基于相互增强和知识继承的多实体动态异构学术网络构建及应用

马建红 宋秦汉
(河北工业大学人工智能与数据科学学院 天津 300401)

摘 要 对科学家、研究组织和研究资助机构来说,准确客观地衡量作者的学术成就、评估论文和地点的学术水平是一项至关重要且具有挑战性的任务。各种基于图的排名方法,如 PageRank 已经被广泛用于在同构网络中对作者、论文和地点进行排名,但是仅限于在同构网络中解决这个问题,不适用于异构网络。为此,基于作者、论文和地点三种类型的实体构建一个多实体动态异构学术网络,并提出一种新的模型 MEKI-Rank。模型中提出知识继承的概念,即作者通过其他作者继承知识,同时提取出动态异构网络中的七种关系,基于这些关系进行迭代排名,并使用每一轮的结果来相互增强排名。

关键词 多实体动态异构学术网络 知识继承 相互增强 迭代排名

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.06.042

CONSTRUCTION AND APPLICATION OF MULTI-ENTITY DYNAMIC HETEROGENEOUS ACADEMIC NETWORK BASED ON MUTUAL ENHANCEMENT AND KNOWLEDGE INHERITANCE

Ma Jianhong Song Qinhan
(*School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China*)

Abstract Accurate and objective measurement of the academic achievements of authors and assessment of the academic quality of papers and sites is a vital and challenging task for scientists, research organizations and research funding agencies. Various graph-based ranking methods, such as PageRank, have been widely used to rank authors, papers and places in homogeneous networks, but they are limited to solve this problem in homogeneous networks and are not applicable to heterogeneous networks. Therefore, this paper constructed a multi-entity dynamic heterogeneous academic network based on three entities of author, paper and place, and proposed a new model, MEKI-Rank. In this model, the concept of knowledge inheritance was proposed, that was, the author inherited knowledge from other authors. It extracted seven relationships in the dynamic heterogeneous network at the same time, carried out iterative ranking based on these relationships, and used the results of each round to enhance the ranking of each other.

Keywords Multi-entity dynamic heterogeneous academic network Knowledge inheritance Mutually reinforcing Iterative ranking

0 引 言

随着科研活动逐渐制度化,每年发表的论文越来越多,对论文和作者影响力的评价和预测也日益重要。在这种背景下,许多科学排名方法被提出来,这些方法主要分为两类:基于引文计数的方法和基于图的方法。引文计数是衡量学术影响力最传统的指标,基于引文计数的科学影响指数包括 h 指数、g 指数、c 指数和 s 指数等。IF 和 SIF-[1]等方法把所有的引用都看作是同等的,不考虑引用的质量,但一些研究人员认为,从更重要的论文或作者的引用应该获得更高的权重[2-4]。然而,基于图的排名方法可能更有效,因为它们不仅考虑引用计数,还考虑来自网络的更多信息,如

收稿日期:2021-03-02。**马建红**,教授,主研领域:软件工程,智能处理,自然语言处理,知识图谱,创新理论与方法。**宋秦汉**,硕士生。

合作作者、被引用论文的声望等。目前已经提出了一些基于图的方法来对论文进行排序[5]或同时排名论文、作者和地点[6]。

虽然科学排名方法最近取得了许多进展,但仍有一些问题值得进一步研究。第一个问题,传统的算法如 PageRank 和中心性测度等都是应用于同构网络中,也就是说,在网络中只有一种单一类型的实体和一种单一类型的关系。已有研究表明,使用混合或异构网络具有更好的性能,所有实体如作者、论文和地点都可以成为网络的一部分。第二个问题,以往的许多研究只考虑了研究者的平均绩效,而忽略了研究者的绩效会随着时间而变化。一方面,对于开始发表论文的年轻研究人员来说,他们的影响力无疑比已有的研究人员要有限得多。然而,随着时间的推移、不断学习和出版物的积累,他们可能会获得更大的影响力[7],他们中的一些人经过一段时间后可能会成为非常有声望的研究人员。另一方面,即使是在某一领域有影响力的研究人员,如果他们在很长一段时间内很少或没有发表论文,他们的影响力也会下降。同样,一个场馆的声望也会随着时间而变化。因此,在任何学术影响的模型中考虑这些时间动态影响是很重要的。

为了克服上述缺点,本文提出一种新的基于知识继承的排名模型 MEKI-Rank,在动态异构网络上提取多种规则通过迭代同时对作者、论文和地点进行排名。

本文的主要贡献如下:

(1) 构建多实体动态异构学术网络,并加入时间感知函数来提高准确性和实时性。

(2) 引入一种在三种实体的动态异构网络中同时对实体进行排序的新框架。

(3) 提出知识继承的概念,用于优化排序规则,根据作者列表的顺序,对作者和论文之间的作者权重进行分配。

1 相关研究

如何客观地评价科学论文、作者和地点,一直是许多研究者研究的一个重要课题。1972 年 Garfield[8]提出了一个引文矩阵来评估期刊的重要性,该论文讨论了用影响因子对期刊进行排名的方法,影响因子指期刊上发表的论文年平均引用次数。此后,引文计数被广泛应用于论文评估、研究人员评估和期刊评估,对于期刊评估有许多研究者提出如下评估指标如 5 年影响系数 IF[9]、特征因子得分 ES、每篇论文的标准化影响来源 SNIP、科学杂志排名 SJR 和子影响因子 SIF-等。同时,对于研究人员的评估指标有 h 指数、g 指数、R 指数、成功指数和 DS[10]指数等。

1998 年,Brin 等[11]发明了著名的基于图的 PageRank 算法来对网页进行排名,谷歌的搜索引擎就是基于此。Kleinberg 将排序工作从同构网络扩展到异构网络,单独提出了超链接诱导主题搜索 HITS,证实了排名可以通过各种类型的实体之间的相互作用来加强。因为学术网络反映了论文、研究人员和场所之间的关系,从某种意义上说,学术网络比只关注网页的网络更复杂,PageRank 和 HITS 等算法需要进行一定的修改才能用于学术绩效评估。因此,许多改进的方法如使用影响力和贡献奖励评估论文和作者排名[12]、用相似优先机制对科学出版物进行排序[13]等方法被相继提出。许多不同类型的方法已经被应用于学术能力排名和预测,Yu 等[14]的论文中,使用了四种模型,包括线性回归、k-近邻、支持向量回归和分类回归树模型来进行论文被引预测,Cao 等[15]提出了一种引文模式匹配方法,Abrishami 等[16]使用基于神经网络的方法进行论文引文预测。

最近的工作已经开始整合多个实体的异构网络,以提高排名效率。Co-Rank 结合了引文网络和相应的合作作者网络,声称对作者和文献都有更好的排名结果。FutureRank 同时对论文及其作者进行排名,不仅考虑作者合作关系和引文关系,而且还考虑了时间因素。PV-Rank 利用论文和发布场所之间的相互制约关系,同时对论文和发布场所进行排名。Liu 等[17]提出 Tri-Rank,可以同时对研究者、论文和地点进行排名。Kong 等[18]提出 TAPRank 对研究者进行排名,使用论文引用和作者信息,其模型中新论文和新引文的权重更大。Jiang 等[19]提出 MutualRank,可以同时对论文、作者和地点进行排序,据观察 PageRank 偏向于旧论文,而 HITS 偏向于新论文。因此,在 MutualRank 中采取了更平衡的解决方案。Wang 等[20]提出 MRCoRank,通过相互强化对论文、作者、地点和术语四类实体的未来流行度进行排名,多实体的最终排名是通过利用构建的文本特征和时间加权引用来生成的。

2 多实体动态异构学术网络构建

2.1 作者合著网络

合著是科研的基本特性,科研对象的跨学科性和复杂性使得专家的研究越来越趋向于合著。因此,作者合著网络[21]是学术网络框架中,最为常见的类型,其主要反映作者之间的合著信息。作者合著网络能够较为形象、直观地展示学者之间的合作关系。通过不

同类型的作者合著网络的构建和相互比较,可以发现学者在不同领域的合作关系的变化,通过社会网络分析,能确定学者的学术地位和权威性。合著网络中还包含作者间的引用关系,如图 1 所示,其中 a 表示作者,可见 a_1、a_2 同时引用 a_3 的文章,a_3、a_5 同时引用 a_4 的文章,a_1 单独引用 a_4 的文章。

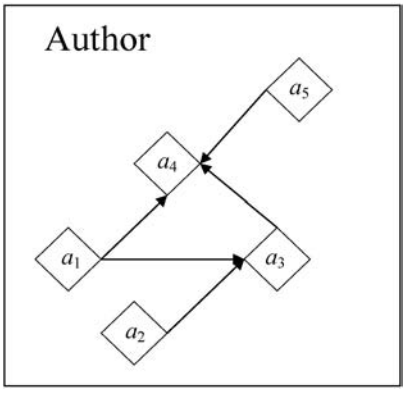

图 1　作者合著网络中的作者引用关系

2.2　论文引文网络

引文对引用论文本身和被引论文都有不同程度的微观评价作用,科学文献间的引用与被引用关系构成了论文引文网络。因此,引用关系的构成要素是实体和联系,实体可以是论文,联系可以是正向关系(引用),也可以是反向关系(被引),论文自引和论文被引是相互制约的。常见的研究方法利用作者自引网络与内容分析相结合,并根据点度中心度等指标发现了研究进程中具有重要作用的文献。论文的他引排除了作者的自引,所以往往也作为学术评价的重要指标,通过学者被引网络和学者自引网络的比较,可以更好地区分学术能力,同时区分出潜在的研究共现关系,通过网络构建,能够进一步发掘学者的学术合作关系。论文引文网络中的论文引用关系如图 2 所示,其中 p 表示论文,可见 p_1 引用 p_2、p_3,p_2 引用 p_3,p_3 引用 p_4。

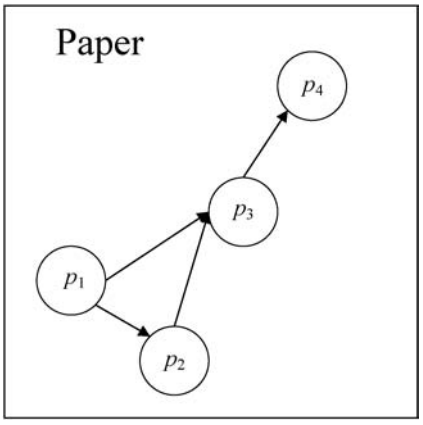

图 2　论文引文网络

2.3　地点引用网络

传统的同构网络往往会忽视地点间的引用关系,其实地点引用网络同样起着至关重要的作用。地点间的引用关系可能不如论文或作者那样复杂,但是论文和地点之间有很强的互动性,一个地点的学术水平主要是由该地点内发表的论文的平均学术水平决定。因此,被引次数多的地点自然就学术影响力高,地点引用网络如图 3 所示,其中 v 表示地点,可见 v_1、v_3 之间相互引用,v_1、v_3 又同时引用 v_2。

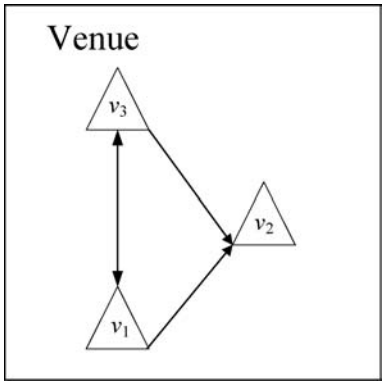

图 3　地点引用网络

2.4　多实体动态异构网络

作者、论文和地点的多实体动态异构学术网络可以表示为 $G=(V,E,W)$,如图 4 所示,其中:V 代表实体的集合;$E=V\times V$ 代表边的集合;W 代表连接每一对实体的边的权重集合。实体有三种类型:作者 V_{A}、论文 V_{P}、地点 V_{V}。边有七种关系:作者与作者间的引用关系 $E_{\mathrm{A\text{-}A}}$,论文与论文间的引文关系 $E_{\mathrm{P\text{-}P}}$,地点与地点间的引用关系 $E_{\mathrm{V\text{-}V}}$,这几种关系是在内部网络中且有方向性的,而其他类型的关系则是在外部网络中且无向的,作者与论文间的撰写关系 $E_{\mathrm{A\text{-}P}}$,多名作者同时撰写论文时的合著关系 $E_{\mathrm{A\text{-}A\text{-}P}}$,作者与地点间的发表关系 $E_{\mathrm{A\text{-}V}}$,论文与地点间的出版关系 $E_{\mathrm{P\text{-}V}}$。动态异构学术网络共有六种子网络:包括三个内部网络,作者合著网络 G_{A}、论文引文网络 G_{P}、地点引用网络 G_{V},以及三个外部网络,G_{P} 与 G_{A} 之间的二分作者网络 G_{PA}、G_{P} 与 G_{V} 之间的二分网络 G_{PV}、G_{A} 与 G_{V} 之间的二分网络 G_{AV}。

本文中动态异构学术网络支持的第一个特性是给予不同贡献的作者不同的评价。如果一篇论文是合著的作品,那么通过作者在论文中的署名顺序来区分他们的贡献。为了实现这一点,采用“1st author”“2 nd author”等来标记每一篇论文的作者的贡献等级。另一个特性是将每一个实体的特定年份视为一个独立实体来体现时间动态性,如图 4 所示,$a_1(2015)$、$a_1(2016)$ 为同一作者不同年份,因此可在网络中表示为不同的实体。

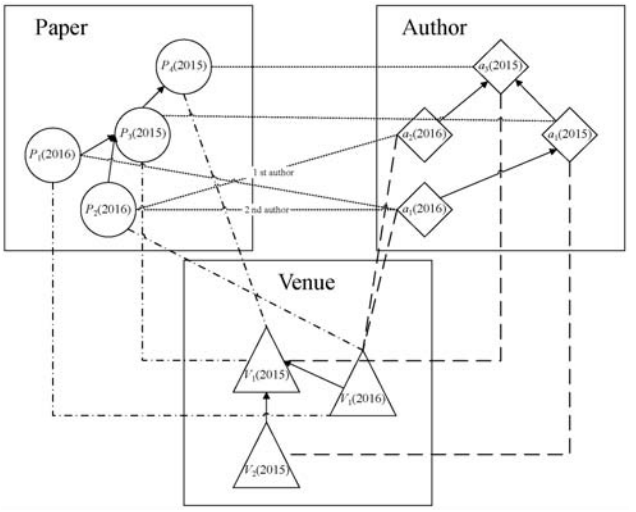

图 4 多实体动态异构学术网络

3 MEKI-Rank 模型

3.1 知识继承

本文基于社会网络分析方法中的知识场模型提出符合学术网络的知识继承概念,并根据知识继承提出 MEKI-Rank 模型。依据知识场模型,两个知识源之间知识继承的强度取决于传递者的知识能量、知识场强和继承者的知识能量。学术网络中,知识源指具有合著关系的作者,知识场强为作者之间的合著强度,知识能量为作者现有的学术产出。因此,作者之间传递的知识能量取决于合著双方的个人知识能量和合著强度。

综合考虑作者的文献数量和发表地点的层次两个因素建立衡量作者知识能量的指标,其中地点层次分为 1 区、2 区、3 区和 4 区,对应的值为 1、2、3、4,用 $Power(i)$ 表示作者 i 的知识能量,如式(1)所示。

$$Power(i) = \frac{1}{r(p)} \tag{1}$$

式中:$p \in |p|$,p 为作者 i 的第 p 篇论文,$r_{(p)}$ 为第 p 篇论文的地点层次,$|p|$为作者 i 的论文总数。

作者在合著论文时所占的贡献等级可以借助论文的署名顺序来度量,本文在构建异构学术网络时已经使用“1st author”等标记来记录每位作者对应的署名顺序,用 $order(i,p)$ 表示作者 i 在论文 p 中的署名顺序,对应的第一署名取值 1、第二署名取值 2 以此类推。则作者 i 在论文 p 中的贡献等级 $Contribution(i,p)$ 如式(2)所示。

$$Contribution(i,p) = \frac{1}{order(i,p)} \tag{2}$$

合著强度取决于论文中的所有作者的贡献等级,因此用 $Co(i,j,p)$ 表示作者 i 与 j 在论文 p 中的合著强度,如式(3)所示。

$$Co(i,j,p) = contribution(i,p) \times contribution(j,p) \tag{3}$$

基于上述作者知识能量和合著强度的描述,用 $Kn(i \rightarrow j)$ 表示作者 j 从作者 i 所继承的知识总量,即每篇合著论文中 i 传递给 j 的知识能量之和,如式(4)所示。

$$Kn(i \rightarrow j) = \sum_{p \in |p|} Power(i) \times Power(j) \times Co(i,j,p) \tag{4}$$

式中:$|p|$为作者 i 与 j 合著的全部论文;$Power(i)$ 和 $Power(j)$ 分别表示 i 和 j 的知识能量。

为确保从每个作者实体引出的所有边的权重之和为 1,需要进行标准化处理,因此基于知识继承的边加权方法可用 W_{ij} 表示,如式(5)所示。

$$W_{ij} = \frac{Kn(j \rightarrow i)}{\sum_{m=1}^{|N|} Kn(m \rightarrow i)} \tag{5}$$

式中:$|N|$为 i 的全部合著作者;$Kn(m \rightarrow i)$ 为 i 从第 m 个合著作者继承到的知识总量。

3.2 多实体动态异构学术网络中的七种关系

本文中三种实体间的关系可以分为同构网络内相同实体间的有向关系以及异构二分网络内不同实体间的无向关系两大类型。

同构网络内的关系如下:

关系 1:论文引文关系,如果论文 P_j 引用论文 P_i,用 $P_j \rightarrow P_i$ 表示,则对应权重 W_{PP} 如式(6)所示。

$$W_{PP}(p_i,p_j) = \begin{cases} 1 & p_j \rightarrow p_i \\ 0 & 其他 \end{cases} \tag{6}$$

关系 2:作者引用关系,由于作者引用一定是通过论文引文间接实现的,则对应权重 W_{AA_Cite} 如式(7)所示。

$$W_{AA\_Cite}(a_k,a_l,p_i,p_j) = \frac{1}{order(a_k,p_i) \times order(a_l,p_j)} \tag{7}$$

式中:$order(a_k,p_i)$ 为作者 a_k 在论文 P_i 中的署名顺序。

为保证论文 P_i 所有作者从论文 P_j 所有作者处获得的引用权重和为 1,因此有必要对式(7)进行标准化,如式(8)所示。

$$W_{AA}(a_k,a_l,p_i,p_j) = \frac{W_{AA\_cite}(a_k,a_l,p_i,p_j)}{\sum_{\substack{p_j \rightarrow p_i \\ a_m \in P_A(p_i) \\ a_n \in P_A(p_j)}} W_{AA\_cite}(a_m,a_n,p_i,p_j)} \tag{8}$$

式中:a_m 表示论文 P_i 的所有作者集合。

一个作者 a_l 可以多次引用另一个作者 a_k,因此 a_k 从 a_l 获得的引用总权重是所有相关论文的总和,如式(9)所示。

$$W_{AA}(a_k,a_l) = \sum_{\substack{a_l \rightarrow a_k \\ p_i \in A_P(a_k) \\ p_j \in A_P(a_l)}} W_{AA}(a_k,a_l,p_i,p_j) \tag{9}$$

式中:P_i 表示作者 a_k 的一组被引论文集合。

关系 3:地点引用关系,如果地点 v_j 引用地点 v_i,用 $v_j \rightarrow v_i$ 表示,则对应权重 W_{VV} 如式(10)所示。

$$W_{VV}(v_i,v_j)=\frac{VenueNum(v_i,v_j)}{VenueNum(v_j)} \tag{10}$$

式中:$VenueNum(v_j)$ 表示地点 v_j 所包含的所有引用总数;$VenueNum(v_i,v_j)$ 表示地点 v_j 所包含的来自地点 v_i 的引用数量。

异构网络内的关系如下:

关系 4:作者与论文间的撰写关系,假设作者 a_i 在论文 p_j 的所有作者 N 中处于第 M 位,则对应权重 W_{AP} 如式(11)所示。

$$W_{AP}(a_i,p_j)=\frac{2^{N-M}}{2^N-1} \tag{11}$$

关系 5:多名作者实体连接到同一个论文实体时会产生作者间的知识继承关系,基于本文提出的知识继承概念,对传统合著关系的加权方式进行了改进,则基于知识继承的合著权重 W_{AAP} 如式(12)所示。

$$W_{AAP}(a_i,a_j,p)=\frac{Kn(j\rightarrow i)}{\sum_{m=1}^{|N|}Kn(m\rightarrow i)} \tag{12}$$

关系 6:论文与地点间的出版关系,如果论文 P_i 在地点 v_j 中被出版它们就可以互相加权,则对应权重 W_{PV} 如式(13)所示。

$$W_{PV}(p_i,v_j)=\begin{cases}\dfrac{1}{PaperNum(v_j)} & p_i\in V_p(v_i)\\ 0 & \text{其他}\end{cases} \tag{13}$$

式中:$PaperNum(v_j)$ 表示地点 v_j 出版的论文总数。

关系 7:作者与地点间的发表关系,如果作者 a_i 在地点 v_j 发表了一篇以上论文,则对应权重 W_{AV} 应为其发表的所有论文的权重总和,如式(14)所示。

$$W_{AV}(a_i,v_j)=\sum_{\substack{p_m\in A_P(a_i)\\ p_m\in V_P(v_j)}}W_{AP}(a_i,p_m) \tag{14}$$

为保证任何地点的所有权重总和为 1,标准化公式如式(15)所示。

$$W_{AV}(a_i,v_j)=\frac{\sum_{\substack{p_m\in A_P(a_i)\\ p_m\in V_P(v_j)}}W_{AP}(a_i,p_m)}{PaperNum(v_j)} \tag{15}$$

3.3　动态时间感知

本文将三种实体的动态时间信息加入到异构网络中,能更好地预测相关实体的未来影响。

对于论文 p_i,假设其出版年份为 $T_{published}(p_i)$,评估年份为 $T_{evaluate}(p_i)$,将年份差 l 初始化为 $l=T_{evaluate}(p_i)-T_{published}(p_i)$,$l$ 年内每年对论文 p_i 的最新引用次数为 t_1,$t_2,\cdots,t_l$,标准化后的平均引用时间为 $T_{avgCitation\_l}(p_i)=t_1+t_2+\cdots+t_l/l$,则论文时间感知权重 $T(p_i)$ 如式(16)所示。

$$T(p_i)=\frac{e^{-\sigma_1(T_{evaluate}(p_i)-T_{avgCitation\_l}(p_i))}}{T_{evaluate}(p_i)-T_{published}(p_i)} \tag{16}$$

式中:σ_1 是正衰减参数。当评估年份和出版年份差值越小 $T(p_i)$ 值越大,可见论文越新权重越大,但一篇老论文在最新的年份还能吸引引用,说明其具有权威性,因此使用 σ_1 来平衡时间加权模式更合理地对待新旧论文。

对于作者 a_i,假设其发表年份为 $T_{published}(a_i)$,评估年份为 $T_{evaluate}(a_i)$,他在特定的年份发表了一组论文 $p_1,p_2,\cdots,p_m$,每篇论文的引用次数为 $t_1,t_2,\cdots,t_m$,标准化后的平均引用时间为 $T_{avgCitation\_l}(a_i)=t_1+t_2+\cdots+t_m/m$,则作者时间感知权重 $T(a_i)$ 如式(17)所示。

$$T(a_i)=\frac{e^{-\sigma_2(T_{evaluate}(a_i)-T_{avgCitation\_m}(a_i))}}{T_{evaluate}(a_i)-T_{published}(a_i)} \tag{17}$$

对于地点 v_i,假设其举办年份为 $T_{published}(v_i)$,评估年份为 $T_{evaluate}(v_i)$,用 n 表示地点发表的论文总数,标准化后的平均引用时间为 $T_{avgCitation\_l}(a_i)=t_1+t_2+\cdots+t_n/n$,则地点时间感知权重 $T(v_i)$ 如式(18)所示。

$$T(v_i)=\frac{e^{-\sigma_3(T_{evaluate}(v_i)-T_{avgCitation\_n}(v_i))}}{T_{evaluate}(v_i)-T_{published}(v_i)} \tag{18}$$

3.4　作者、论文和地点间的相互增强规则

基于动态异构网络中的七种关系提出如下的相互增强规则:

规则 1:论文的得分受到引用该论文的论文分数影响。

规则 2:论文的得分受到论文出版的地点的分数和作者的分数影响。

规则 3:作者的得分受到其合著作者的分数和引用作者的分数影响。

规则 4:作者的得分受到作者撰写的论文的分数和作者发表的地点的分数影响。

规则 5:地点的得分受到引用该地点的地点分数影响。

规则 6:地点的得分受到该地点出版的所有论文的平均分数和在该地点进行发表的所有作者的平均分数影响。

由于论文是科学研究的核心,因此从论文排名开始进行迭代是最为合理的,MEKI-Rank 对异构网络中的作者、论文和地点实体进行相互增强的步骤如下:

(1) 根据规则 1,首先在论文引文网络 G_P 中进行第一次迭代。

(2) 根据规则 2,将作者、论文和地点的得分进行更新,并对作者得分进行标准化处理,然后根据规则 3

在作者合著网络 G_A 中进行一次迭代。

(3) 根据规则4,用最新的作者、论文得分更新地点得分,并对地点得分进行标准化处理,然后根据规则5在地点引用网络 G_V 中进行一次迭代。

(4) 根据规则6,用最新的作者、地点得分更新论文得分,并对论文得分进行标准化处理,返回到步骤(1)。

3.5 MEKI-Rank 排序算法

基于异构网络提供的信息,MEKI-Rank 排序算法对每个涉及的实体进行打分。在为所有实体设置初始值之后,将对它们应用一个迭代过程,在每个步骤中,每个实体都获得一个更新的分数。注意,所有涉及的实体都相互影响。该算法在满足终止条件时停止(例如,给定的迭代次数或更新的阈值),算法的伪代码细节如算法1所示。

算法1 MEKI-Rank 算法

输入:实体集合 V_P、V_A、V_V,边集合 E_{PP}、E_{AA}、E_{VV}、E_{AP}、E_{AAP}、E_{AV}、E_{PV},权重 W_{PP}、W_{AA}、W_{VV}、W_{AP}、W_{AAP}、W_{PV}、W_{AV},参数 α、β、γ、σ_1、σ_2、σ_3、λ、ε、μ、t_a、t_v。

输出:论文、作者和地点的得分 P_S、A_S、V_S。

1 for

2 //初始化所有实体得分

3 $PS(p_i)^0 = \frac{1}{|V_P|}, AS(a_j)^0 = \frac{1}{|V_A|}, VS(v_k)^0 = \frac{1}{|V_V|}$

4 end

5 $t = -1, \Delta = 2\varepsilon$

6 while $\Delta > \varepsilon$ do

7 $t = t + 1$

8 for: //计算作者得分

9 $pas(a_i)^{t+1} = \frac{1}{t_a + 1}[\sum_{a_k \in p_n}(a_i, t_a) AS(a_k)^t + AS(a_i)^t]$

10 end for //计算地点得分

11 $pvs(v_i)^{t+1} = \frac{1}{t_v + 1}[\sum_{v_k \in p_n}(v_i, t_v) VS(v_k)^t + VS(V_i)^t]$

12 end for //更新论文得分

13 $temp(p_i)^{t+1} = \alpha \sum W_{AP} \times pas(a_j)^{t+1} + (1 - \alpha) \times W_{PV} \times pvs(v_k)^{t+1}$

14 $temp(p_i)^{t+1} = \sum W_{PP} \times temp(p_j)^{t+1} + \mu \times temp(p_i)^{t+1}$

15 $PS(p_i)^{t+1} = \lambda \times T(p_i) \times temp(p_i)^{t+1} + (1 - \lambda) \times \frac{1}{|V_P|}$

16 end for //更新作者得分

17 $temp(a_i)^{t+1} = \beta \sum W_{AP} \times PS(p_j)^t + (1 - \beta) \sum W_{AV} \times pvs(v_k)^{t+1}$

18 $temp(a_i)^{t+1} = \sum W_{AA} \times temp(a_j)^{t+1} + \mu \times temp(a_i)^{t+1}$

19 $AS(a_i)^{t+1} = \lambda \times T(a_i) \times temp(a_i)^{t+1} + (1 - \lambda) \times \frac{1}{|V_A|}$

20 end for //更新地点得分

21 $temp(v_i)^{t+1} = \gamma \sum W_{PV} \times PS(p_i)^t + (1 - \gamma) \sum W_{AV} \times pas(a_k)^{t+1}$

22 $VS(a_i)^{t+1} = \lambda \times T(v_i) \times \sum W_{VV} \times temp(v_j)^{t+1} + (1 - \lambda) \times \frac{1}{|V_V|}$

23 end

24 //标准化论文、作者和地点得分

25 $\Delta = \sum |PS(p_i)^{t+1} - PS(p_i)^t| + \sum |AS(a_i)^{t+1} - AS(a_i)^t| + \sum |VS(v_i)^{t+1} - VS(v_i)^t|$

26 end

第1-第4行初始化三种类型的实体,对于属于同一类别的所有实体,都给予相同的分数 $1/n$,其中 n 为该类别的实体数量。第5行初始化参数 t、Δ,第9行计算了每个作者的平均得分,第11行计算了每个场馆的平均表现。

算法的主要部分包含在一个 while 循环中,在循环内部(第6-第26行),迭代更新所有涉及的实体的分数。论文的新分数在第12-第15行计算,需要考虑三个因素:作者(第13行)、地点(第13行)和引文(第13行)。作者的新分数在第16-第19行计算,考虑四个因素:已发表论文(第17行)、论文发表地点(第17行)、作者被引用次数(第18行)和已发表论文的合著者(第18行)。场馆的新分数在第20-第22行计算,需要考虑三个因素:发表论文(第21行)、作者(第21行)和地点引用(第22行)。对于这三种类型的节点,都考虑了时间因素:第15行为论文,第19行为作者,第22行为地点。

算法中对于出版年份接近评估年份的论文给予了鼓励,在第14行,参数 μ 用于调整这个因子。近五年内发表的任何论文,即使没有被任何论文引用,也可以获得一定的分数。同样地,作者也被以同样的方式对待(第18行)。

4 实 验

4.1 数据集预处理

本文采用微软学者数据集 Microsoft Academic Graph 作为研究对象,该数据集由微软官方发布,包含多种不同类型的实体及其交互关系信息,如论文标题、作者、研究领域、时间、引用情况和所发表的会议或期刊等数据,共收录了1.26亿篇文章,发表年份从1800年到2016年不等。完整的微软学术网络数据集规模庞大,为了避免处理过程过于复杂,从中提取出数据挖掘领域子数据集进行分析研究,并使用计算机科学图书馆 DBLP 数据集对各实体间引用关系进行补充。实

验前对数据集进行如下预处理:首先删除那些既不引用任何其他论文也不接受任何引用的论文,因为它们是完全分离的很难评估它们的影响;其次对作者和地点实体进行人工消歧;最后对作者人数超过 5 位的论文只保留前 5 位作者。结果如表 1 所示,共包含论文 157 305 篇,作者 187 834 名,地点 9 755 处,引用次数 2 578 724 次。

表 1　数据集信息

论文数量	157 305
作者数量	187 834
地点数量	9 755
引用次数	2 578 724

4.2　参数设置和评估指标

为了对作者、论文和地点进行迭代排名,本文从动态异构学术网络中提取七种关系,并对 MEKI-Rank 模型进行如下参数设置,首先让 $\alpha=0.70$、$\beta=0.60$、$\gamma=0.10$、$\lambda=0.85$、$\sigma_1=0.5$、$\sigma_2=1.0$、$\sigma_3=1.1$,然后对 μ 分别设置 0.8、0.6、0.4 和 0.2 四种值对应发表年份不同的新旧论文例如 2004 年、2003 年、2002 年和 2001 年等,最后通过 3.4 节中的相互增强规则进行算法迭代。

(1) 作者评估指标。选择了三个度量标准来评估作者排名的结果,第一个是精度,如式(19)所示。

$$P_{\text{recision@N}}=\frac{|S_{\text{GT}}\cap S_{\text{A}}|}{|S_{\text{A}}|} \tag{19}$$

式中:S_{GT} 是查询的基本事实;S_{A} 是方法 A 返回的前 top-N 结果。

第二个是二元偏好测度,如式(20)所示。

$$B_{\text{pref}}=\frac{1}{|R|}\sum_{r\in R}1-\frac{|R\text{ 个作者中前 }N\text{ 名}|}{M} \tag{20}$$

式中:R 表示作者总数;M 表示关系总数。

第三个是折扣累计收益 DCG,使用二进制形式,在特定的等级位置 P 处累积的 DCG 如式(21)所示。

$$D_{\text{CG}_p}=\sum_{i=1}^{p}\frac{2^{r_{\text{el}_i}}-1}{\log_2(i+1)} \tag{21}$$

式中:$r_{\text{el}_i}\in\{0,1\}$。

(2) 论文评估指标。评估论文排名更加困难,本文通过分析不同方法之间的相关性来评估论文排名的结果,两种方法的相关性 $C(A,B)$ 如式(22)所示。

$$C(A,B)=\frac{|r_A(k)\cap r_B(k)|}{|r_A(k)\cup r_B(k)|} \tag{22}$$

式中:A、B 是两种排名方法;$r_A(k)$ 和 $r_B(k)$ 是 A 和 B 返回的两组 Top-k 论文。

(3) 地点评估指标。使用修正影响因子 MIF 来衡量其影响,对于给定的地点 v,其 MIF(v)如式(23)所示。

$$MIF(v)=\frac{CitationNum(v)}{|v|} \tag{23}$$

式中:$CitationNum(v)$ 是所有在 v 地点发表的论文获得的引文总数,$|v|$ 为该地点的论文总数,因此 $MIF(v)$ 是 v 地点每篇论文获得的平均引文数。

4.3　对比方法

MEKI-Rank 模型进行对比的方法如下:

(1) PageRank。著名的排名算法,最初是为网页排名而设计的,后被广泛应用于其他网络节点权威排名中,包括学术评估、作者和地点排名等。

(2) RandHITS。一种基于图表的作者排名算法,并加入了随机种子进行优化。

(3) Co-Rank。同时对作者、论文进行迭代排名。

(4) Tri-Rank。在异构网络中同时对作者、论文和地点进行排名。

(5) MutualRank。通过集成异构网络中的相互增强关系来同时对论文、作者和地点进行排名。

(6) MEKI-Rank。基于知识继承,在动态异构学术网络中对作者、论文和地点进行迭代排名。

4.4　实验结果及分析

作者排名:为了对作者的权威进行排名,从 ArnetMiner 中收集的前 20 名专家被选为该领域的基准排名。ArnetMiner 是一个著名的系统,为学术界提供全面的搜索和挖掘服务。首先将数据挖掘领域的作者按 4.3 节排名算法进行排名,得到前 10 名的作者及其论文,如表 2 所示,括号中的数字是网络基准中相关专家的排名情况。

表 2　数据挖掘领域十大作者及论文

序号	顶类作者	顶尖论文
1	Jiawei Han(1)	Mining interesting knowledge using DM-II
2	Pedro Domingos	Mining association rules with multiple minimum supports
3	Jian Pei(6)	Mining the most interesting rules
4	Philip S. Yu(2)	MetaCost: a general method for making classifiers cost-sensitive
5	Christos Faloutses(7)	CACTUS-clustering categorical data using summaries
6	Mohammed J. Zaki(8)	On the merits of building categorization systems by supervised clustering
7	Johannes Gehrke(4)	Prediction with local patterns using cross-entropy
8	Rakesh Agrawal	Data mining: crossing the Chasm

续表 2

序号	顶类作者	顶尖论文
9	Bing Liu(9)	Horting hatches an egg: a new graph-theoretic approach to collaborative filtering
10	Jon Kleinberg	Pruning and summarizing the discovered associations

可见,MEKI-Rank 对作者排名后的结果还是比较准确的,将基准排名第一的作者找出,并且排名的前 10 位作者有 7 位都是基准排名以内的作者。

再使用精度 Precision@ N、二元偏好测度 Bperf、DCG 分别对六种对比方法进行比较,图 5 展示了六种方法的对应精度结果,分别取 5、10、15、20 和 100 次为迭代上限。表 3 显示六种方法的 Bperf、DCG20 和 DCG100 结果。结果表明,MEKI-Rank 在所有三个指标上都优于三个竞争对手。以 DCG_{20}为例,MEKI-Rank 的平均提高率分别为 41.0%、38.2% 和 79.1%。

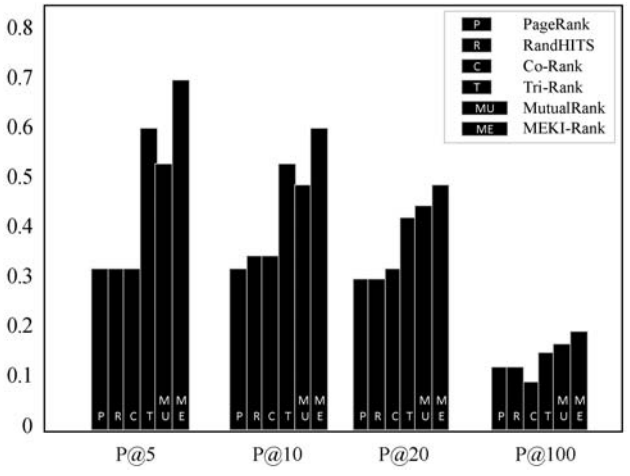

图 5 作者排名精度对比

表 3 作者排名参数对比

指标	Page Rank	Rand HITS	Co-Rank	Tri-Rank	Mutual Rank	MEKI-Rank
Bpref	0.817	0.833	0.752	0.873	0.876	0.877
DCG_{20}	2.877	2.981	2.335	4.762	3.973	4.843
DCG_{100}	4.008	4.215	3.845	5.816	5.657	5.948

论文排名:表 4 显示了在“数据挖掘”和 $k=100$ 的条件下,六种排名方法之间的相关性。

表 4 六种方法相关性

排名方法	Page Rank	Rand HITS	Co-Rank	Tri-Rank	Mutual Rank	MEKI-Rank
Page Rank	1.000	0.459	0.098	0.149	0.176	0.153
Rand HITS	0.459	1.000	0.047	0.212	0.155	0.231
Co-Rank	0.098	0.047	1.000	0.063	0.078	0.054

续表 4

排名方法	Page Rank	Rand HITS	Co-Rank	Tri-Rank	Mutual Rank	MEKI-Rank
Tri-Rank	0.149	0.212	0.063	1.000	0.262	0.314
Mutual Rank	0.176	0.155	0.078	0.262	1.000	0.203
MEKI- Rank	0.153	0.231	0.054	0.314	0.203	1.000

六种排名方法之间的相关性较低,因为不同的方法有其自身的特点。PageRank,论文是按引文排序的;RandHITS,论文以一种强化的方式进行排名;Co-Rank,论文的排名受作者排名的影响;而对于 MEKI-Rank 来说,论文的排名受作者排名和发表地点的影响,MEKI-Rank 返回的排名靠前的论文大多是排名靠前的作者写的或者是在排名靠前的地点发表的,其论文排名结果如表 5 所示。

表 5 论文排名前 10

序号	ID	顶尖论文
1	J93-2003[9]	The Mathematics of Statistical Machine Translation: Parameter Estimation
2	P02-1040[4]	Pruning and summarizing the discovered associations
3	J92-4003[3]	Prediction with local patterns using cross-entropy
4	H91-1060[2]	A Maximum Entropy Model for Part-of-Speech Tagging
5	P95-1026[6]	Unsupervised Word Sense Disambiguation Rivaling Supervised Methods
6	J92-1002[3]	Text Chunking Using Transformation-Based Learning
7	P96-1025[2]	CACTUS-clustering categorical data using summaries
8	J98-1004[3]	Mining interesting knowledge using DM-II
9	P05-1012[3]	Mining the most interesting rules
10	E91-1060[2]	Learning Accurate, Compact, and Interpretable Tree Annotation

地点排名:在每个领域中,总有几个知名的顶级场地,MEKI-Rank 返回的地点排名结果符合我们的常识。例如,在“数据挖掘”领域,由三阶返回的排名前 10 位的地点是 SIGMOD、ICDE、VLDB、KDD、ICDM 等,如表 6 所示。

表 6 地点排名前 10

序号	顶尖地点	$MIF(v)$	论文总数
1	SIGMOD	1.873	121 502
2	ICDE	1.758	114 812
3	VLDB	1.564	10 685
4	KDD	1.521	96 812
5	ICDM	1.395	86 364
6	SIGIR	1.152	75 894

续表 6

序号	顶尖地点	$MIF(v)$	论文总数
7	PODS	1.107	65 345
8	SDM	1.069	64 217
9	ICML	0.988	56 689
10	ECML	0.972	45 871

综上所述,在对比的方法中,无论是精度还是排名准确率本文提出的基于知识继承的排名模型 MEKI-Rank 都有着最优秀的结果,可见知识继承概念合理地应用到学术网络中是可以起到优化作用的。

5 结　语

本文基于社会网络分析方法中的知识场模型提出符合学术网络的知识继承概念,通过构建多实体动态异构学术网络,从中提取各实体间的 7 种相互关系,提出知识继承理论模型即 MEKI-Rank 模型,并与现有排名方法进行对比实验,实验结果分别从作者排名、论文排名和地点排名几方面证明本文模型更有效。准确的排名结果对于提高学术研究效率、节约研究成本至关重要。将来,可在更多更复杂的领域进行测试,并通过优化 MEKI-Rank 中的相互关系和参数来提高性能。此外,未来的研究方向是进一步考虑其他实体信息来完善多实体动态异构学术网络,如主题实体和领域实体等,并对本文构建的学术网络进行更广泛的应用,如学者影响力的评估、学者个性化评估、学术成功因素挖掘和学者年龄预测等。

参 考 文 献

[1] Xu F, Liu W, Rousseau R. Introducing sub-impact Factor (SIF-) sequences and an aggregated SIF-indicator for journal ranking[J]. Scientometrics,2015,102:1577 – 1593.

[2] Li L, Wang X, Zhang Q, et al. A quick and effective method for ranking authors in academic social network[C]//Multimedia and Ubiquitous Engineering,2014:179 – 185.

[3] Wan X, Liu F. Are all literature citations equally important? Automatic citation strength estimation and its applications [J]. Journal of the Association for Information Science and Technology,2014,65(9):1929 – 1938.

[4] Zhu X, Turney P, Lemire D, et al. Measuring academic influence: Not all citations are equal[J]. Journal of the Association for Information Science and Technology, 2015, 66 (2):408 – 427.

[5] Zhou J, Zeng A, Fan Y, et al. Ranking scientific publications with similarity-preferential mechanism[J]. Scientometrics,2016,106:805 – 816.

[6] Wang S, Xie S, Zhang X, et al. Coranking the future influence of multiobjects in bibliographic network through mutual reinforcement[J]. ACM Transactions on Intelligent Systems and Technology,2016,7(4):1 – 28.

[7] Lee D. Predicting the research performance of early career scientists[J]. Scientometrics,2019,121:1481 – 1504.

[8] Garfield E. Citation analysis as a tool in journal evaluation [J]. American Association for the Advancement of Science, 1972,178(4060):471 – 479.

[9] Pajić D. On the stability of citation-based journal rankings [J]. Journal of Informetrics,2015,9(4):990 – 1006.

[10] Farooq M, Khan H, Iqbal S, et al. DS-index: Ranking authors distinctively in an academic network[J]. IEEE Access,2017(5):19588 – 19596.

[11] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine[J]. Computer Networks and ISDN Systems,1998,30(1):107 – 117.

[12] Dunaiski M, Visser W, Geldenhuys J. Evaluating paper and author ranking algorithms using impact and contribution awards[J]. Journal of Informetrics,2016,10(2):392 – 407.

[13] Zhou J, Zeng A, Fan Y, et al. Ranking scientific publications with similarity-preferential mechanism[J]. Scientometrics,2016,106(2):805 – 816.

[14] Yu T, Yu G, Li P, et al. Citation impact prediction for scientific papers using stepwise regression analysis[J]. Scientometrics,2014,101(2):1233 – 1252.

[15] Cao X, Chen Y, Liu K. A data analytic approach to quantifying scientific impact[J]. Journal of Informetrics,2016,10 (2):471 – 484.

[16] Abrishami A, Aliakbary S. Predicting citation counts based on deep neural network learning techniques[J]. Journal of Informetrics,2019,13(2):485 – 499.

[17] Liu Z, Huang H, Wei X, et al. Tri-Rank: An authority ranking framework in heterogeneous academic networks by mutual reinforce[C]//2014 IEEE 26th International Conference on Tools with Artificial Intelligence,2014:493 – 500.

[18] Kong X, Zhou J, Zhang J, et al. TAPrank: A time-aware author ranking method in heterogeneous networks[C]//2015 IEEE International Conference on Smart City/SocialCom/SustainCom,2015:242 – 246.

[19] Jiang X, Sun X, Yang Z, et al. Exploiting heterogeneous scientific literature networks to combat ranking bias: Evidence from the computational linguistics area[J]. Journal of the Association for Information Science and Technology, 2016,67(7):1679 – 1702.

[20] Wang S, Xie S, Zhang X, et al. Coranking the future influence of multiobjects in bibliographic network through mutual reinforcement[J]. ACM Transactions on Intelligent Systems and Technology,2016,7(4):1 – 28.

[21] 王炎,魏瑞斌. 基于多数据源的专家学术网络构建研究[J]. 情报杂志,2016,35(12):121 – 126,138.