IMNE 模型的具体框架

Fig. 3　The specific framework of the IMNE model

定义 4　全局影响力。给定异质信息网络 G 中的节点 u，若节点 u 在整个网络中具有影响力，那么 I_u 被定义为节点 u 在 G 上的全局影响力。

全局影响力与直接/间接影响力有着密切关系。如果对象具有很强的直接和间接影响力，则该对象通常在社会网络中具有较强的全局影响力。IMNE 算法考虑了不同类型对象之间的影响(如作者对论文的影响或论文对作者的影响)，通常具有影响力的对象其相关行为也具有影响力。例如，在数据挖掘领域具有较强影响力的研究人员，他发表的论文在数据挖掘领域通常也具有较高的影响力。

2.2.1 直接影响力度量

在社交网络中，影响力与许多潜在因素相互作用，如相似性和相关性，相似对象之间的相互影响往往更强，例如，在学术网络中，作者 i 与作者 j、作者 m 均有合作，若作者 i 与作者 j 的研究领域更相似，那么作者 i 和 j 间的相互影响往往更强。同时，度中心性作为图论中度量节点相对重要性的评价指标，也是社会影响力的潜在影响因素。

根据 HIN 嵌入得到的各类型节点的嵌入信息，可获得 HIN 中各类型节点的相似度 Sim_{ij}，即：

$$\mathrm{Sim}_{ij}=\frac{\boldsymbol{e}_i\cdot\boldsymbol{e}_j}{\|\boldsymbol{e}_i\|\|\boldsymbol{e}_i\|} \tag{2}$$

式中：$\boldsymbol{e}_i$ 和 $\boldsymbol{e}_j$ 分别对应对象 i 和 j 的潜在表征向量；$(\cdot)$ 表示向量的点积；$\|\boldsymbol{e}\|$ 表示向量的长度。

HIN 网络嵌入不仅将不同类型节点映射于同一空间便于度量不同类型节点间的社会影响，还保留 HIN 原始结构，故给定异质信息网络 G，令 N_i 表示对象 i 的度，对象 i 的直接影响力的定义如下：

$$\mathrm{De}_i=-\sum_{i=1,j\in V}^{N_i}\left(\frac{1}{N_i}\lg\frac{1}{N_i}+\frac{\mathrm{Sim}_{ij}}{\sum_{k=1}^{N_i}\mathrm{Sim}_{ik}}\lg\frac{\mathrm{Sim}_{ij}}{\sum_{k=1}^{N_i}\mathrm{Sim}_{ik}}\right) \tag{3}$$

其中 $i\neq j$ 且 $i\neq k$。

2.2.2 间接影响力度量

除了直接影响力，对象之间往往也具有某种间接影响，例如，在学术网络中，具有影响力的作者可以通过论文或会议影响其他作者。

给定异质信息网络 G，令 N_{ik} 表示对象 i 和 k 间公共对象的数量，则对象 i 和 k 之间的间接影响力描述如下：

$$II_i(k)=\sum_{k=1}^{N_{ik}}\mathrm{De}_i\times\mathrm{De}_k/N_{ik} \tag{4}$$

令 M_i 表示对象 i 的多跳连接对象的数量，对象 i 的间接影响力 II_i 的定义如式 (5) 所示。

$$II_i=\frac{\sum_{k=1}^{M_i}II_{ik}}{M_i} \tag{5}$$

2.2.3 全局影响力度量

根据直接影响力和间接影响力的度量可得，对象 i 的全局影响力如式 (6) 所示。

$$I_i=\alpha\mathrm{De}_i+\beta II_i \tag{6}$$

式中：α 和 β 分别表示直接影响力 De_i 和间接影响力 II_i 的权重，且 $\alpha+\beta=1$。

2.3 IMNE 算法描述

算法 1 IMNE 算法

输入 异质信息网络 $G=\{V,E\}$；参数 α 和 β

输出 种子集 S

1) 初始化 $S=\mathrm{null}$；

2) 基于异质信息网络嵌入学习 G 中节点的表征向量；

3) For $i=1$ to k：

4) For each $v\in V$ do：

5) 利用公式 (3) 计算直接影响力 De_v；

6) 利用公式 (5) 计算间接影响力 II_v；

7) 利用公式 (6) 计算全局影响力 I_v；

8) End For

9) $u=\max_v I_v$；

10) $S\leftarrow S\cup\{u\}$；

11) End For

12) Return S

该算法中语句 3~11 用以选择最具影响力的 k 个对象作为种子节点集 S。IMNE 算法的复杂度为 $O(m+nd+n\log(k))$，其中 n 表示异质信息网络中对象的数量，$n=|V|$；k 表示最具影响力的对象数量；d 表示一跳链接对象的平均数量；m 表示异质信息网络中边的总数，$m=|E|$。

3 实验评估

3.1 实验准备

3.1.1 数据集

本文实验部分共使用了 3 个真实的 HIN 数据集，来自两种不同领域，其中 4-area 数据集来自学术领域，Yelp 数据集和 Amazon 数据集来自商业领域。4-area[24] 是从 DBLP 网站收集的子数据集，涉及数据库、数据挖掘、机器学习和信息检索 4 个研究领域，共包含 20 场会议、排名前 5 000 的作者、14 328 篇论文和 8 789 个术语，其中作者与论文、论文与会议、论文与术语之间存在联系。Yelp 数据集记录了 1 268 条用户对坐落于 47 个城市、3 种类型的 2 614 条商户的评分情况，其中用户与商户、商户与城市、商户与类型之间存在联系。Amazon 是一种商业网络，该网络记录了 6 170 个用户对来自 334 个品牌旗下的 22 种类别共 2 753 项产品的 3 857 条评论情况，其中用户与产品、产品与评论、产品与类别、产品与品牌之间存在联系。以上 3 个数据集，除了来自不同的领域，还具有不同的稀疏性，其中 Yelp 的数据分布最稀疏，Amazon 数据分布最密集。在实验过程中选择 HIN2Vec[13] 模型嵌入 HIN 数据集，使不同类型节点处于同一度量空间。

3.1.2 扩散模型

影响力最大化旨在正确识别一组种子集以使它们在特定扩散模型下影响力的扩散范围最大化。本文选择 SIR 模型[25] 和线性阈值模型共两

种模型作为扩散模型，在 SIR 模型实验中，感染概率 γ 设为 0.8，恢复概率 θ 设为 0.5，种子集大小设为 1~100；在线性阈值模型实验中，其阈值设置为 0.5，种子集大小设为 0~50。

3.1.3　对比算法

采用 DC、PageRank、Entropy-based、MPIE 和 IMNE-D 算法作为对比算法，验证 IMNE 算法实现 HIN 中影响力最大化的有效性。其中 DC、PageRank 算法是常见的影响力最大化中选择种子集的方法；Entropy-based 算法[11] 是同质网络中不仅考虑直接影响和间接影响，还考虑影响力扩散的一种方法，有助于对比异质信息对种子集选择的影响；其次，IMNE 算法除了实现 HIN 中的影响力最大化，还可以度量 HIN 中节点的社会影响力，而 MPIE 算法是基于元路径度量 HIN 中节点影响力的方法，因此，本文还选择 MPIE 算法[18] 作为对比算法，分析 IMNE 算法在 HIN 中度量节点社会影响力的效果。

在实验中，DC、PageRank 和 Entropy-based 方法将忽略 HIN 中节点和链接的异质性，直接在整个网络 (4-area、Yelp 和 Amazon) 上运行，实验结果包含各种不同类型的节点。IMNE-D 是 IMNE 算法的变体，主要通过计算节点的直接影响力来选择种子集，未考虑非直接相连的节点间的影响，有助于对比间接影响力对影响力最大化的作用。特别地，基准算法中的参数均选择实验结果最优参数。

3.1.4　评估指标

1) 在 SIR 模型实验中，将感染率 (Infection rate) 和感染时间 (Infection time) 作为评估指标验证 IMNE 算法的性能，其中感染率表示种子集在 SIR 模型下感染节点占所有节点的比例；感染时间代表种子集在 SIR 模型下达到最大感染率所需时间，验证 IMNE 算法是否在 SIR 扩散模型下使得影响力扩散范围在较短时间内达到最大。令 NI 表示在影响力扩散过程中从易感染状态转变为感染状态的节点数量，$|V|$ 表示整个 HIN 包含的节点数量，则感染率定义为

$$\text{infection rate} = \frac{\text{NI}}{|V|} \tag{7}$$

通常情况下，感染率越大、感染周期越短，算法性能越强。

2) 在线性阈值模型实验中，将使用种子节点激活的其他节点数目来评估不同算法的性能，激活的节点数量越多，表示该影响力最大化算法的性能越强。

3.2　IMNE 算法在 SIR 模型下的性能

本节将从感染率和感染时间两个方面分析 IMNE 算法在 SIR 模型下的有效性。

1) 最大感染率。

由于存在随机性，每次实验中被感染节点的数量通常不同，因此本实验将每组种子集在 SIR 模型上运行 50 次，取平均感染率作为该组种子集的感染率。图 4 显示了不同算法的 top-k 个不同类型的影响力节点获得的感染率。

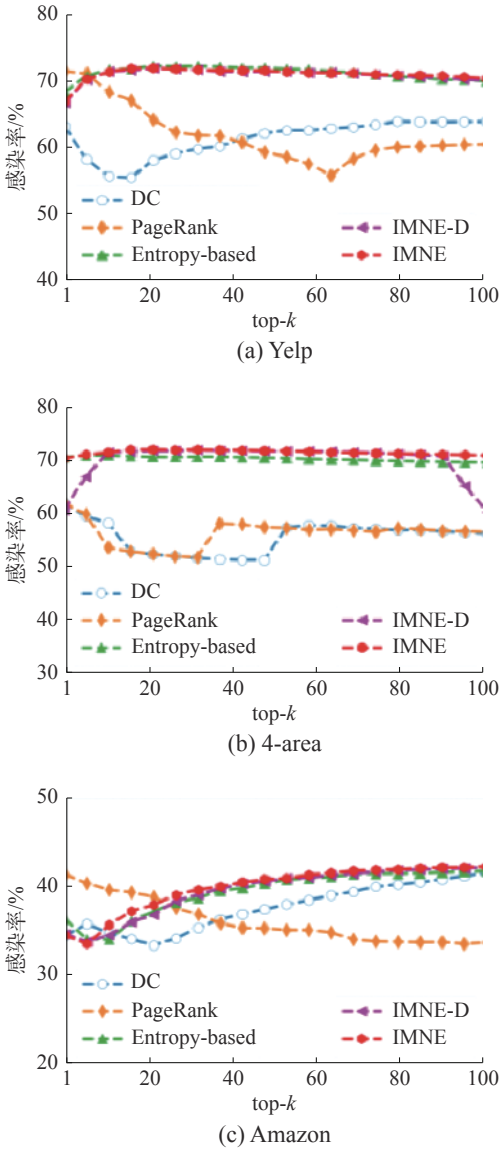


图 4　不同算法的 top-k 影响力节点的感染率比较

Fig. 4　Comparison of infection rate of different methods with different top-k influential nodes

首先，从图 4 可以看出，IMNE 算法得到的感染率优于基线算法 (DC、PageRank 和 Entropy-based)，这表明不同类型节点之间具有影响且这些异质信息有益于影响力最大化。其次，由于不同数据集数据的分布情况不同，相同的方法在不同的数据集上具有不同的感染率。特别地，DC 和 PageRank 更加依赖于数据的分布情况，例如 4-area 中节点度的差异比 Amazon 和 Yelp 更明显，因此，在 4-area 中 PageRank 的性能比在 Amazon 和

Yelp 更稳定。同时，还可观察到，当 k 值较小时，在 Yelp 数据集中，Entropy-based 优于 IMNE 算法，但是随着 k 值的增加，IMNE 算法的感染率逐渐优于 Entropy-based。

其次，关于 IMNE-D，其与 IMNE 的主要区别在于影响力度量组件，IMNE-D 仅考虑了直接链接节点之间的影响，忽略了社交网络中朋友的朋友之间的某种间接影响。根据结果 (IMNE> IMNE-D)，可以发现间接影响力对改善影响力的传播范围具有重要意义。

2) 达到最大感染率的时间。

影响力最大化不仅要求种子节点的影响范围最广，还要求在短时间内达到影响力扩散范围最广，因此，本实验从感染时间验证 IMNE 算法的性能。图 5 显示了不同模型的 top-k 影响力节点达到最大感染率的周期（周期即种子完成一次感染和恢复所需时间），可以看出 IMNE 算法在 3 个数据集上的影响力传播过程中，尤其是 Yelp 和 Amazon，IMNE 达到最大感染率的时间均小于其他基线算法，表示 IMNE 算法能在较短的时间内达到较大的感染率,即 IMNE 算法具有效性。在 4-area 数据集中，当 k 值超过 60 时，尽管 IMNE-D 感染时间小于 IMNE，但此时从图 4 可以发现 IMNE 算法的感染率大于 IMNE-D。

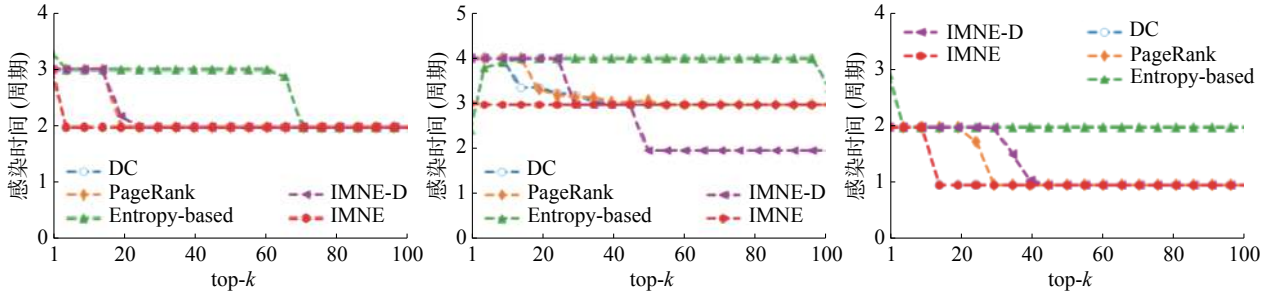


图 5　不同算法的 top-k 影响力节点感染时间比较

Fig. 5　Comparison of infection time of different methods with different top-k influential nodes

综上所述，通过分析最大感染率和达到最大感染率所需时间，可以发现，IMNE 算法相较于其他基准算法能在更短的时间实现影响力最大化。

3.3　IMNE 在线性阈值模型下的性能

为了更好地验证 IMNE 算法的有效性，本实验也验证了 IMNE 算法在线性阈值模型下的有效性。图 6 显示了在 4-area 和 Yelp 数据集上不同算法的 $k = (5, 10, 20, 30, 40, 50)$ 个有影响力作者的影响范围。

从图 6 可看出，IMNE 算法的影响范围均大于同质网络中的方法 (DC、Entropy-based 和 PageRank)，这表明 IMNE 算法在 LT 模型下也能较好地实现影响力最大化。其次，IMNE 算法的影响范围接近 MPIE 算法，这表明 IMNE 算法可以在不指定特定的元路径的情况下也能较为有效地度量 HIN 中节点的社会影响力。

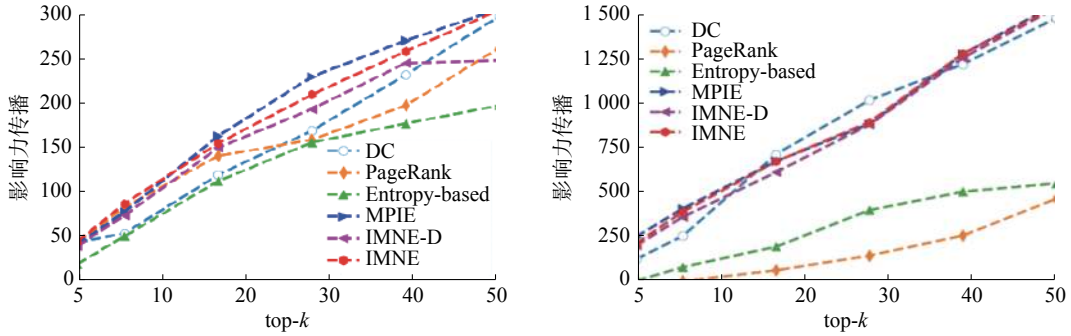


图 6　不同算法的 top-k 影响力节点影响范围

Fig. 6　Comparison of the range of top-k influential nodes of different methods

3.4　计算效率

本实验主要从节点社会影响力的计算时间分析 IMNE 算法的效率。表 1 展示了不同算法在 4-area、Yelp 和 Amazon 数据集上计算节点社会影响力耗费的时间。

首先，从表 1 可以看出，由于不同数据集的数据分布情况不同，不同算法在不同数据集上的计算时间不同，IMNE 算法和其他基准算法在

Yelp 数据集上的计算时间最少，4-area 数据集上　　的计算时间最多。

表 1　不同算法的计算时间比较

Table 1　Comparison of computation time of different algorithm　　s

数据集	DC	PageRank	Entropy-based	MPIE	IMNE
4-area	3.6055	11.2153	652.6717	10194.2154	6560.0586
Amazon	2.0362	9.7238	329.6371	9652.9749	1149.1622
Yelp	0.7461	4.8529	283.2934	750.2954	340.1706

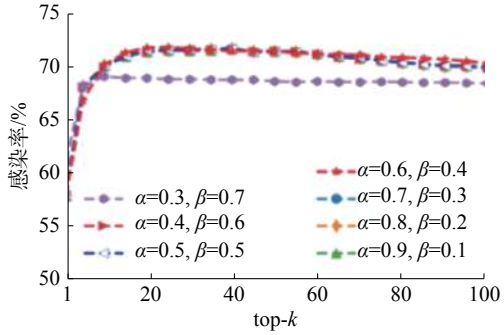
其次，IMNE 算法花费的计算时间高于同质网络中的方法 (DC、PageRank 和 Entropy-based)，这是因为 IMNE 算法既考虑了 HIN 中节点的异质性又度量了节点的直接影响力和间接影响力，而 DC、PageRank 和 Entropy-based 忽略了 HIN 中节点类型和边类型，仅做简单的计算；IMNE 算法的计算时间少于 MPIE 算法，这是因为 MPIE 算法需要迭代计算不同元路径下节点的社会影响力进行融合，而 IMNE 算法通过网络嵌入已将 HIN 映射于同一向量空间，节省了社会影响力的计算时间。

通过分析 IMNE 算法的计算效率和有效性可以发现，IMNE 算法不仅相较于其他基准算法能在更短的时间内实现影响力最大化，而且在社会影响力的计算效率上，也具有一定的优势。

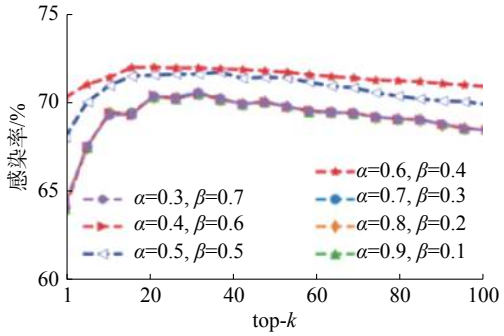
3.5　参数分析

1) 权重的影响。

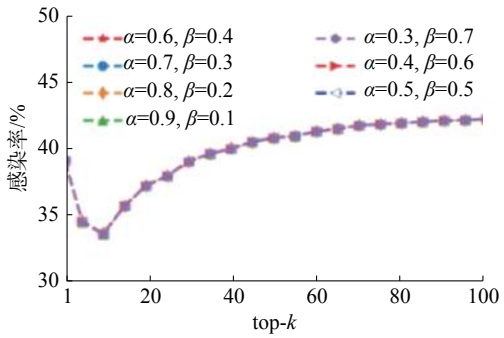
图 7 展示了 IMNE 算法中直接影响力和间接影响力的各种线性组合对影响力最大化的影响。从图 7(a)、(b) 可以看出直接影响力和间接影响力的权重分别为 0.6 和 0.4 时，感染率优于其他组合。这表明区分直接影响力和间接影响力对影响力最大化具有重要意义。此外，在 Amazon 数据集下，不同直接影响力和间接影响力权重组合，其感染率变化不明显，这是因为 Amazon 是一个密集数据集，在密集数据集中，具有较高直接影响力的节点其也具有较高的间接影响力。因此，本文将直接影响力和间接影响力的权重分别设置为 0.6 和 0.4。



(a) Yelp



(b) 4-area



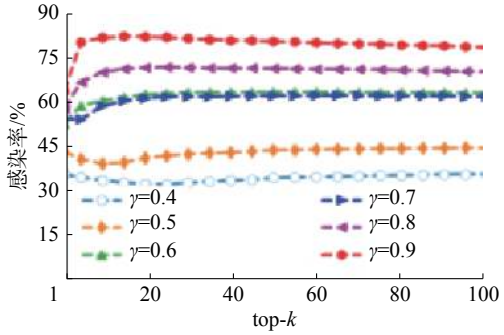
(c) Amazon

图 7　不同的权重对 IMNE 算法的影响

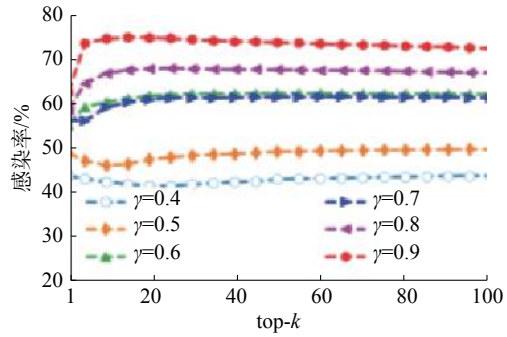
Fig. 7　Comparison of IMNE with different weight of direct and indirect influence on three datasets.

2) 感染概率 γ 的影响。

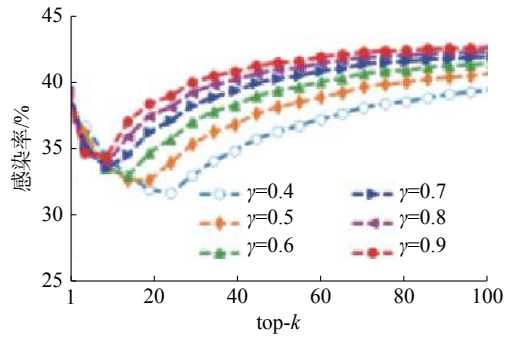
SIR 模型主要包含感染概率 γ 和恢复概率 θ 两个参数，本节实验测试了感染概率 $\gamma=(0.4,0.5,0.6,0.7,0.8,0.9)$ 对 IMNE 算法的影响，测试结果如图 8 所示。



(a) Yelp



(b) 4-area



(c) Amazon

图 8　SIR 中不同的感染概率对 IMNE 算法的影响

Fig. 8　Comparison of IMNE with different γ on three datasets.

从图 8 可以看出，随着感染概率 γ 的增加，感染节点的比例也增加，这与事实相符。但是，从图 8(a) 可以看出，当感染概率分别为 0.6 和 0.7 时，感染率接近，因此本文选择 0.8 作为感染概率。

4　结束语

本文提出了一种 HIN 中基于网络嵌入的影响力最大化算法 IMNE，该模型首先基于网络嵌入方法将不同类型的节点映射于低维向量空间，保留 HIN 的网络结构以及语义信息，使得不同类型节点处于同一度量空间，然后通过扩展传统信息熵模型度量 HIN 中不同类型节点的影响力选择最具影响力的节点作为种子集，实现了 HIN 中的影响力最大化。但本文选择了已知的 SIR 模型和线性阈值模型作为影响力扩散模型，未提出新的扩散模型，在将来的工作中，将考虑提出基于博弈论的扩散模型，不仅考虑网络结构对影响力扩散的影响，还考虑信息本身对影响力扩散的影响。

本文的主要贡献如下：

1) 提出了一种 HIN 中基于网络嵌入的影响力最大化模型 IMNE，该模型利用网络嵌入将 HIN 中所有节点映射于同一向量空间，不仅揭示了 HIN 中丰富的语义信息，还保留了更多的上下文信息，同时还解决了 HIN 中不同类型间的异质问题，保持不同类型节点处于同一度量空间；

2) 扩展传统的信息熵模型，考虑多种社会影响力的影响因素，度量 HIN 中不同类型节点的直接影响和间接影响，有效地描述了社会影响力的复杂性和不确定性。

3) 在 3 个真实数据集和两个扩散模型上评估了 IMNE 算法的性能，实验结果表明，IMNE 算法相较于其他基准算法能在更短的时间内实现影响范围最大。

参考文献：

[1] HEIDARI M, ASADPOUR M, FAILI H. SMG: fast scalable greedy algorithm for influence maximization in social networks[J]. Physica a: statistical mechanics and its applications, 2015, 420: 124–133.

[2] ORIEDI D, DE RUNZ C, GUESSOUM Z, et al. Influence maximization through user interaction modeling[C]//Proceedings of the 35th Annual ACM Symposium on Applied Computing. Brno, Czech Republic, 2020: 1888−1890.

[3] LIU Wei, CHEN Ling, CHEN Xin, et al. An algorithm for influence maximization in competitive social networks with unwanted users[J]. Applied intelligence, 2020, 50(2): 417–437.

[4] KIANIAN S, ROSTAMNIA M. An efficient path-based approach for influence maximization in social networks[J]. Expert systems with applications, 2021, 167: 114–168.

[5] SHANG Jiaxing, ZHOU Shangbo, LI Xin, et al. CoFIM: a community-based framework for influence maximization on large-scale networks[J]. Knowledge-based systems, 2017, 117: 88–100.

[6] 胡庆成, 张勇, 邢春晓. 基于有重叠社区划分的社会网络影响最大化方法研究 [J]. 计算机科学, 2018, 45(6): 32–35.
HU Qingcheng, ZHANG Yong, XING Chunxiao, et al. K-clique heuristic algorithm for influence maximization in social network[J]. Computer science, 2018, 45(6): 32–35.

[7] KEMPE D, KLEINBERG J, TARDOS É. Maximizing the spread of influence through a social network[C]//Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2003: 137−146.

[8] LESKOVEC J, KRAUSE A, GUESTRIN C, et al. Cost-effective outbreak detection in networks[C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA, 2007: 420−429.

[9] GOYAL A, LU Wei, LAKSHMANAN L V S. CELF^{++}: optimizing the greedy algorithm for influence maximization in social networks[C]//Proceedings of the 20th International Conference Companion on World Wide Web. Hyderabad, India, 2011: 47−48.

[10] TANG Jang, TANG Xueyan, YUAN Junsong. An efficient and effective hop-based approach for influence maximization in social networks[J]. Social network analysis and mining, 2018, 8(1): 10.

[11] PENG Sancheng, YANG Aimin, CAO Lihong, et al. Social influence modeling using information theory in mobile social networks[J]. Information sciences, 2017, 379: 146–159.

[12] SHI Chuan, KONG Xiangnan, HUANG Yue, et al. HeteSim: a general framework for relevance measure in heterogeneous networks[J]. IEEE transactions on knowledge and data engineering, 2014, 26(10): 2479–2492.

[13] WANG Chenguang, SUN Yizhou, SONG Yanglei, et al. RelSim: relation similarity search in schema-rich heterogeneous information networks[C]//Proceedings of the 2016 SIAM International Conference on Data Mining. Philadelphia, USA, 2016: 621−629.

[14] CHEN Lu, GAO Yunjun, ZHANG Yuanliang, et al. Efficient and incremental clustering algorithms on star-schema heterogeneous graphs[C]//Proceedings of 35th International Conference on Data Engineering. Macao, China, 2019: 256−267.

[15] CHEN Junxiang, DAI Wei, SUN Yizhou, et al. Clustering and ranking in heterogeneous information networks via gamma-Poisson model[C]//Proceedings of the 2015 SIAM International Conference on Data Mining. Vancouver, Canada, 2015: 425−432.

[16] BANGCHAROENSAP P, MURATA T, KOBAYASHI H, et al. Transductive classification on heterogeneous information networks with edge betweenness-based normalization[C]//Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. San Francisco, USA, 2016: 437−446.

[17] WAN Mengting, OUYANG Yunbo, KAPLAN L, et al. Graph regularized meta-path based transductive regression in heterogeneous information network[C]//Proceedings of the 2015 SIAM International Conference on Data Mining. Vancouver, Canada, 2015: 918−926.

[18] YANG Yudi, ZHOU Lihua, JIN Zhao, et al. Meta path-based information entropy for modeling social influence in heterogeneous information networks[C]//Proceedings of the 20th IEEE International Conference on Mobile Data Management. Hong Kong, China, 2019: 557−562.

[19] LIU Zemin, ZHENG V W, ZHAO Zhou, et al. Distance-aware DAG embedding for proximity search on heterogeneous graphs[C]//Proceedings of 32nd AAAI Conference on Artificial Intelligence (AAAI). New Orleans, USA, 2018: 2355−2362.

[20] SHI Chuan, HU Binbin, ZHAO W X, et al. Heterogeneous information network embedding for recommendation[J]. IEEE transactions on knowledge and data engineering, 2019, 31(2): 357–370.

[21] SHI Yu, ZHU Qi, GUO Fang, et al. Easing embedding learning by comprehensive transcription of heterogeneous information networks[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK, 2018: 2190−2199.

[22] LUO Chen, GUAN Renchu, WANG Zhe, et al. HetPathMine: a novel transductive classification algorithm on heterogeneous information networks[C]//Proceedings of the 36th European Conference on Information Retrieval. Amsterdam, The Netherlands, 2014: 210−221.

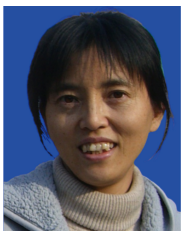
[23] FU Taoyang, LEE W C, LEI Zhen. HIN2Vec: explore meta-paths in heterogeneous information networks for representation learning[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore, 2017: 1797−1806.

[24] SUN Yizhou, HAN Jiawei, JING Gao, et al. iTopicModel: information network-integrated topic modeling[C]//Proceedings of the 9th IEEE International Conference on Data Mining. Miami, USA, 2009: 493−502.

[25] 马知恩, 周义仓, 王稳地, 等. 传染病动力学的数学建模与研究 [M]. 北京: 科学出版社, 2014.

作者简介：

杨宇迪，硕士研究生，主要研究方向为社会网络分析、数据挖掘。



周丽华，教授，博士生导师，CCF 会员，主要研究方向为数据挖掘、多视角学习、异质社交网络分析。主持国家自然科学基金项目 3 项、云南省重点基金项目 1 项。发表学术论文 80 余篇，出版学术著作 2 部。

杜国王，博士研究生，主要研究方向为数据挖掘、多视角聚类。