中国科学: 信息科学 2020 年 第 50 卷 第 2 期: 221–238
SCIENTIA SINICA Informationis
论文

基于层次化混合特征图的链路预测方法

李冬[1,2], 申德荣[1*], 寇月[1], 林梦儿[1], 聂铁铮[1], 于戈[1]

1. 东北大学计算机科学与工程学院, 沈阳 110004
2. 东软集团股份有限公司, 沈阳 110179
* 通信作者. E-mail: shenderong@cse.neu.edu.cn

收稿日期: 2018–08–19; 接受日期: 2018–11–15; 网络出版日期: 2020–02–12

国家重点研发计划课题 (批准号: 2018YFB1003404)、国家自然科学基金 (批准号: 61472070, U1435216, 61672142) 和中国国家留学基金委 (批准号: 201806085016) 资助项目

摘要 现实世界中的实体连同关联关系构成了一种网络关系结构即异构信息网络. 利用链路预测技术可以预测出异构信息网络中存在但未被观察到, 或者未来可能会出现的链路, 更好地帮助用户理解网络的结构生成和演化规律. 然而, 目前链路预测技术缺乏对多种特征的有效融合而影响预测准确性, 且难以适应异构信息网络的异构性和动态性. 本文提出了一种层次化混合特征图模型 (hierarchical hybrid feature graph, HHFG), 充分考虑了异构信息网络的拓扑特征、语义特征和时序特征. 提出了一种基于 HHFG 的链路预测算法, 基于混合特征在 HHFG 上做随机游走, 并采用梯度下降法学习特征权重, 转移系数等参数, 有效地保证了链路预测的准确性. 通过实验验证了本文所提出的关键技术的可行性和有效性.

关键词 链路预测, 层次化混合特征图, 异构信息网络, 随机游走, 参数学习

1 引言

在现实世界中, 实体连同关联关系构成了一种网络关系结构即信息网络. 这里的实体可以是商品、文章、会议、人、图片、电影或者导演等个体, 关联关系可以是发表、购买、观看、出演或者指导等关联. 链路预测是指通过已知的网络节点以及网络结构等信息来预测网络中尚未产生连边的两个节点之间产生链接的可能性. 这些被预测出的链路可以是实际存在但未被观察到的链路, 也可以是未来可能会出现的链路. 链路预测已经成为数据挖掘领域中一个重要的研究方向, 它可以更好地帮助用户理解网络的结构生成和演化规律.

随着实体类别的多样化和实体间关联关系的复杂化, 信息网络正在向异构化方向发展, 即形成了异构信息网络. 这种异构信息网络在现实生活中是普遍存在的, 例如学术论文网络 DBLP、电影资料

引用格式: 李冬, 申德荣, 寇月, 等. 基于层次化混合特征图的链路预测方法. 中国科学: 信息科学, 2020, 50: 221–238, doi: 10.1360/N112018-00223
Li D, Shen D R, Kou Y, et al. Research on a link-prediction method based on a hierarchical hybrid-feature graph (in Chinese). Sci Sin Inform, 2020, 50: 221–238, doi: 10.1360/N112018-00223

网络 IMDB, 以及图片共享网络 Flickr 等. 在新形势下, 异构信息网络上的链路预测呈现如下特点. 首先, 异构信息网络中的实体类型具有多样化的特点. 例如论文与书籍、专家与项目均为不同类型的对象. 其次, 异构信息网络中的关联关系具有多元化的特点. 例如两个人之间可以具有项目合作关联, 也可以具有论文合作关联. 第三, 异构信息网络随着时间、节点位置的变化而具有动态性, 主要表现为节点本身信息发生改变、节点加入退出等因素造成的网络结构变化等. 信息网络的动态变化使链路预测结果必然会受到影响, 甚至可能衍生出新的链路类型.

尽管当前的链路预测技术有很多, 但面对上述特点, 仍存在不足. 一方面, 这些技术通常将所有实体和关联关系都同等对待或独立分析, 忽略了不同实体类型及关联关系之间的相关性. 另一方面, 已有技术通常仅考虑单一特征 (如: 网络拓扑特征、实体属性特征或时序特征) 来评估节点之间产生链接的可能性, 缺乏对这些特征的有效融合. 此外, 这些技术一般只适用于静态网络, 而异构信息网络具有动态性, 单一时刻的预测结果不足以说明网络的结构生成和演化规律.

因此, 异构信息网络中的实体和关联关系可表现出更为丰富的特征, 如果在链路预测时能够有效地利用并融合这些特征, 将有助于提升预测的准确性. 针对异构信息网络的链路预测面临一些新的挑战. 首先, 如何同时考虑不同类型的实体和链接的信息, 如何同时考虑实体及其关联关系在拓扑、语义、时序等多种不同特征, 如何对这些信息和特征进行模型化表示. 其次, 如何有效地为不同的特征设置合适的权重, 以有效地评估各个特征对于链路预测的重要性. 第三, 如何充分地利用实体及其关联关系的时序特征来预测某时段链路存在的可能性, 以适应异构信息网络的动态性.

针对上述问题, 本文提出了一种基于层次化混合特征图 (hierarchical hybrid feature graph, HHFG) 的链路预测方法, 主要贡献点包括:

(1) 提出了一种层次化混合特征图模型 HHFG. 不同于传统的链路预测技术, HHFG 利用实体特征和边特征来区分不同类型的实体及关联关系, 并将异构信息网络的拓扑特征、语义特征和时序特征进行层次化表示, 对异构信息网络所表达的丰富特征进行有效组织.

(2) 提出了一种基于 HHFG 的链路预测算法. 一方面, 算法基于混合特征在 HHFG 上做随机游走, 通过计算随机游走概率来评估节点之间产生链接的可能性. 另一方面, 采用梯度下降法学习特征权重和转移系数等参数, 有效地保证了链路预测的准确性.

(3) 通过实验验证了本文所提出的关键技术的可行性和有效性.

本文第 2 节介绍链路预测的相关工作; 第 3 节对本文的研究问题进行定义; 第 4 节提出层次化混合特征图模型; 第 5 节提出基于 HHFG 的链路预测算法; 第 6 节为实验部分, 对提出的模型及算法进行测试; 最后对全文进行总结, 并指出下一步研究计划.

2 相关工作

本节首先介绍了链路预测的相关工作, 然后将本文提出的关键技术与这些技术进行了比较. 按照所考虑特征的不同, 链路预测技术可分为: 基于拓扑特征的链路预测、基于语义特征的链路预测、基于时序特征的链路预测, 以及基于混合特征的链路预测.

基于拓扑特征的链路预测是从网络拓扑结构角度出发, 计算网络中节点之间的相似性 [1]. 这类方法利用共同邻居 [2]、Jaccard 指标 [3] 等相似性指标来衡量节点的相似性. 例如, 文献 [2] 提出了基于邻域的链路预测方法, 其本质是两个节点如果有较多的共同邻居, 则它们之间更倾向于存在链接. 文献 [4] 基于深度学习理论提出了一种条件时间受限玻尔兹曼机 (conditional temporal restricted Boltzmann machine, ctRBM), 考虑了节点自身及其邻域的拓扑特征, 并利用 ctRBM 进行特征学习. 文

献 [5] 提出了一种网络嵌入方法, 基于网络拓扑特征将节点和边投影到低维向量空间, 通过计算向量的相似度来进行链路预测. 此外, 一些文献提出了基于路径的链路预测方法, 采用的相似性指标包括 LP 指标 [6]、随机游走 [7] 和 SimRank 指标 [8] 等.

基于语义特征的链路预测是通过计算两个节点的属性相似度来评估其链接概率. 例如, 文献 [9] 针对社交网络提出了一种链路预测方法, 通过考察两个节点之间的年龄、职业、教育、兴趣、地理位置、性别、信仰等属性的相似程度, 对节点对之间产生联系或者节点对之间关系的演化做出预测. 文献 [10] 在此基础上, 针对用户属性缺失的问题提出了标签补充策略. 文献 [11] 利用网络上用户收藏 (发表) 的内容、参与的话题等信息来计算用户之间的话题相似性并进行好友预测.

基于时序特征的链路预测考虑了链路的生成时间、出现频率、变化趋势等时序特征. 例如, 文献 [12] 提出了一种时间序列模型 ARIMA 来预测未来链路出现的频率. 但该方法仅针对已有链接的出现频率进行预测, 而无法针对链路的有无及新链接进行预测. 文献 [13] 将两个节点的最近活动时间作为它们间边的权值, 并基于这些权值进行节点相似性度量. 文献 [14] 研究了异构信息网络中关系建立时间的预测问题, 使用广义线性模型来预测链路是否形成以及链路形成的时间. 文献 [15] 利用社交向量时钟来描述节点交互的顺序及时间间隔, 以此预测两个节点未来交互的可能性. 文献 [16] 提出了基于序列行为的链路预测方法, 通过累积信息来预测序列中下一个或多个时刻的链路.

基于混合特征的链路预测是将多种特征相结合, 综合评估节点间的相似性. 目前, 比较常见的是将拓扑特征与语义特征相结合来进行链路预测. 例如, 文献 [17] 将语义特征建模为实体属性图, 并利用网络的拓扑特征在实体属性图上进行随机游走, 以此计算节点间形成链路的概率. 文献 [18] 首次提出了元路径的概念, 一方面通过不同元路径来表达不同的语义, 进而描述不同类型节点之间的接近度; 另一方面也考虑了网络模式等拓扑特征. 文献 [19] 提出了一种基于元路径的合作关系预测方法, 在考虑作者共同邻居数, 共同发表的会议数等拓扑特征的同时, 还考虑了其研究主题的语义相似性. 文献 [20] 提出了一种异构信息网络链路预测模型, 通过组合对象之间在不同元路径上建立链接的概率来进行链路预测. 文献 [21] 将语义特征和拓扑特征相结合, 提出了一种基于广义关系主题模型的链路预测方法, 可有效解决非对称网络上的链路预测问题. 文献 [22] 提出一种 Bayes 深度学习模型, 将节点属性特征和拓扑特征作为该模型中的隐变量进行学习. 文献 [23] 将链接预测看作一个有监督矩阵去噪问题, 基于用户特征和网络拓扑结构来最小化权重矩阵范数. 除此之外, 一些文献将网络表示学习技术应用于异构信息网络分析当中. 例如, 文献 [24] 提出了一种异构信息网络表示模型, 考虑了异构信息网络的拓扑结构和关联关系所具有的语义信息. 首先按照节点间关联关系的不同, 将异构信息网络分解为多个子空间; 然后利用网络表示技术, 对各个子空间进行训练, 从而得到节点的向量化表示; 最后通过计算向量间相似度来评估节点间产生链接的概率. 文献 [25] 提出了一种基于用户时空数据和语义信息的社区发现方法, 该方法可应用于用户行为 (如: 签到行为、交友行为等) 的预测. 文献 [26, 27] 考虑了用户间关系、事件内容及时空信息, 提出了基于用户社交关系的事件推荐模型, 有效解决了冷启动问题, 可应用于用户与事件间参与关系的预测.

本文提出的链路预测技术与上述技术的不同之处在于:

(1) 传统的链路预测技术主要作用于同构信息网络上的预测, 虽然一些文献针对异构信息网络上的链路预测技术进行了研究, 但通常将所有实体和关联关系同等对待或独立分析, 而忽略了不同类型实体及关联之间的相关性. 本文提出了一种层次化混合特征图模型, 利用实体特征和边特征来区分不同类型的实体及关联关系, 充分地利用了异构信息网络所表达的丰富特征.

(2) 传统的链路预测技术侧重于考虑拓扑特征、语义特征或时序特征, 由于考虑的因素有限, 很难保证预测的准确性. 虽然一些文献同时考虑了拓扑特征与语义特征, 但仅局限于简单的语义范畴 (如:

语义相似、元路径中边的语义等), 且没有将不同类型的特征加以区分, 也无法适应异构信息网络的动态性. 本文充分考虑了实体的拓扑特征、语义特征与时序特征, 将这些特征以 HHFG 模型进行有效组织, 其层次化特征可清晰表示不同特征之间的相关性. 通过融合这些特征来计算随机游走概率, 以此来综合评估节点之间产生链接的可能性.

(3) 虽然一些文献基于图模型来表达节点之间的连边概率, 并根据模型中概率依赖关系进行链路预测, 但在计算转移概率时并没有区分连边类型. 然而, 不同类型的连边的重要性应该是不同的, 如果被同等对待, 必然会影响预测的准确率. 本文提出了一种基于 HHFG 的链路预测算法: 一方面, 按照连边两侧端点的不同, 将连边划分为不同类型, 并提出了不同的转移概率计算策略; 另一方面, 采用梯度下降法学习特征权重和转移系数等参数, 有效地保证了链路预测的准确性.

(4) 利用网络表示学习技术可以将大规模的异构信息网络进行低维度表示, 可有效降低网络分析 (包括分类、聚类、链路预测、可视化等) 的代价. 但目前大多数文献面向特定应用需求 (如: 事件推荐、社区划分等), 侧重于解决特定链路类型 (如: 签到关系、好友关系) 的预测问题. 这些文献虽然在考虑拓扑特征、语义特征的同时, 考虑了时序特征, 但忽略了这些特征间的相互影响及网络的演化规律. 本文综合考虑了拓扑特征、语义特征和时序特征, 利用层次间的跳转来衡量它们之间的相互影响, 并分析了网络的演化结构生成和演化规律, 可针对异构信息网络中不同类型的链路进行预测.

3 问题定义

本节先介绍几个相关概念, 然后针对异构信息网络的链路预测问题进行形式化定义.

定义1 (实体) 实体是指现实世界中客观存在的并可以相互区分的个体, 每个实体具有某种实体类型, 且与一组属性取值相对应.

现实生活中实体多种多样, 且不是孤立存在的, 将多样化的实体及其关联关系以图数据的形式表示出来即形成了异构信息网络. 例如, DBLP 主要包含了论文、作者、会议、关键词, 以及出版社等不同类型的实体, 它们及其关联关系 write, written, publish, published, press, pressed, ... 构成了一个异构信息网络.

定义2 (异构信息网络) 异构信息网络以实体为核心, 用来表示实体之间的关联关系, 用四元组 $G=(V,E,F_V,F_E)$ 表示. 其中, V 表示所有的实体节点, E 表示实体间的关联边集合, F_V 用来表示实体的特征集合 (如: 实体类型、描述属性等), F_E 用来表示边的特征集合 (如: 边类型、建立时间等). 存在一个实体特征映射函数 $\Psi_V: V\to F_V$ 和一个边特征映射函数 $\Psi_E: E\to F_E$, 其中任一节点 $v\in V$ 的实体特征为 $\Psi_V(v)\in F_V$, 任一边 $e\in E$ 的边特征为 $\Psi_E(e)\in F_E$.

定义3 (链路预测) 给定一个异构信息网络 $G=(V,E,F_V,F_E)$, 源节点集 S 和整数 k, 链路预测的目标是学习一个预测函数 P, 该函数能够为每个源节点 $s\in S$ 预测出可链接的目标节点集 N_s ($N_s\subseteq V$, 且 $|N_s|=k$).

$$P:(G,S,k)\longrightarrow\{N_s\}. \tag{1}$$

预测函数将针对每个源节点 $s\in S$ 与 V 中各个节点之间的链接输出一个预测概率值. 按照概率值由高到低的顺序, N_s 将截取前 top-k 个节点作为预测结果.

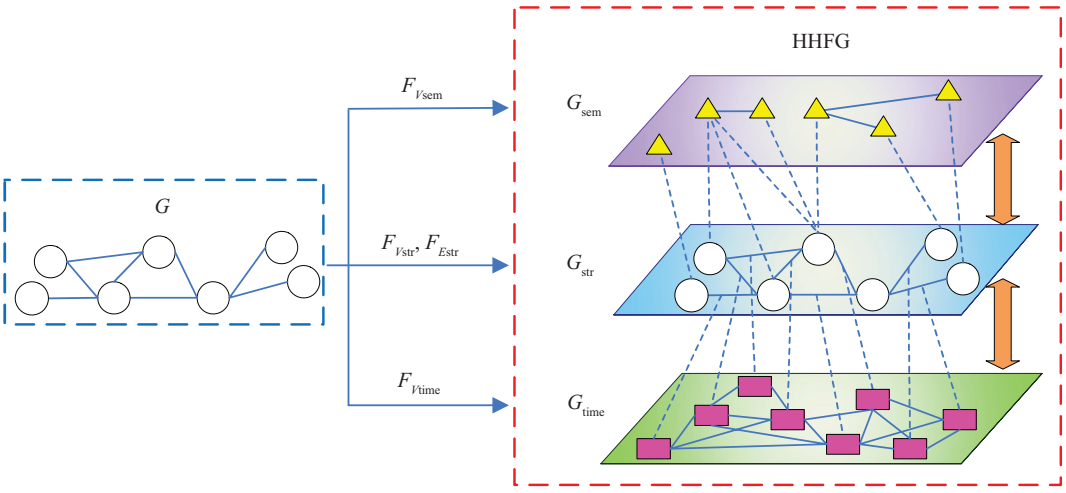

图 1 (网络版彩图) 层次化混合特征图模型 HHFG

Figure 1 (Color online) Hierarchical hybrid feature graph (HHFG)

4 层次化混合特征图模型

为了充分利用并整合异构信息网络的多种特征, 以提高链路预测的准确性, 我们提出了一种层次化混合特征图模型 HHFG. 本节首先介绍模型的基本思想, 然后提出直接关联强度的量化方法, 最后针对特征选取策略进行讨论.

4.1 HHFG 模型的基本思想

本文基于异构信息网络 G 来构建 HHFG 模型, 该模型分为语义特征层、拓扑特征层和时序特征层 3 个层次 (如图 1 所示).

- 语义特征层用来描述实体属性、实体类型等语义特征及其它们之间的语义相关性, 表示为 $G_{\text{sem}} = (V_{\text{sem}}, E_{\text{sem}}, F_{V\text{sem}})$. 其中, V_{sem} 表示语义节点集, 每个语义节点代表一个具体的属性值或实体类型值. E_{sem} 表示边集合, 若两个语义节点是语义相关的, 则它们之间存在边. $F_{V\text{sem}}$ 表示语义特征集, 存在一个语义特征映射函数 $\Psi_{\text{sem}} : V_{\text{sem}} \to F_{V\text{sem}}$, 其中任一语义节点 $v \in V_{\text{sem}}$ 的语义特征为 $\Psi_{\text{sem}}(v) \in F_{V\text{sem}}$.
- 拓扑特征层用来描述异构信息网络的拓扑结构, 其节点和边等同于异构信息网络中的节点和边, 用四元组 $G_{\text{str}} = (V, E, F_{V\text{str}}, F_{E\text{str}})$ 表示. 这里的 $F_{V\text{str}}$ 和 $F_{E\text{str}}$ 主要用来描述实体和边的拓扑特征 (包括共同邻居、节点入度、节点出度、网络稠密度等).
- 时序特征层用来描述实体间连边的时序特征, 表示为 $G_{\text{time}} = (V_{\text{time}}, E_{\text{time}}, F_{V\text{time}})$. 其中, V_{time} 表示时序节点集, 每个节点代表某对实体在时间序列中的交互情况. E_{time} 表示边集合, G_{time} 中两节点有一条边相连当且仅当它们在拓扑特征层所对应的边在 G_{str} 中相邻, 即 G_{time} 是 G_{str} 的边图. $F_{V\text{time}}$ 表示时序特征集, 存在一个时序特征映射函数 $\Psi_{\text{time}} : V_{\text{time}} \to F_{V\text{time}}$, 其中任一时序节点 $v \in V_{\text{time}}$ 的时序特征为 $\Psi_{\text{time}}(v) \in F_{V\text{time}}$.

例如, 在学术论文网络中, 为了预测作者之间的合作关系, 可将作者、论文作为实体, 即拓扑特征层中的节点集合 V, 论文与作者之间的发表关系、作者与作者间合作关系构成了拓扑特征层中的边集合 E. 这些作者和论文可以通过姓名、所在机构、论文名、主题词、出处等属性特征来描述, 这些属性

特征构成了语义特征层的节点集合 V_{sem}, 它们之间的包含、等价、相似等关联关系构成了语义特征层的边集合 E_{sem}. 将论文发表时间、作者合作时间作为时序特征, 这些特征构成了时序特征层中的节点集合 V_{time}, 并根据它们在拓扑特征层中的毗邻情况来构建时序特征层中的边集合 E_{time}.

HHFG 模型具有如下优点:

首先, HHFG 模型将异构信息网络的拓扑特征, 语义特征与时序特征以层次化形式进行组织, 充分地利用了异构信息网络所表达的丰富特征. 其中, 语义特征层是基于异构信息网络中的实体语义特征 $F_{V\text{sem}}$ 构建的, 拓扑特征层是基于实体和关联边的拓扑特征 $F_{V\text{str}}$, $F_{E\text{str}}$ 构建的, 时序特征层是基于关联边的时序特征 $F_{V\text{time}}$ 构建的 ($F_V = F_{V\text{sem}} \cup F_{V\text{str}}$, $F_E = F_{E\text{str}} \cup F_{V\text{time}}$).

其次, HHFG 中的 3 个层次不是孤立存在的, 不同层次节点之间存在着关联关系, 可清晰表示不同特征之间的相关性. 例如, 在拓扑特征层, 利用映射函数 Ψ_V 可以获取 V 中任一实体的一组语义特征值, 通过分析 $F_{V\text{str}}$ 与 V_{sem} 的包含关系即可得到拓扑特征层中实体节点 V 与语义特征层中语义节点 V_{sem} 的对应关系. 同样, 在拓扑特征层, 利用映射函数 Ψ_E 可以获取到 E 中任一边 e 的边特征, 进而可以找到其在时序特征层所对应的时间序列节点及时序特征 $F_{V\text{time}}$. 这种层次化的结构将有助于根据连边特征的不同而采取不同的转移概率计算策略.

第三, 基于 HHFG 模型既可以表示出丰富的语义, 同时也有助于对网络状态的演化分析. 一方面, 利用 HHFG 模型可以衡量实体之间的语义相似性 (如: 可通过统计实体 V 与语义特征层中 V_{sem} 的交互频率来衡量) 和语义相关性 (如: 通过计算 V_{sem} 在语义特征层中的相似度来衡量). 另一方面, 利用时序特征层可以统计分析链接产生的时间及稳定性, 使预测结果能够较好地适应异构信息网络的动态性.

4.2 直接关联强度量化

若要预测两个实体之间是否存在链路, 需要评估它们之间的关联强度, 即它们间存在链路的概率. 理论上, 在异构信息网络 G 中, 实体 v_i 与 v_j 间的关联强度取决于 v_i 与 v_j 间关联路径的长度、条数, 以及路径上各条边的关联强度. 其中, 关联路径上的每条边都是由直接相连的节点所构成. 我们将这些节点间的关联强度称为直接关联强度 (记为 $S(i,j)$), 它与节点 i, j 的特征 $(\Psi_V(i), \Psi_V(j))$ 以及边特征 $(\Psi_E(ij))$ 均相关 (如式 (2) 所示). 其中, 特征向量 $\Psi = \Psi_V(i) \cup \Psi_V(j) \cup \Psi_E(ij)$, W 是 Ψ 的参数即权重向量. f 是 Sigmoid 函数, 用于将特征向量和其权重的内积映射到 0 到 1 之间.

$$S(i,j) = f(\Psi, W) = \frac{1}{1 + \mathrm{e}^{-\Psi \cdot W}}. \tag{2}$$

然而, 式 (2) 并没有对不同的特征加以区分, 即没有对这些特征进行层次化组织. 在 HHFG 模型中, 构成某关联路径的连边可分为 4 种类型: 实体 – 语义、语义 – 语义、语义 – 实体、实体 – 实体. 相应地, 可进一步将直接关联强度 $S(i,j)$ 分为 $S_{V,V\text{sem}}$, $S_{V\text{sem},V\text{sem}}$, $S_{V\text{sem},V}$ 和 $S_{V,V}$ 4 种类型:

$$\begin{cases} S_{V,V\text{sem}}(i,j) = f(\Psi_V(i) \cup \Psi_{V\text{sem}}(j), W_{\text{I}}), & i \in V \wedge j \in V_{\text{sem}}, \\ S_{V\text{sem},V\text{sem}}(i,j) = f(\Psi_{V\text{sem}}(i) \cup \Psi_{V\text{sem}}(j), W_{\text{II}}), & i \in V_{\text{sem}} \wedge j \in V_{\text{sem}}, \\ S_{V\text{sem},V}(i,j) = f(\Psi_{V\text{sem}}(i) \cup \Psi_V(j), W_{\text{III}}), & i \in V_{\text{sem}} \wedge j \in V, \\ S_{V,V}(i,j) = f(\Psi_{V\text{str}}(i) \cup \Psi_{V\text{str}}(j) \cup \Psi_{E\text{str}}(ij) \cup \Psi_{V\text{time}}(ij), W_{\text{IV}}), & i \in V \wedge j \in V. \end{cases} \tag{3}$$

与式 (2) 中直接关联强度的计算方式不同, 这种细粒度的计算方式将特征按层次加以区分, 并设置不同的权重 ($W_{\text{I}} - W_{\text{IV}}$), 使直接关联强度的计算更具针对性. 这里, $S_{V,V\text{sem}}$ (或 $S_{V\text{sem},V}$), $S_{V\text{sem},V\text{sem}}$

和 $S_{V,V}$ 分别考虑了节点的语义相似性、语义相关性和实体相关性, 其中实体相关性主要体现在拓扑结构相关性和链路稳定性两个方面.

4.3 特征选取

为了能够准确评估节点间的直接关联强度, 需要选取区分能力强、有代表性的特征来构建特征向量. 这些特征往往是特定于某网络的, 本小节以论文合作网络为例, 针对这些特征的选取进行讨论.

(1) 针对 $S_{V,V\mathrm{sem}}$ (或 $S_{V\mathrm{sem},V}$) 的特征选取. $S_{V,V\mathrm{sem}}$ 用来衡量实体节点与语义节点间的语义相似性, 若要评估某作者 (v_i) 与某主题 (V_{sem}) 的关联强度, 可考虑如下语义特征: v_i 所发表的关于主题 V_{sem} 的论文数、v_i 的合作者中发表了关于主题 V_{sem} 的人数、主题 V_{sem} 的稀有程度等. 若 v_i 所发表的关于主题 V_{sem} 的论文数越多, 发表了 V_{sem} 且为 v_i 的合作者越多, V_{sem} 越稀有, 将对 $S_{V,V\mathrm{sem}}$ (或 $S_{V\mathrm{sem},V}$) 具有促进作用.

(2) 针对 $S_{V\mathrm{sem},V\mathrm{sem}}$ 的特征选取. $S_{V\mathrm{sem},V\mathrm{sem}}$ 用来衡量节点的语义相关性, 这里仍以论文合作网络为例, 若要评估两个主题 (V_{sem1} 与 V_{sem2}) 的关联强度, 可考虑如下语义特征: 与 V_{sem1} (或 V_{sem2}) 在论文中共现的其他主题数、V_{sem1} 与 V_{sem2} 共现的论文数、V_{sem1} 与 V_{sem2} 在语义特征层中的共同邻居数等. 这里, 若与 V_{sem1} (或 V_{sem2}) 在论文中共现的主题数越多, 则表明 V_{sem1} (或 V_{sem2}) 越平凡, 将对 $S_{V\mathrm{sem},V\mathrm{sem}}$ 具有削弱作用. 若 V_{sem1} 与 V_{sem2} 经常在一篇论文中同时出现, 或者有许多共同的邻居, 则表明 V_{sem1} 与 V_{sem2} 越相关, 将对 $S_{V\mathrm{sem},V\mathrm{sem}}$ 具有促进作用.

(3) 针对 $S_{V,V}$ 的特征选取. $S_{V,V}$ 用来衡量实体相关性, 具体体现在拓扑结构相关性和链路稳定性两个方面. 针对拓扑结构相关性的评估, 目前已有很多相关工作可以借鉴. 例如: 若要评估两个作者实体 (v_1 与 v_2) 的拓扑结构相关性, 可考虑 v_1 (或 v_2) 发表的论文数、v_1 与 v_2 的共同合作者数等特征. 本文考虑了异构信息网络的动态性, 链路的存在与否是动态变化的. 例如, 若两个作者在之前合作过, 并不代表他们以后会一直合作下去. 为此, 需要考虑链路的时序特征, 即链路稳定性. 针对 v_1 与 v_2 的链路稳定性评估, 本文考虑了链路产生的时间和稳定因子两个特征. 下面, 针对稳定因子的计算进行介绍.

假定初始时刻为 t_1, 当前时刻为 t, 实体 v_1 与 v_2 间的链路对应于 G_{time} 中的时序节点 v_i ($v_i \in V_{\mathrm{time}}$). 首先将 G_{time} 划分为 N 个时间窗口, 形成 N 个时序特征图 $G_1,\ldots,G_N$, 每个窗口的时长为 $(t-t_1)/N$. 然后针对每两个相邻的窗口, 根据 G_{time} 的变化计算时序节点 v_i 的影响分值 $Af(v_i;j)$ (v_i 在第 j 个窗口的影响分值).

$$Af(v_i,j)=\begin{cases}1, & v_i\in G_j\wedge v_i\notin G_{j-1},\\ -1, & v_i\notin G_j\wedge v_i\in G_{j-1},\\ 0, & \text{otherwise.}\end{cases}\tag{4}$$

若时序节点 v_i 存在于时间窗口 j 中, 而在其前一个窗口 $j-1$ 中未存在, 则表明实体 v_1 与 v_2 间的链路是新生成的, 可推测该链路在下一窗口消亡的概率较小, 因此 G_{time} 的变化对链路的形成具有促进作用. 若时序节点 v_i 存在于时间窗口 $j-1$ 中, 而在其后的窗口 j 中消亡, 则可推测实体 v_1 与 v_2 间的链路极有可能在下一窗口仍然处于消亡状态, 因此 G_{time} 的变化对链路的形成具有削减作用. 若时序节点 v_i 在时间窗口 $j-1$ 和 j 中的状态一致, 则可推测实体 v_1 与 v_2 间链路的状态极有可能继续保持, 此时 G_{time} 对链路的形成不具备影响. 例如, 通过分析 G_{time} 可以获取作者间合作关系的时序信息. 若两个作者在时间窗口 $j-1$ 中未合作而在时间窗口 j 中有所合作, 或者他们在时间窗口 $j-1$ 中合作而在时间窗口 j 中未合作, 都表明时序特征对链路的有无具有较大的影响作用. 前者表示时序特

征对链路的存在性具有促进作用, 即该合作关系在下一时间窗口极有可能被继续保持; 而后者表示时序特征对链路的存在性具有消极作用, 即该合作关系在下一时间窗口极有可能继续处于消亡状态.

除了时序节点 v_i 自身的影响分值对实体 v_1 与 v_2 间链路的稳定性有影响, v_i 的邻居节点的影响分值也会影响到链路的稳定性. 这是因为, 若两个实体与它们的共同邻居交互紧密, 间接反映了它们的交互也紧密. 假定实体 v_1 与 v_2 在拓扑特征层中的共同邻居是实体 v_3, v_1 与 v_3, v_2 与 v_3 间连边分别对应于时序特征层的时序节点 v_p 和 v_q, 则时序节点 v_i 的稳定因子如下所示:

$$St(v_i) = \sum_{j=2}^{N} \mathrm{e}^{-\beta(t-t_j)} Af(v_i, j) + \tau \sum_{p,q} Af(v_p, j) + Af(v_q, j), \tag{5}$$

其中 t 代表待预测时间窗口, t_j 代表时间窗口 j, β 代表时间惩罚因子 ($\beta \in [0, 1]$). 若 t_j 距离当前时刻越久远, 表明其产生的影响作用越微弱. 这里, 我们既考虑了 v_i 自身的稳定性, 也考虑了其邻居的稳定性, τ 为 v_i 的邻域影响系数.

特征选取后, 即可计算节点间的直接关联强度. 首先构建特征向量, 并通过训练, 学习这些特征的权重 (详见 5.3 小节). 然后, 计算特征向量与权重向量的内积, 并利用式 (3) 将其映射为 0 到 1 之间的数, 即节点间的直接关联强度.

5 基于 HHFG 的链路预测算法

基于 HHFG 模型, 本节提出一种链路预测算法. 首先介绍算法的设计思想, 然后针对算法中的随机游走概率计算和参数学习两个核心部分进行详细介绍.

5.1 算法设计思想

基于 HHFG 的链路预测算法的基本思想如图 2 所示. 该算法包括训练和预测两个阶段:

训练阶段. 首先, 将训练集作为输入, 并构建 HHFG 模型. 然后, 为 HHFG 所表达的特征定义初始权重, 将其作为预测函数的初始参数值, 并计算 HHFG 中节点间的随机游走概率. 最后, 基于梯度下降的思想, 在渐变下降中使用微积分迭代调整参数值, 使它们最小化给定的损失函数.

预测阶段. 首先, 将待预测数据集作为输入, 并构建 HHFG 模型. 然后, 基于训练出的参数计算 HHFG 中节点间的随机游走概率, 以此来预测某时段节点间链路存在的可能性. 最后, 针对每个源节点, 按照概率值由高到低的顺序, 截取前 top-k 个目标节点作为预测结果.

基于 HHFG 的链路预测算法包括如下步骤 (如算法 1 所示).

步骤 1. 数据集划分 (第 1 行). 假定初始时刻为 t_1, 当前时刻为 t, 为了训练模型, 需要将整个时段 $[t_1, t]$ 划分为两个片段 ($T_F = [t_1, t']$ 和 $T_L = [t', t]$). 相应地, 图 G 被划分为 G_F 和 G_L 两部分: G_F 用于特征提取; G_L 用于形成正例和反例的标签, 通过最大化 G_L 中新形成链路的似然来学习参数.

步骤 2. 基于 G_F 构建 HHFG 模型并抽取特征 (第 2, 3 行). 将 G_F 的拓扑特征、语义特征与时序特征以层次化的 HHFG 模型组织, 并抽取特征 $F' = F_{V\mathrm{sem}} \cup F_{V\mathrm{str}} \cup F_{E\mathrm{str}} \cup F_{V\mathrm{time}}$.

步骤 3. 获取训练集 (第 4 行). 在 G_F 上选择与源节点集 S 中节点具有相似特征的节点作为训练集, 记为 S'. 这是因为, S' 具有与 S 相似的特征, 训练出的参数也更能符合 S 的特点, 以确保预测的准确性.

步骤 4. 初始化参数并对其学习 (第 5~11 行). 首先, 设置参数初始值 (本算法将 0 作为初始值). 然后, 针对 S' 中的每个节点 s', 基于特征 F' 和初始参数计算 s' 到其他节点的随机游走概率 (详见 5.2 小节). 接下来, 使用梯度下降方法迭代地调整参数, 直到收敛 (详见 5.3 小节).

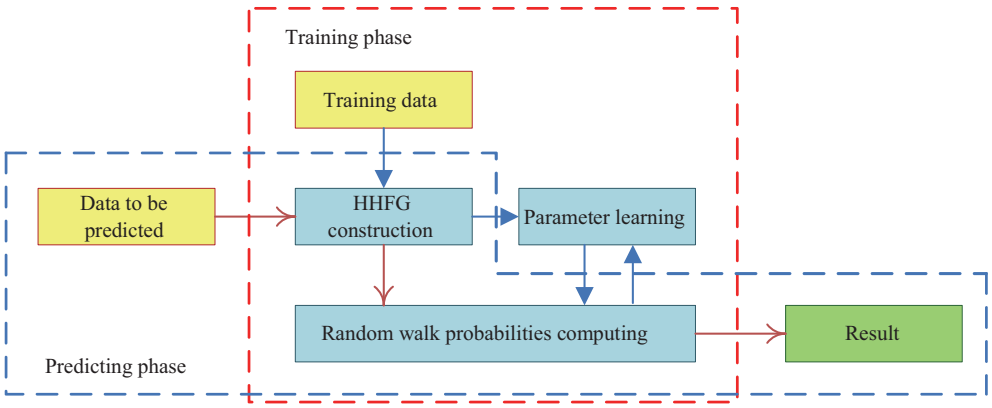

图 2 (网络版彩图) 基于 **HHFG** 的链路预测算法的基本思想

Figure 2 (Color online) The basic idea of HHFG-based link prediciton algorithmn

步骤 5. 基于 G 构建 HHFG 模型并抽取特征 $F = F_{V\text{sem}} \cup F_{V\text{str}} \cup F_{E\text{str}} \cup F_{V\text{time}}$ (第 12, 13 行).

步骤 6. 生成预测结果 (第 14~18 行). 首先, 针对待预测源节点集 S 中的每个节点 s, 基于特征 F 和训练后的参数计算 s 到其他节点的随机游走概率. 然后, 根据概率值选取前 top-k 个节点作为 s 的预测结果. 最终, 将 S 中各个节点的预测结果汇总并返回.

Algorithm 1 HHFG-based link prediciton algorithmn

Input: Heterogeneous information network G, source node set S, integer k;

Output: $\{N_1, \ldots, N_{|S|} | N_i \subseteq V \wedge |N_i| = k \wedge 1 \leqslant i \leqslant |S|\}$;

1: Divide G into G_F and G_L;
2: Construct HHFG (denoted as $G'_{\text{sem}} \cup G'_{\text{str}} \cup G'_{\text{time}}$) based on G_F;
3: Extract features F' from $G'_{\text{sem}} \cup G'_{\text{str}} \cup G'_{\text{time}}$;
4: $S' = \text{findSimNodes}(G_F, S)$;
5: Initialize W_{I}, W_{II}, W_{III} and W_{IV};
6: **repeat**
7: **for** $s' \in S'$ **do**
8: $P = \text{compRWProb}(F', W_{\text{I}}, W_{\text{II}}, W_{\text{III}}, W_{\text{IV}})$;
9: **end for**
10: Update W_{I}, W_{II}, W_{III} and W_{IV} based on G_L;
11: **until** Convergence
12: Construct HHFG (denoted as $G_{\text{sem}} \cup G_{\text{str}} \cup G_{\text{time}}$) based on G;
13: Extract features F from $G_{\text{sem}} \cup G_{\text{str}} \cup G_{\text{time}}$;
14: **for** $s \in S$ **do**
15: $P = \text{compRWProb}(F, W_{\text{I}}, W_{\text{II}}, W_{\text{III}}, W_{\text{IV}})$;
16: $N_i = \text{sort}(P, k)$;
17: **end for**
18: Return $\{N_1, \ldots, N_{|S|}\}$.

5.2 随机游走概率计算

在 HHFG 模型中, 每条关联路径都是由一组连边构成的. 不同类型连边的重要性是有差别的, 为此本文按照连边两侧端点的不同, 将连边划分为实体 – 语义、语义 – 语义、语义 – 实体、实体 – 实体 4 种类型, 并分别提出了转移概率计算策略.

(1) 以实体为起点的转移概率计算. 以实体为起点的连边, 终点可能是实体类型的节点 V, 也可能是语义特征类型的节点 V_{sem}. 为了加以区别, 我们在计算转移概率时设置了实体转移系数 μ ($\mu \in [0,1]$), 表示转移到 V 和 V_{sem} 的概率分别是 μ 和 $1-\mu$, 该系数用来区分拓扑特征和语义特征的重要性. 实体 – 实体, 实体 – 语义的转移概率计算分别如式 (6) 和 (7) 所示. 其中, $\Gamma_V(i)$ 和 $\Gamma_{V\text{sem}}(i)$ 分别表示节点 i 的邻域中的实体节点集合和语义节点集合.

$$Q_{V,V}(i,j)=\begin{cases}\mu\dfrac{S_{V,V}(i,j)}{\sum_m S_{V,V}(i,m)+\sum_n S_{V,V\text{sem}}(i,n)}, & |\Gamma_V(i)>0|\wedge|\Gamma_{V\text{sem}}(i)>0|,\\ \dfrac{S_{V,V}(i,j)}{\sum_m S_{V,V}(i,m)}, & |\Gamma_V(i)>0|\wedge|\Gamma_{V\text{sem}}(i)=0|,\\ 0, & \text{otherwise.}\end{cases}\tag{6}$$

$$Q_{V,V\text{sem}}(i,j)=\begin{cases}(1-\mu)\dfrac{S_{V,V\text{sem}}(i,j)}{\sum_m S_{V,V}(i,m)+\sum_n S_{V,V\text{sem}}(i,n)}, & |\Gamma_V(i)>0|\wedge|\Gamma_{V\text{sem}}(i)>0|,\\ \dfrac{S_{V,V\text{sem}}(i,j)}{\sum_n S_{V,V\text{sem}}(i,n)}, & |\Gamma_{V\text{sem}}(i)>0|\wedge|\Gamma_V(i)=0|,\\ 0, & \text{otherwise.}\end{cases}\tag{7}$$

(2) 以语义特征为起点的转移概率计算. 以语义特征为起点的连边, 终点可能是实体类型的节点 V, 也可能是语义节点 V_{sem}. 同样, 通过设置语义转移系数 η ($\eta \in [0,1]$) 来区分语义相似性和语义相关性的重要性, 由语义节点转移到 V 和 V_{sem} 的概率分别是 η 和 $1-\eta$. 语义 – 语义、语义 – 实体的转移概率计算分别如式 (8) 和 (9) 所示:

$$Q_{V\text{sem},V\text{sem}}(i,j)=\begin{cases}\eta\dfrac{S_{V\text{sem},V\text{sem}}(i,j)}{\sum_m S_{V\text{sem},V}(i,m)+\sum_n S_{V\text{sem},V\text{sem}}(i,n)}, & |\Gamma_V(i)>0|\wedge|\Gamma_{V\text{sem}}(i)>0|,\\ \dfrac{S_{V\text{sem},V\text{sem}}(i,j)}{\sum_n S_{V\text{sem},V\text{sem}}(i,n)}, & |\Gamma_V(i)=0|\wedge|\Gamma_{V\text{sem}}(i)>0|,\\ 0, & \text{otherwise.}\end{cases}\tag{8}$$

$$Q_{V\text{sem},V}(i,j)=\begin{cases}(1-\eta)\dfrac{S_{V\text{sem},V}(i,j)}{\sum_m S_{V\text{sem},V}(i,m)+\sum_n S_{V\text{sem},V\text{sem}}(i,n)}, & |\Gamma_V(i)>0|\wedge|\Gamma_{V\text{sem}}(i)>0|,\\ \dfrac{S_{V\text{sem},V}(i,j)}{\sum_m S_{V\text{sem},V}(i,m)}, & |\Gamma_{V\text{sem}}(i)=0|\wedge|\Gamma_V(i)>0|,\\ 0, & \text{otherwise.}\end{cases}\tag{9}$$

(3) 随机游走概率计算. 若要预测两个实体之间是否存在链路, 需要评估它们之间的关联强度. 转移概率用来衡量直接由边相连的节点之间的关联强度即直接关联强度. 若两个节点未直接相连, 则可以通过连接它们的路径来衡量它们间的关联强度, 它与关联路径的条数、长度及路径上的直接关联强度均相关. 本文针对源节点集 S 中的每个源节点, 采用带重启的随机游走策略来估计它与 V_{sem}, V 中各个节点的关联强度, 即随机游走概率. 最终将随机游走概率作为节点间产生链路的概率值. 本文基于带重启随机游走模型来计算随机游走概率, 该模型的收敛性已在相关文献 [28] 中得以证明. 假设从某个源节点出发, 采用向量 P 表示它与 V_{sem}, V 中每个节点的关联强度, 利用随机游走的过程可以使向量 P 收敛到向量 P^*. 如式 (10) 所示, 第 n 次迭代后所有节点的向量记为 P^n, 其中 $Q=Q_{V,V}\cup Q_{V,V\text{sem}}\cup Q_{V\text{sem},V\text{sem}}\cup Q_{V\text{sem},V}$ 是转移概率矩阵, P^0 是初始向量 (例如: 可将各节点度的倒数作为其初始值).

$$P^n=(1-\alpha)\times Q\times P^{n-1}+\alpha\times P^0.\tag{10}$$

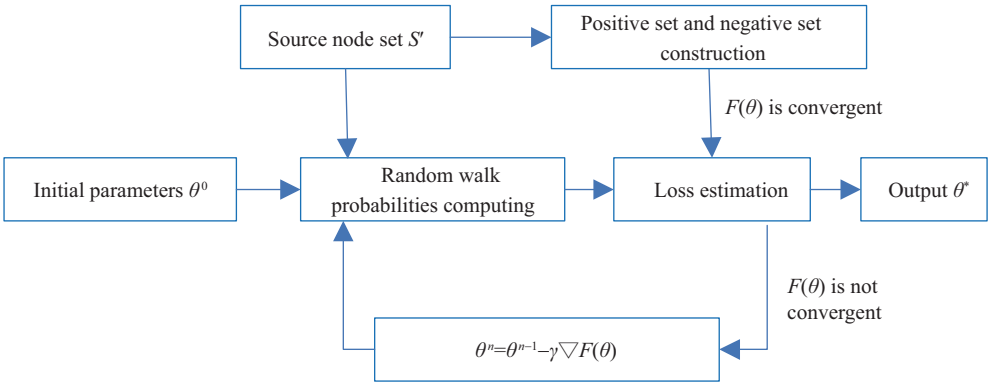

图 3 (网络版彩图) 基于梯度下降的参数更新策略

Figure 3 (Color online) The update strategy of parameters based on gradient descent

5.3 参数学习

前文在计算直接关联强度及转移概率时, 使用的参数主要包括特征权重 $(W_{\mathrm{I}}-W_{\mathrm{IV}})$、实体转移系数 μ 及语义转移系数 η 等, 这些参数构成了参数向量 θ. 为了学习这些参数, 本文基于梯度下降思想来迭代调整 θ 取值, 最终得到 θ^*, 使损失函数 $F(\theta)$ 最小化 (如式 (11)). 首先, 针对训练集 S' 中的每个源节点 s', 构建出正例集 $D_{s'}$ 和反例集 $L_{s'}$. 然后, 计算 s' 到其他节点的随机游走概率, 并迭代地对 θ 进行调整, 尽可能保证 $D_{s'}$ 中节点的随机游走概率 p_d 大于 $L_{s'}$ 中节点的随机游走概率 p_l. 其中, h 是一个非负函数 (如式 (12)), 当 p_l-p_d 小于 0 时, $h(p_l-p_d)$ 为 0; 当 p_l-p_d 大于等于 0 时, $h(p_l-p_d)$ 大于 0.

$$\theta^*=\mathrm{argmin}_\theta F(\theta)=\mathrm{argmin}_\theta\left(\|\theta\|^2+\lambda\sum_{s'\in S'}\sum_{d\in Ds',l\in Ls'}h(p_l-p_d)\right), \tag{11}$$

$$h(x)=\frac{1}{1+\mathrm{e}^{-x/b}}, \tag{12}$$

$$\theta^n=\theta^{n-1}-\gamma\nabla F(\theta). \tag{13}$$

为了求得满足损失函数的解, 本文使用梯度下降算法更新 θ, 直到 $F(\theta)$ 收敛. 更新策略如图 3 所示: 首先, 随机选取 θ 的初始值 θ^0; 然后分别计算初始点处各个参数对 F 的偏微分 $\nabla F(\theta)$ (即梯度), 并将 θ^0 减去速率因子 γ 乘以偏微分的值 (如式 (13), θ^n 表示第 n 次迭代后的参数向量); 对 θ 进行多次迭代更新, 最终得到 θ^*, 使 $F(\theta)$ 收敛并达到最小.

6 实验测试

6.1 数据集

本文从 arXiv e-print archive[1)] 中抽取了部分文献的信息作为数据集, 该数据集包含了在 1992 年至 1998 年期间所发表的关于高能物理理论的论文信息和作者信息. 这些论文和作者构成了一个异构信息网络, 本文的目标是预测作者间的合作关系.

1) https://arxiv.org/.

表 1 实验数据集

Table 1 Dataset

Level	Number of nodes	Number of edges
Semantic feature level	2867	98677
Structure feature level	9682	15864
Time feature level	15864	39656

对于数据集的划分, 采用如下策略. (1) 将 1992 年至 1997 年期间所发表的论文信息作为训练数据集. 为了训练模型, 又将训练数据集划分为两部分 ($T_F = [1992, 1995]$ 和 $T_L = [1996, 1997]$), 分别用于特征提取和参数学习. (2) 将 1998 年至 1999 年期间所发表的论文信息作为测试数据集.

在构建 HHMG 模型时, 通过抽取论文的主题词作为语义特征, 通过分析作者的合作关系来获取拓扑特征, 将论文发表的时间作为时序特征. HHMG 模型的数据规模如表 1 所示, 语义特征层包含的节点数和边数分别为 2867, 98677; 拓扑特征层包含的节点数和边数分别为 9682, 15864; 时序特征层包含的节点数和边数分别为 15864, 39656.

为了充分利用网络所表达的各种特征, 应选取特征较鲜明的节点作为源节点. 为此, 我们选取了发表论文数大于 2 篇以及合作者数大于 5 的作者作为待预测数据集, 即源节点集 S. 本文选取了 206 个源节点, 将候选集中最终与源节点形成链路的节点集作为正例集, 将没有与源节点形成链路的节点集作为反例集. 其正例集的平均基数 D_{avg} 和反例集的平均基数 C_{avg} 分别为 0.496 和 11.01, $D_{\text{avg}}/C_{\text{avg}}$ 为 0.045.

6.2 评价指标

本文使用 Precision (式 (14)), Recall (式 (15)), AUC [29] (式 (16)), ROC 曲线 [30], P@K 作为评价指标对实验结果进行度量.

Precision 等于预测结果中真实的正例占整个预测结果的比例, 用于衡量预测结果的精确性. Recall 等于预测结果中真实的正例占整个真实正例的比例, 用于衡量预测结果是否完备.

除了 Precision 和 Recall 以外, 本文还使用 ROC 曲线和 AUC 值作为评价指标来克服网络的类不平衡性. ROC 曲线的横纵坐标分别是假正例率和真正例率, 将阈值 k 从最大的链路分数开始下降, 每下降一步可以得到新的真正例率和假正例率的值, 作为 ROC 曲线上的一点. AUC 可以理解为从正例 (即实际存在但缺失的链路集) 中随机选择一条链路的分数比从反例 (即实际不存在的链路集) 中随机选择一条链路的分数高的概率. 为了简化操作, 我们借鉴了文献 [29] 中计算 AUC 的方法 (如式 (16)), 其中 n 为比较次数, n' 表示正例中所选边的分数大于反例中所选边分数的次数, n'' 表示二者相等的次数. n' 和 n'' 越大, 说明正例分数高于反例分数的概率越大, 预测准确性越高.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{14}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{15}$$

$$\text{AUC} = \frac{n' + 0.5n''}{n}. \tag{16}$$

6.3 参数学习测试

首先, 我们评估了参数的迭代更新对链路预测准确性的影响. 本文提出了基于梯度下降的参数学

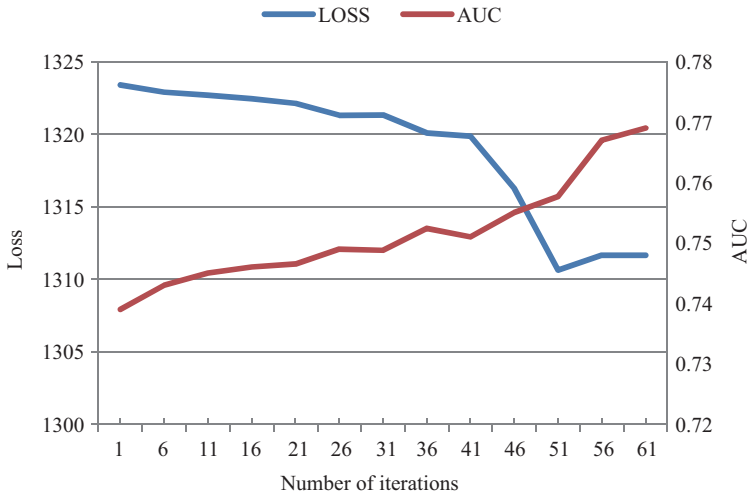

图 4 (网络版彩图) 参数优化过程中损失值和 AUC 值的变化

Figure 4 (Color online) The variety of Loss and AUC during parameter optimization

表 2 参数设置

Table 2 Parameters setting

F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	μ	η
0.0143	0.0115	−0.0258	−0.0371	0.0083	0.0236	−0.003	0.0601	0.0251	0.054	0.255	0.211

习算法, 图 4 表示随着参数的迭代更新, LOSS 值 (损失值) 和 AUC 值的变化情况. 从图 4 可以看出, 随着迭代次数的增加, 总体上 AUC 值保持上升趋势, LOSS 值保持下降趋势. 这表明, 随着迭代的进行, 调整后的参数使得预测精度在不断提高.

其次, 我们针对特征的选取及特征权重、实体转移系数 μ、语义转移系数 η 等参数的设置进行了测试. 在实验中, 针对 $S_{V,V\text{sem}}$ (或 $S_{V\text{sem},V}$) 的计算, 选取的特征包括: 作者所发表的关于某主题的论文数 (记为 F_1); 主题的稀有程度 (记为 F_2). 针对 $S_{V\text{sem},V\text{sem}}$ 的计算, 选取的特征包括: 与主题 (连边源点) 在论文中共现的主题数 (记为 F_3); 与主题 (连边终点) 在论文中共现的主题数 (记为 F_4); 两个主题共现的论文数 (记为 F_5); 两个主题在语义特征层中的共同邻居数 (记为 F_6). 针对 $S_{V,V}$ 的计算, 选取的特征包括: 作者 (连边源点) 发表的论文数 (记为 F_7); 作者 (连边终点) 发表的论文数 (记为 F_8); 链路产生的时间 (记为 F_9); 稳定因子 (记为 F_{10}).

经过训练后, 得到的最佳参数如表 2 所示. 其中, 权值为正的特征将对链路的生成具有促进作用, 而权值为负的特征将具有削弱作用. 例如, 特征 F_5 对链路的生成具有促进作用, 这是因为两个主题共现的论文数越多, 说明它们在语义上越相关. 而特征 F_4 对链路的生成起到了削弱作用, 这是因为在论文中频繁出现的主题, 其区分能力较差.

6.4 有效性验证

首先, 从特征选取的角度, 我们比较了基于不同特征的链路预测方法的准确性. 具体包括:

方法 1. 基于共同邻居的链路预测方法 (CN) [2];

方法 2. 基于 Jaccard 的链路预测方法 (Jaccard) [3];

方法 3. 基于随机游走的链路预测方法 (RW) [7];

表 3 基于不同特征的链路预测方法的 Precision, Recall 和 AUC 比较

Table 3 Comparision of link prediction methods based on different features with Precision, Recall and AUC

Method	Precision	Recall	AUC
CN	0.103	0.099	0.608
Jaccard	0.105	0.099	0.652
RW	0.106	0.105	0.741
DW	0.115	0.112	0.746
SS	0.123	0.123	0.753
GT	0.125	0.124	0.759
SF	0.129	0.124	0.767
HHFG	0.131	0.124	0.769

方法 4. 基于 DeepWalk 的链路预测方法 (DW) [31];

方法 5. 语义特征与拓扑特征相结合的链路预测方法 (SS);

方法 6. 基于生成时间的链路预测方法 (GT);

方法 7. 基于稳定因子的链路预测方法 (SF);

方法 8. 本文提出的基于带邻域影响系数的稳定因子的链路预测方法 (HHFG).

其中, 方法 1~4 是传统的链路预测方法, 方法 5~8 是对本文所提出的链路预测方法的不同实现方式. 针对性能指标 Precision, Recall 和 AUC 的测试结果如表 3 所示. 方法 1~4 分别基于共同邻居、Jaccard 系数、随机游走概率和向量距离来评估节点间产生链路的概率. 其中, 方法 4 是一种典型的基于网络表示学习的链路预测方法, 其思想是利用随机游走序列表示各个节点的近邻并作为语言模型的输入, 以获取节点的向量表示, 最终通过计算向量间距离来进行链接预测. 由于方法 4 在随机游走的基础上采用了网络表示学习技术, 可有效缓解信息网络的稀疏性问题, 因此略优于方法 3. 然而, 方法 1~4 在预测时仅考虑了网络的拓扑特征, 而忽略了节点之间的语义相似性, 语义相关性, 因此 Precision, Recall 和 AUC 均较低. 方法 5 在此基础上考虑了主题的共现程度, 稀有程度等语义特征, 同方法 1~4 比, Precision, Recall 和 AUC 均有所提高. 方法 6~8 在预测时考虑了拓扑特征、语义特征和时序特征. 方法 6 在计算直接关联强度时仅将链路的生成时间作为时序特征. 方法 7, 8 在方法 6 的基础上, 考虑了时序节点的稳定因子, 不同在于: 方法 7 在计算稳定因子时未考虑邻域系数, 而方法 8 既考虑了当前时序节点的稳定性, 也考虑了其邻居节点的稳定性, 更能有效地评估关联强度.

这些算法的 ROC 曲线如图 5 所示. 方法 1~7 侧重于考虑拓扑特征、语义特征或时序特征, 由于考虑的因素有限, 很难保证预测的准确性. 方法 8 充分考虑了实体的拓扑特征、语义特征与时序特征, 将这些特征以 HHFG 模型进行有效组织, 通过融合这些特征来计算随机游走概率, 以此来综合评估节点之间产生链接的可能性. 可以看出, 本文提出的 HHFG 算法优于其他算法.

此外, 本文针对性能指标 P@K 进行了测试 (如表 4 所示). 在链路预测过程中, 本文分别为每个源节点截取分数排在前 K 的链路作为预测结果, P@K 用来衡量所有源节点的平均预测准确率. 从实验结果可知, K 的选取对预测准确率的影响较大. 目前, 本文仅通过实验评估了 K 值对预测准确率的影响, 在将来的工作中, 本文将针对 K 值的设置进行理论分析, 进一步提高预测性能.

其次, 从特征区分的角度, 我们比较了不同的特征划分策略及转移概率计算策略对链路预测的影响. 具体包括:

方法 1. 基于传统异构信息网络的链路预测方法 (HIN);

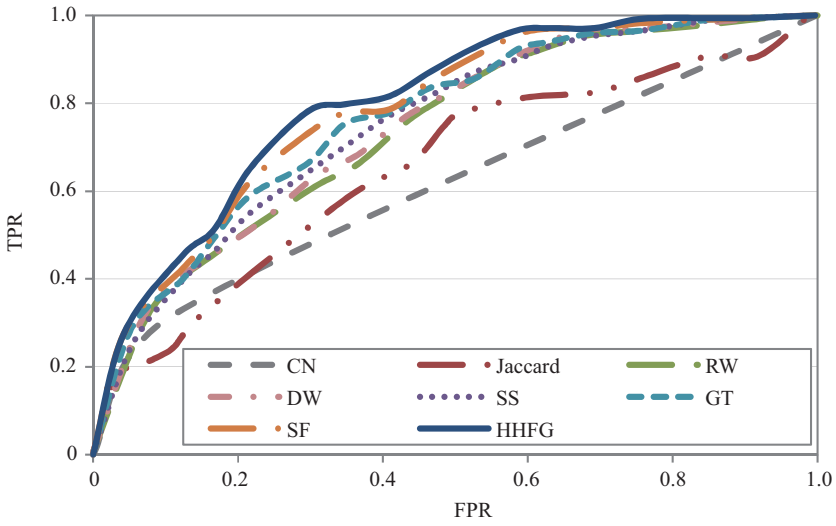

图 5 (网络版彩图) 基于不同特征的链路预测方法的 ROC 曲线

Figure 5 (Color online) Comparision of link prediction methods based on different features with ROC curve

表 4 基于不同特征的链路预测方法的 P@K 比较

Table 4 Comparision of link prediction methods based on different features with P@K

Method	P@1	P@5	P@10	P@20	P@50
CN	0.142	0.131	0.102	0.099	0.043
Jaccard	0.147	0.128	0.112	0.099	0.037
RW	0.152	0.128	0.121	0.099	0.029
DW	0.163	0.142	0.135	0.099	0.037
SS	0.175	0.161	0.129	0.108	0.043
GT	0.179	0.171	0.131	0.101	0.043
SF	0.183	0.176	0.142	0.106	0.037
HHFG	0.191	0.177	0.139	0.106	0.043

方法 2. 基于层次化混合特征图 (不考虑转移系数) 的链路预测方法 (NTC);

方法 3. 本文提出的基于层次化混合特征图 (考虑转移系数) 的链路预测方法 (HHFG).

实验结果如表 5 所示, 其中, HIN (heterogeneous information network) 方法是在传统的异构信息网络上进行链路预测, 虽然也考虑了拓扑特征、语义特征和时序特征, 但将它们同等对待 (即不考虑权重), 影响了预测的准确性. 方法 2, 3 基于梯度下降方法对这些特征的权重进行了学习, 能够有效地对它们加以区分. 对于方法 2, 当节点在层内及层间进行随机游走时, 不区分边的类型. 方法 3 在方法 2 的基础上, 利用实体转移系数和语义转移系数对边的特征加以区分, 提高了链路预测的准确性. 这些算法的 ROC 曲线如图 6 所示, 可见本文提出的算法 (方法 3) 具有一定优势.

7 总结

针对现有链路预测技术的不足, 本文针对异构信息网络上的链路预测技术进行了研究. 为了更加准确地评估实体间产生链接的概率, 本文充分考虑了异构信息网络的拓扑特征、语义特征和时序特征,

表 5　不同的特征划分策略及转移概率计算策略对链路预测的 Precision, Recall 和 AUC 影响

Table 5　Comparision of link prediction methods based on different features division and transition probability computing strategies with Precision, Recall and AUC

Method	Precision	Recall	AUC
HIN	0.124	0.124	0.747
NTC	0.126	0.124	0.752
HHFG	0.131	0.124	0.769

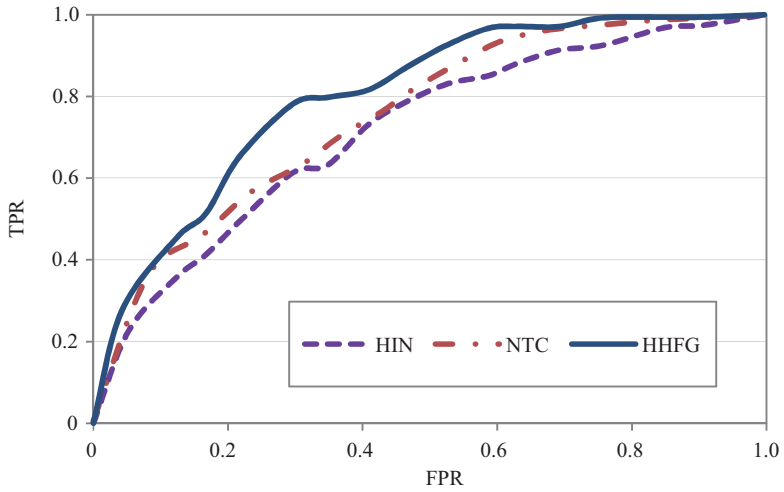

图 6　(网络版彩图) 不同的特征划分策略及转移概率计算策略对链路预测的 ROC 曲线影响

Figure 6　(Color online) Comparision of link prediction methods based on different features division and transition probability computing strategies with ROC curve

提出了一种层次化混合特征图模型 HHFG. 该模型利用实体特征和边特征来区分不同类型的实体及关联关系, 并将异构信息网络的拓扑特征、语义特征和时序特征进行层次化组织, 充分地利用了异构信息网络所表达的丰富特征. 此外, 本文提出了一种基于 HHFG 的链路预测算法, 通过计算随机游走概率来评估节点之间产生链接的可能性, 并采用梯度下降法学习参数, 有效地保证了链路预测的准确性.

随着网络规模的增加, 链路预测算法的计算效率及准确率都将会受到很大挑战. 为此, 在下一步的工作中, 我们将力求从两方面来提高算法的可扩展性. 首先借鉴网络表示学习技术, 综合 HHFG 模型所表达的语义特征、拓扑特征及时序特征, 将各层次节点映射为低维度向量, 以此提升节点关联强度的计算效率. 其次, 按照特征的不同, 将异构信息网络划分为不同子空间, 并采用并行计算模型来协同计算各个子空间, 以此提高链路预测算法的并行性. 此外, 还将针对特征选择策略及 K 值设置进行研究.

参考文献

1　Zhou T, Lü L, Zhang Y C. Predicting missing links via local information. Eur Phys J B, 2009, 71: 623–630
2　Mitzenmacher M. A brief history of generative models for power law and lognormal distributions. Internet Math, 2004, 1: 226–251
3　Salton G, McGill M J. Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1986
4　Li X Y, Du N, Li H. A deep learning approach to link prediction in dynamic networks. In: Proceedings of SIAM International Conference on Data Mining, 2014. 289–297

5 Wang D X, Cui P, Zhu W W. Structural deep network embedding. In: Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2016. 1225–1234

6 Lü L, Jin C H, Zhou T. Similarity index based on local paths for link prediction of complex networks. Phys Rev E, 2009, 80: 046122

7 Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst, 1998, 30: 107–117

8 Jeh G, Widom J. SimRank: a measure of structural-context similarity. In: Proceedings of the 8th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2002. 538–543

9 Leskovec J, Horvitz E. Planetary-scale views on a large instant-messaging network. In: Proceedings of the 17th International Conference on World Wide Web, 2008. 915–924

10 Akcora C G, Carminati B, Ferrari E. User similarities on social networks. Soc Netw Anal Min, 2013, 3: 475–495

11 Aiello L M, Barrat A, Schifanella R, et al. Friendship prediction and homophily in social media. ACM Trans Web, 2012, 6: 1–33

12 Huang Z, Lin D K J. The time-series link prediction problem with applications in communication surveillance. Inform J Comput, 2009, 21: 286–303

13 Tylenda T, Angelova R, Bedathur S. Towards time-aware link prediction in evolving social networks. In: Proceedings of the 3rd Workshop on Social Network Mining and Analysis, 2009

14 Sun Y, Han J, Aggarwal C C, et al. When will it happen? — relationship prediction in heterogeneous information networks. In: Proceedings of the 5th ACM International Conference on Web Search and Data Mining, 2012. 663–672

15 Lee C, Nick B, Brandes U, et al. Link prediction with social vector clocks. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013. 784–792

16 Liu B Q, Liu F, Wang X L, et al. Link value and event-result prediction for sequence behavior in social networks. Sci Sin Inform, 2015, 45: 1558–1573 [刘秉权, 刘峰, 王晓龙, 等. 社会网络中序列行为的链接值及事件结果预测. 中国科学: 信息科学, 2015, 45: 1558–1573]

17 Bao Z F, Zeng Y, Tay Y C. SonLP: social network link prediction by principal component regression. In: Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013. 364–371

18 Sun Y Z, Han J W, Yan X F, et al. PathSim: meta path-based top-K similarity search in heterogeneous information networks. In: Proceedings of the VLDB Endowment, 2011. 992–1003

19 Sun Y Z, Barber R, Gupta M, et al. Co-author relationship prediction in heterogeneous bibliographic networks. In: Proceedings of International Conference on Advances in Social Networks Analysis and Mining, 2011. 121–128

20 Huang L W, Li D Y, Ma Y T. A meta path-based link prediction model for heterogeneous information networks. Chinese J Comput, 2014, 37: 848–858 [黄立威, 李德毅, 马于涛. 一种基于元路径的异质信息网络链路预测模型. 计算机学报, 2014, 37: 848–858]

21 Chen N, Zhu J, Xia F, et al. Discriminative relational topic models. IEEE Trans Pattern Anal Mach Intell, 2015, 37: 973–986

22 Wang H, Shi X J, Yeung D Y. Relational deep learning: a deep latent variable model for link prediction. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, 2017. 2688–2694

23 Hao Z G, Zhang W X, Chen Z. Link prediction in online social networks based on supervised joint denoising model. Sci Sin Inform, 2017, 47: 1551–1565 [郝占刚, 章伟雄, 陈政. 基于监督联合去噪模型的社交网络链接预测. 中国科学: 信息科学, 2017, 47: 1551–1565]

24 Chen H X, Yin H Z, Wang W Q, et al. PME: projected metric embedding on heterogeneous networks for link prediction. In: Proceedings of the 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2018. 1177–1186

25 Yin H Z, Hu Z T, Zhou X F, et al. Discovering interpretable geo-social communities for user behavior prediction. In: Proceedings of the 32nd IEEE International Conference on Data Engineering, 2016. 942–953

26 Yin H Z, Zou L, Nguyen Q, et al. Joint event-partner recommendation in event-based social networks. In: Proceedings of the 34th IEEE International Conference on Data Engineering, 2018

27 Xie M, Yin H Z, Xu F J, et al. Learning graph-based POI embedding for location-based recommendation. In: Proceedings of ACM International Conference on Information and Knowledge Management, 2016. 15–24

28 Fujiwara Y, Nakatsuji M, Onizuka M, et al. Fast and exact top-K search for random walk with restart. In: Proceedings of the VLDB Endowment, 2012. 442–453

29 Lü L, Zhou T. Link prediction in complex networks: a survey. Phys A-Stat Mech Appl, 2011, 390: 1150–1170
30 Powers D M W. Evaluation: from precision, recall and F-Measure to ROC, informedness, markedness & correlation. J Mach Learn Technol, 2011, 2: 37–63
31 Bryan P, Rami A, Steven S. DeepWalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2014. 701–710

Research on a link-prediction method based on a hierarchical hybrid-feature graph

Dong LI[1,2], Derong SHEN[1*], Yue KOU[1], Menger LIN[1], Tiezheng NIE[1] & Ge YU[1]

1. *School of Computer Science & Engineering, Northeastern University, Shenyang* 110004*, China*;
2. *Neusoft Corporation, Shenyang* 110179*, China*
* Corresponding author. E-mail: shenderong@cse.neu.edu.cn

Abstract Entities in the real world are often interconnected, forming heterogeneous information networks. Link-prediction is a necessary technique for predicting the existence of unobserved or future links in heterogeneous information networks. It is useful to make users better understand the generation and evolution of networks. However, current techniques lack the effective fusion of multiple features, often leading to nonsensical results. Also, it is difficult for them to adapt to the heterogeneity and dynamics of heterogeneous-information networks. In this paper, we present a hierarchical-hybrid-feature-graph (HHFG) model by fully considering structural, semantic, and time features. Also, an HHFG based link-prediction algorithm is proposed to effectively guarantee accuracy. On one hand, it performs a random walk on HHFG based upon hybrid features. On the other hand, parameters such as feature weights and transition coefficients are learned by the gradient-descent method. The experiments demonstrate the feasibility and effectiveness of our key techniques.

Keywords link-prediction, hierarchical hybrid-feature graph, heterogeneous information networks, random walk, parameters learning

Dong LI was born in 1979 and is currently a Ph.D. candidate. He received his master's degree in computer technology from Northeastern University, Shenyang, in 2008. His research interests include social-network analysis and data mining.

Derong SHEN was born in 1964. She received her Ph.D. in computer software and theory from Northeastern University, Shenyang, in 2004. Currently, she is a professor in the School of Computer Science & Engineering, Northeastern University. Her research interests include social-network analysis and data integration.