基于主题声望和动态异构网络的学术影响力排序算法

陈潘, 陈红梅, 罗川

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230100037

基于主题声望和动态异构网络的学术影响力排序算法

陈　潘[1]　陈红梅[2,3,4,5]　罗　川[6]
1 西南交通大学唐山研究院　河北 唐山 063000
2 西南交通大学计算机与人工智能学院　成都 611756
3 可持续城市交通智能化教育部工程研究中心　成都 611756
4 综合交通大数据应用技术国家工程实验室　成都 611756
5 四川省制造业产业链协同与信息化支撑技术重点实验室　成都 611756
6 四川大学计算机学院　成都 610065
(pchen@my.swjtu.edu.cn)

摘　要　有效地挖掘学术大数据,分析论文的学术影响力,有助于科研工作者获取重要的信息。文本内容与学术网络结构的动态变化,会对论文的学术影响力排名结果产生重要的影响。但现有的论文学术影响力排序算法或是缺乏对文本内容的考虑,或是缺乏对学术网络结构的动态变化的考虑。针对该问题,提出了一种学术影响力排序算法,称之为基于主题声望和动态异构网络的学术影响力排名(TND-Rank)。TND-Rank衡量了论文主题在某一时间对论文的影响,并将其嵌入考虑时间因素的论文影响力排序算法中。TND-Rank通过考虑影响主题声望水平、期刊、作者、时间等多种因素的综合影响来计算论文的动态学术影响力相关排名。在实验中,对AMiner数据集1936—2014年间发表且信息保存完整的文章进行了分析,将所提算法与近年来的4种相关算法进行了比较,采用Spearman相关系数、归一化折损累积增益(NDCG)和分级平均精度(GAP)对算法性能进行了评估。实验结果验证了TND-Rank算法的可行性和有效性,其可以有效地综合各种信息对论文的学术影响力进行排序。

关键词:异构网络;学术影响力;学术大数据;主题声望;论文排序
中图分类号　TP391

Academic Influence Ranking Algorithm Based on Topic Reputation and Dynamic Heterogeneous Network

CHEN Pan[1],CHEN Hongmei[2,3,4,5] and LUO Chuan[6]
1 Tangshan Research Institute,Southwest Jiaotong University,Tangshan,Hebei 063000,China
2 School of Computing and Artificial Intelligence,Southwest Jiaotong University,Chengdu 611756,China
3 Engineering Research Center of Sustainable Urban Intelligent Transportation,Ministry of Education,Chengdu 611756,China
4 National Engineering Laboratory of Integrated Transportation Big Data Application Technology,Southwest Jiaotong University,Chengdu 611756,China
5 Manufacturing Industry Chains Collaboration and Information Support Technology Key Laboratoryof Sichuan Province,Southwest Jiaotong University,Chengdu 611756,China
6 College of Computer Science,Sichuan University,Chengdu 610065,China

Abstract　Effectively mining academic big data and analyzing academic influence of papers are benefical for researchers to obtain important information. The dynamic changes of text content and academic network structure have an important impact on the ranking results of academic impact. However,the existing ranking algorithms of academic influence of papers either lack consideration of text contents or the dynamic changes of academic network structure. To solve this problem,this paper proposes an algorithm for ranking academic influence,which is called TND-Rank,based on topic reputation and dynamic heterogeneous network. In TND-Rank,the impact of the topic on the paper at a certain time is measured and embedded to the paper influence ranking algorithm that takes into account the time factor. The dynamic ranking related to the academic impact of a paper is calculated by

到稿日期:2023-01-09　返修日期:2023-05-25
基金项目:国家自然科学基金(61976182,62076171);四川省自然科学基金(2022NSFSC0898);四川省科技成果转移转化示范项目(2022ZHCG0005)
This work was supported by the National Natural Science Foundation of China(61976182,62076171),Natural Science Foundation of Sichuan Province(2022NSFSC0898) and Sichuan Science and Technology Achievements Transfer and Transformation Demonstration Project (2022ZHCG0005).
通信作者:陈红梅(hmchen@swjtu.edu.cn)

comprehensively considering the influence of various factors, i. e, the level of topic prestige, journal, author, and time etc. In the experiments, the AMiner data set published between 1936 and 2014 with complete information are analyzed, and compared with four related algorithms in recent years. Spearman correlation coefficient, normalized discounted cumulative gain(NDCG) and graded average precision(GAP) are adopted to evaluate performance of the algorithm. Experimental results verify the feasibility and effectiveness of the proposed algorithm TND-Rank, which can effectively synthesize various information to rank the academic influence of papers.

Keywords Heterogeneous network, Academic influence, Academic big data, Thematic prestige, Thesis ranking

1 引言

学术大数据是研究者成果的积累，也是研究者获取新兴知识的重要渠道。每个研究人员均需查阅其研究领域的相关论文，深入了解相关领域的热门问题、最新研究方向等。随着科学技术的不断进步，学术成果不断涌现，论文数量大且随时间增加呈指数型增长[1]。论文爆炸性的增长增加了研究人员快速准确地获取相关领域感兴趣论文的难度，综合评价论文的影响力已经成为学术大数据研究的重要课题之一。针对论文学术影响力进行分析和研判，帮助研究人员聚焦重要的科研成果，推动科学技术的不断发展，有着重要的意义。

引文计数因能简单而有效地对学者或论文的重要性进行排序[2]，是目前应用最广泛的学术影响力评价指标。基于引文计数在衡量学术影响力方面的有效性，研究人员提出了一系列基于引文计数的影响评估指标。期刊影响因子[3]、g 指数[4]、h 指数[5]等都是具有代表性的基于引文的指标。针对以上指标存在的问题，研究人员提出了它们的各种变体，例如 t 指数[6]、k 指数[7]、Mo 指数[8]等，它们都从不同的角度考虑了出版物的数量和质量来评价学术影响力。但这些方法均存在一定的局限性，现在的学者们更倾向于基于图论的排序算法，例如 HITS[9]和 PageRank[10]算法，来衡量不同学术网络拓扑下的科学影响力。

虽然现在的许多研究在学术影响力影响评价方面取得了多方面的成果，但仍有许多具有挑战性的问题未得到解决。1)学术大数据的异构性和动态性导致了一个高度多样化的学术网络，这对探索作者、论文与期刊等其他学术实体之间的关系提出了挑战。2)文本内容对学术实体的排序结果产生了影响，论文的文本内容是研究人员选择一篇论文的关键。当论文主题与研究人员研究兴趣相匹配时，该论文才会得到研究人员的关注。同时，我们认为当前流行的主题对论文在当前年份的影响也应该更大。

基于以上问题，我们提出了一种基于主题声望和动态异构网络的学术影响力排序算法。该方法根据每年论文的主题，借助每篇论文的共现关系构建主题网络，利用介数中心性计算主题词在该年份的声望值。同时，考虑到学术网络的动态性，在以往的学术实体排序算法及学术网络各实体之间的边上加入时间函数，以拟合网络的动态性。最后，将主题声望加入改进的 HITS 算法中，计算出论文基于主题声望的当前年份学术影响力排名。

2 相关工作

在过去的几十年里，针对学术实体的排序算法取得了长足的进步。在最初的学术影响力评价中，经常使用一些经典的引文计量学指标，例如评价作者的 h 指数[5]、g 指数[4]，评价期刊的期刊影响因子[3]、五年影响因子[11]和论文的被引量等。然而，这些都是一些结构化的指标，其虽然对相关学术实体的影响力评价做出了贡献，但单一的衡量方式不能精确地评价论文的重要性。它们忽略了一些重要的论文的元数据，如作者和期刊，而且由简单的引文计数演变而来的指标难免会受到引文操纵[12]的影响。对此，研究人员进一步提出了引文的结构化评价方法。

基于引文的结构化评价是近年来流行的一种学术影响力评价的方法。传统的结构化方法是将论文集合视为静态引文网络，将论文视为网络节点，将论文之间的引用关系视为论文节点之间的边，在该网络上利用 PageRank 算法来计算论文的威望，并将此作为最终的论文学术影响力得分。Wang 等利用论文、作者、期刊这 3 个学术实体作为节点建立了一种异构网络，并在相关边加入了时间感知权重，使用 PageRank 和 HITS 改进的算法对论文进行了学术影响力的排序[13]。该方法一定程度上解决了 PageRank 偏向旧论文的弊端，但其缺乏对论文文本内容的考虑。Amjad 等在论文、作者和期刊构成的异构网络中加入主题信息[14]，分别讨论了论文、作者和期刊中论文关键词在特定主题领域内出现的频率，并以此来定义该实体在具体领域的得分，然后将这个得分作为相关实体初始值代入算法进行迭代。但是，该方法未考虑相关网络的权重。Yu 等在异构网络中还加入了一种合著者网络，更加丰富了网络中的结构及内容信息[15]。Bai 等在其中加入了机构信息作为一种新的节点，在新的异构网络中计算学术影响力[16]。Lu 等提出了一种基于 PageRank 和 HITS 特征的引文网络影响力模型[17]，将计算的 PageRank 和 HITS 中的权威值与中心值等特征加入学习排序算法 SVMRank 中，根据所得的影响力分数来评估论文的重要性。

上述方法都是基于静态网络的方法，它们在完整的引文网络上进行分析，对引文网络多年来的演变一无所知，这是基于静态网络方法的严重缺陷。Jiang 等提出了一个能够很好地捕捉科学领域中的知识演化的模型，发现论文的声望随着科学领域的发展而不断演变[18]。他将引文网络按时间分为各个时间快照，在各个时间快照下的网络中计算论文的学术影响力得分，在上一代引文网络中的得分加权的基础上加上新一代网络的得分加权。Franceschet 按时间顺序递增处理引文。在 $t>0$ 的任意时间，对于在时间 t 引用的学者的评级，根据他们在时间 $t-1$ 的先前评级和引用学者先前评级更新，来动态地对学者进行评价[19]。Rakoczy 将期刊或会议作为节点，利用发布在上面的论文的引文关系得出期刊/会议的引文网络，利用格兰杰因果关系考查动态网络中各节点对彼此的影响[20]。Zhou 等充分考虑了学术网络上各实体之间边的

权重关系，并通过加入时间函数缓解了传统排序算法对旧论文的偏差问题，提出了一种基于异构学术网络的加权 P-Rank 算法[21]。

动态网络中可以使用异构网络和语义信息来丰富网络中包含的信息。Zhang 等利用论文、作者、期刊节点构建出 7 个同质网络，利用马尔可夫链模型对各实体的未来学术影响力进行预测[22]。Zhou 等在研究人员-研究人员、研究人员-论文和论文-论文 3 种基本关系的基础上，引入并定义了一套度量方法，用于量化构建多维网络模型中的一系列学术实体（如研究人员和论文）之间基于活动的协作关系、专业感知链接和话题感知引文适合度之间的相关性，提出了一种基于 RWR (Random Walk With Restart)的改进算法。该算法在一定的社会背景下重新定义和度量了时变的学术影响力[23]。Zhou 等考虑了文献计量学中的老化效应，在模拟引文网络中随机游走算法的过程中引入了指数衰减函数，提出了一种基于年龄扩散的模型来识别未来更可能被引用的论文[24]。

论文的文本内容是研究人员寻找论文的一个重要标准，因此越来越多的研究人员关注这一点并创新了评价方法。比如，Zhang 等将网络结构与论文语义相结合，提出了一种新的论文引文网络的权重量化方法，将 HITS 算法与异构文献计量学网络相结合，对论文进行排名[25]。Amjad 等通过考虑作者声望水平和学术网络中论文主题的影响，提出了一种基于主题的学术影响力异构排名算法[26]。Xu 等提出了一个基于论文文本内容和时间敏感的随机文献浏览模型，通过模拟研究人员的阅读行为得出论文的声望得分[27]。Huang 等认为论文链接在引文网络中与被引论文的主题和被引论文的重要性有关，根据这两个因素提出了一种可扩展的产生式模型 MTID，将文本主题和主题在引文网络上的影响力相结合，以发现有影响力的文章[28]。Wang 等针对文章主题是否会影响其未来的被引趋势问题，建立了一个机器学习框架，研究了论文的哪些特征将对其今后的引文行为产生重要的价值。其中，主题特征在未来高影响力文章预测中有着必不可少的作用[29]。

但以往对文本内容的研究缺少对主题声望的考虑。对此，本文讨论了异构动态网络中学术实体基于主题的声望，并考虑了作者、论文以及期刊和会议等场所的相互影响，提出了一种基于主题声望的异构排名算法 TND-Rank，用于衡量学术实体在异构动态网络中的学术影响力。

3　基于主题声望的异构排名算法 TND-Rank

本文利用 LDA[30]模型在论文摘要和标题中得出论文的主题关键词，利用主题在论文中的共现关系建立主题的共现网络图，利用介数中心度等网络的结构指标发现主题在动态网络中的声望，将其加入由 HITS 改进的算法中求得论文的学术影响力。

本文方法分为 3 个步骤：计算主题声望，计算论文声望，论文影响力排序。3.2 节将介绍在数据集中提取出论文的主题词并构建网络的过程，3.3 节介绍如何在主题网络中计算主题的声望，3.4 节介绍计算论文影响力的方法。本文的基本框架如图 1 所示。

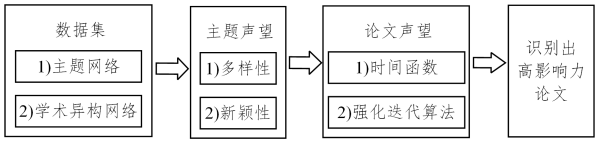


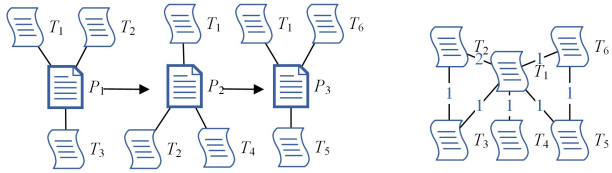
图 1　本文框架

Fig. 1　Framework

3.1　主题词提取与主题网络构建

首先使用 LDA[30]在 Aminer[31]数据集提供的摘要与文章标题中提取出论文的主题；然后基于整个数据集构建全局主题共现网络，并计算各节点的介数中心度。主题共现网络是一个无向加权图，其中节点是主题，边表示节点之间的共现关系。

主题共现网络如图 2 所示。图 2(a)中，P_1，P_2 和 P_3 表示 3 篇论文，它们之间的边表示引用关系。T_1，T_2，…，T_6 表示主题词，它们与论文节点的边表示它们在该篇论文中出现。图 2(b)给出了这些主题词在该引文网络中的主题共现网络，其中边的权重值是由它们在同一篇论文中共同出现的次数进行加权。

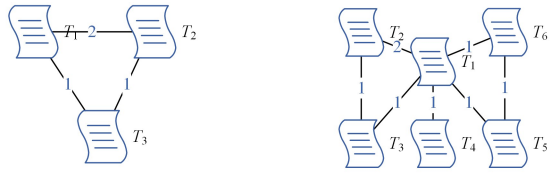


(a) 引文网络　　(b) 主题共现网络

图 2　全局主题网络

Fig. 2　Global theme network

然后，识别出与这些主题相关的论文，并为每个主题建立一个局部主题共现网络，对每个主题的内部知识结构进行建模。局部主题网络本质上是以自我为中心的网络，其中自我节点是主题关键词。局部主题网络如图 3 所示。



(a) 以T_2为中心　　(b) 以T_1为中心

图 3　局部主题网络

Fig. 3　Local theme network

3.2　计算主题声望

本节提出了一种基于主题共现网络的主题声望评价方法。我们认为一个主题如果与其他主题共同出现在一篇论文中的频率高，那么它是具有高跨主题领域性质的多样性的。如果一个主题与主题共现网络中未出现过的主题存在高频率的链接，那么这个主题是存在一定的新颖性的。本文利用主题的多样性和新颖性来评价主题的声望值。

3.2.1　主题的多样性

度中心性[32](Degree Centrality)是在网络分析中刻画节点中心性的最直接度量指标。一个节点的节点度越大，意味着该节点的度中心性越高，该节点在网络中就越重要。而节点介数指一个网络中通过节点的最短路径条数，得出的介数中心度可以反映出该节点与其他节点的连接关系。我们使用

3.2 节中建立的基于主题的全局主题网络计算出各个节点在不同时间的介数中心度，并用其代表该节点(主题)的多样性。

首先，我们计算了全局主题网络中每个节点 n_i 在时间 t 的加权介数中心度 $BC_{n_i}(t)$。如式(1)所示，其中 gi_{pq} 表示全局主题网络中节点 n_p 和 n_q 之间的加权最短路径的总数，而 si_{pq} 表示通过节点 l 的路径的数量。

$$BC_{n_i}(t)=\sum_{p\neq i\neq q}\frac{s_{pq}^{i}}{g_{pq}^{i}} \tag{1}$$

对其进行归一化：

$$BC_{n_i}(t)=\frac{1}{(m-1)(m-2)/2}\sum_{p\neq i\neq q}\frac{s_{pq}^{i}}{g_{pq}^{i}} \tag{2}$$

其中，m 表示全局主题网络节点的数量。

3.2.2 主题的新颖性

主题的新颖性 N_{n_i} 的计算式如式(3)所示，其中 n_{new} 表示在局部主题网络中出现的新节点。新节点定义为在该时间点之前从未出现在局部主题网络中的节点。这里不是简单地统计局部主题网络中的新节点数量，而是根据新节点的介数中心度计算分数。

$$N_{n_i}=\sum_{n_j\in n_{\text{new}}}BC_{n_j} \tag{3}$$

其中，n_j 为与 n_i 直接相连的新节点，即 n_j 与 n_i 出现在同一篇论文中，且 n_j 为当前时间新出现的主题词。

3.2.3 主题的声望值

结合主题的多样性和新颖性，按照式(4)计算某一主题在某一时间的声望值。

$$T_{n_i}(t)=BC_{n_i}(t)\cdot N_{n_i} \tag{4}$$

3.3 计算论文的学术影响力

构建学术异构网络，基于该网络计算论文的学术影响力。

3.3.1 学术异构网络

首先构建论文、作者和期刊的异构网络，将其表示为 $G=(V,E)$，其中 V 表示节点集合，E 表示边集合。

节点类型有 3 种：论文 V_P、作者 V_A、期刊 V_V。对应有 3 种类型的边：作者与论文的边 $E\_(P\text{-}A)$ 表示论文 P 由作者 A 撰写，论文与期刊的边 $E\_(P-V)$ 表示论文 P 在期刊 V 上发表，论文与论文 $E\_(P\text{-}P)$ 表示论文之间的引用关系。其中，$E\_(P-P)$ 是有向的，其他边是无向的。

为有效地比较研究人员在不同年份的表现，将特定年份的每个作者或期刊/会议视为一个单独的实体。期刊/会议的处理方式与此类似。

构建的学术异构网络如图 4 所示。

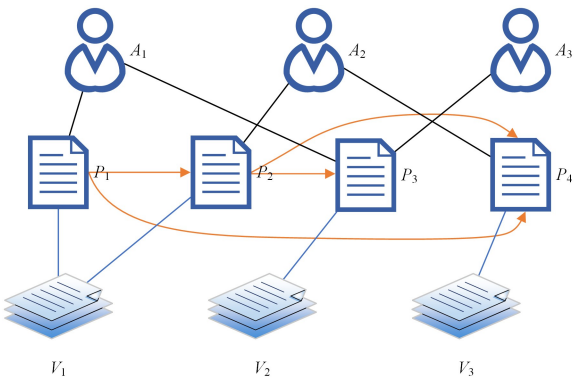


图 4 学术异构图

Fig. 4 Academic heterogeneous network

图 4 中，作者、论文和期刊/会议分别用 A_1-A_3，P_1-P_4 和 V_1-V_3 节点表示。3 种类型的边 E_{P-A}(论文作者)、E_{P-V}(论文期刊)和 E_{P-P}(论文引用)都可以在图 4 中找到。

3.3.2 基于主题声望的排序算法

本文中的方法基于以下现象：

1)声誉高的作者撰写的论文质量可能更高；

2)顶级期刊上所发表的论文质量可能更高；

3)质量高的论文通常会被其他高质量的论文引用。

鉴于此，确定以下规则：

1)如果作者的声望值高，那么他所撰写的论文将获得高的权威值；

2)如果期刊的声望值高，那么发布在该期刊上的论文将获得高的权威值；

3)如果引用该论文的论文声望值高，那么该被引论文将获得高的权威值。

因此，使用改进的 HITS 来计算论文的声望得分。该算法的步骤如下：

1)首先将所有论文的权威值得分初始化为 $1/N_p$，其中 N_p 为集合中论文的数量；

2)根据式(7)在作者-论文网络中计算作者的 hub 分数 $H(A_i)$；

3)根据式(9)在论文-期刊网络中计算期刊的 hub 分数 $H(V_i)$；

4)根据式(10)在论文-论文网络中计算论文的 hub 分数 $H(P_i)$；

5)利用根据式(15)计算出的在引文网络中传播的 PageRank 分数 $PageRank(P_i)$、作者的 hub 分数 $H(A_i)$即作者传播的分数、期刊的 hub 分数 $H(V_i)$即期刊传播的分数、论文的 hub 分数 $H(P_i)$即论文传播的分数，以及根据式(17)计算的论文的时间感知分数 $Time_{P_i}$ 和根据式(18)计算出的当前时间的主题声望 AT_{P_i}，更新论文的权威值得分 A_{P_i}；

6)反复执行步骤 2)－步骤 5)，直到收敛。

HITS[9]算法的全称为“Hyperlink-Induced Topic Search”，其由 Joh Kleinberg 提出，最初用于对网页排序。研究人员将该算法用于学术网络中计算学术影响力。在 HITS 算法中，网络中的节点被赋予了两个属性：权威值(Authority)和枢纽值(Hub)。其中，权威值表示节点的自身重要性，枢纽值表示该节点与其他节点的连接程度。如果一个节点的权威值高，那么会有更多的节点指向它；一个节点的枢纽值越高，那么它指向的节点越多。

权威值和枢纽值是相互影响的，它们的计算方式如式(5)和式(6)所示。

网页 i 的权威值＝所有指向它的网页 Hub 值之和，即：

$$A_i=\sum_{j\in P_{\text{to}}}H_j \tag{5}$$

网页 i 的枢纽值＝网页 i 指向的所有网页 Authority 值之和，即：

$$H_i=\sum_{j\in P_{\text{from}}}A_j \tag{6}$$

其中，P_{to} 表示指向网页 i 的网页集合，P_{from} 表示网页 i 指向的网页集合。在初始状态下，网页节点的权威值和枢纽值均设置为 1。

在本文的学术异构网络中，相关计算式如下。

1)计算 Hub 分数

Hub 分数包括作者的 Hub 分数、期刊的 Hub 分数和论文的 Hub 分数，其定义如下。

(1)作者的 Hub 分数

在作者-论文网络中，作者的 Hub 分数由该作者所撰写的论文的权威值之和计算。这里将其进行了归一化：

$$H(A_i)=\frac{\sum_{P_j\in Set_{A_i}^P}A(P_j)\cdot f(rank_{A_i})\cdot f_1(age)}{|Set_{A_i}^P|\cdot\sum_{P_j\in Set_{A_i}^P}f(rank_{A_i})\cdot f_1(age)} \tag{7}$$

其中，$H(A_i)$表示作者A_i的 Hub 分数；$SetP_{A_i}$表示作者A_i撰写的论文集合；$|SetP_{A_i}|$表示作者A_i撰写的论文数量；$f(rank_{A_i})$为根据作者顺位分配分数的权重函数。

$$f(rank_{A_i})=\frac{\varphi}{\sigma_i}+b \tag{8}$$

其中，φ表示函数的放大倍数，σ_i表示作者i在论文中的顺位，b表示函数的垂直偏移量。

计算中心分数时，考虑到旧论文引用信息更多，Hub 分数应该更高，因此给旧论文更高的时间权重。这里设置$f_1(age)=\omega^{t_{current}-t_j}$，$t_{\text{current}}$为当前时间；$t_j$为论文$P_j$的发布时间；$\omega$为常数，本文设置为 2。

(2)期刊的 Hub 分数

在论文-期刊网络中，期刊的 Hub 分数由该期刊上发布的所有论文的权威值之和计算。

$$H(V_i)=\frac{\sum_{P_j\in Set_{V_i}^P}A(P_j)\cdot f_1(age)}{|Set_{V_i}^P|\cdot\sum_{P_j\in Set_{V_i}^P}f_1(age)} \tag{9}$$

其中，$H(V_i)$表示期刊V_i的 Hub 分数；$Set_{V_i}^P$表示发布在期刊V_i的论文集合；$|Set_{V_i}^P|$表示发布在期刊V_i的论文数量。

(3)论文的 Hub 分数

在论文引文网络中，论文的 Hub 分数由该论文引用的所有论文的权威值之和计算。

$$H(P_i)=\frac{\sum_{P_j\in Set_{P_i}^P}A(P_j)\cdot f_1(age)}{|Set_{P_i}^P|\cdot\sum_{P_j\in Set_{P_i}^P}f_1(age)} \tag{10}$$

其中，$H(P_i)$表示论文P_i的 Hub 分数；$Set_{P_i}^P$表示该论文的参考文献集合；$|Set_{P_i}^P|$表示该论文的参考文献数量。

2)计算论文的 Authority 分数

下文首先给出作者传播给论文的权威分数的计算方式，然后给出改进的 HITS 算法以计算论文的影响力。

(1)作者传播给论文的权威分数

在论文-作者网络中，我们认为论文的重要性主要由撰写论文的作者和作者在论文中的顺位共同决定。顺位高的作者对论文的重要性影响更大。因此论文从其相应作者处收到的权威值的计算式为：

$$A_{P_i}^A=\frac{1}{N(A)}f(rank_{A_i})\cdot\sum_{A_j\in Set_{P_i}^A}H(A_j)\cdot f_2(age) \tag{11}$$

其中，$Set_{P_i}^A$表示论文P_i的作者集合；$H(A_j)$为作者A_j的 Hub 分数；$N(A)$表示所有作者传播给论文的分数之和。计算权威分数时，考虑到新论文应该比旧论文更重要，权威分数应该更大，因此给新论文更大的时间权重。这里设置$f_2(age)=\frac{1}{t_{\text{current}}-t_j+1}$。$f(rank_{A_i})$的计算式如式(8)所示。

(2)期刊传播给论文的权威分数

$$A_{P_i}^V=\frac{1}{N(V)}\sum_{V_j\in Set_{P_i}^V}H(V_j)\cdot f_2(age) \tag{12}$$

其中，$Set_{P_i}^V$表示论文P_i的期刊；$H(V_j)$表示期刊V_j的 Hub 分数；$N(V)$表示所有期刊传播给所有论文的分数之和。

(3)论文传播给论文的权威分数

$$A_{P_i}^P=\frac{1}{N(P)}\sum_{P_j\in Set_{P_i}^P}H(P_j)\cdot f_2(age) \tag{13}$$

其中，$Set_{P_i}^P$表示引用论文P_i的论文集合；$H(P_j)$表示论文的 Hub 分数；$N(P)$表示所有论文传播给论文的分数之和。

3.3.3　改进的 HITS 算法

最终论文的权威分数可以由论文的 Pagerank 值、作者传播给论文的权威分数$A_{P_i}^A$、期刊传播给论文的权威分数$A_{P_i}^V$、论文传播给论文的权威分数$A_{P_i}^P$，以及当前时间快照下论文主题新颖性和多样性传播给论文的分数$A_{P_i}^T$线性组合求得，如式(14)所示：

$$A_{P_i}=\alpha PageRank(P_i)+\beta A_{P_i}^A+\gamma A_{P_i}^V+\delta A_{P_i}^P+\theta Time_{P_i}+\eta A_{P_i}^T+(1-\alpha-\beta-\gamma-\delta-\theta-\eta)\frac{1}{N_P} \tag{14}$$

其中，$\alpha,\beta,\gamma,\delta,\theta,\eta$是范围在(0,1)的参数，实验中设置$\alpha+\beta+\gamma+\delta+\theta+\eta=0.85$；$(1-\alpha-\beta-\gamma-\delta-\theta-\eta)\frac{1}{N_p}$为随机跳跃的概率，$N_p$为网络中的论文数量；$PageRank(P_i)$为论文$P_i$在引文网络中的 Pagerank 值，如式(15)所示：

$$PageRank(P_i)=\sum_{P_j\in Set_{\text{from}}^{P_i}}\frac{A_{P_j}\cdot f_3(age)}{N(Set_{\text{to}}^{P_j})} \tag{15}$$

其中，$Set_{\text{from}}^{P_i}$表示引用论文P_i的论文集合；$N(Set_{\text{to}}^{P_j})$表示论文P_i引用的论文数量；A_{P_j}表示论文P_j更新前的权威分数。由于 PageRank 存在偏向旧论文的缺陷，本文加入了一种时间函数$f_3(age)$对其进行修正。

$$f_3(age)=\frac{1}{t_{\text{current}}-t_j} \tag{16}$$

其中，$Time_{P_i}$为论文的时间感知值，如式(17)所示：

$$Time_{P_i}=e^{-\rho(t_{current}-t_j)} \tag{17}$$

其中，$A_{P_i}^T$为当前时间论文主题新颖性和多样性传播给论文的分数，如式(18)所示：

$$A_{P_i}^T=\frac{\sum_{n_i\in P_i}T_{n_i}(t)}{N(n_i)} \tag{18}$$

其中，$T_{n_i}(t)=BC_{n_i}(t)\cdot N_{n_i}$表示当前时间主题$n_i$在主题网络中基于多样性和新颖性的得分；$N(n_i)$表示论文$P_i$的主题数量。

基于异构网络和主题网络提供的信息，利用主题网络的排序算法 DNT-Rank 对涉及的每个实体进行评分。在为所有实体设置初始值 $1/N$ 之后，将对它们进行迭代，并且在每个步骤中每个实体都获得更新的分数。请注意，所有涉及的实体都会相互影响。当满足终止条件(更新阈值小于 0.000 001)时，该算法停止。最终获得的论文的权威分数即为论文的学术影响力评分。

算法 1 给出了所提出的 DNT-Rank 算法的完整计算过程。

算法 1 DNT-Rank

输入：全局主题网络G_T；学术异构图 G=(V,E)；论文发布时间列表 T；学者顺位列表 σ

参数：$\alpha,\beta,\gamma,\delta,\theta,\eta,t_{current},\rho,b,\varphi$

输出：论文权威分数A_P

步骤：

1. 在全局主题网络中计算主题声望G_T：$T_{n_i}(t)=BC_{n_i}(t)\cdot N_{n_i}$；
2. 通过式(3)－式(16)计算主题传播分数A_P^T；
3. 计算学者的权重函数 $f(rank_A)$；
4. 计算论文的时间感知分数$Time_P$；
5. 初始化论文的权威分数$A_P=\frac{1}{N_p}$；
6. While 未收敛 do
7. 更新 PageRank 值：
   $PageRank(P)^t \leftarrow PageRank(P)^{t-1}$
8. 更新 Hub 分数：
9. 通过式(7)更新作者 Hub 值：
   $H(A)^t \leftarrow GetHubScore(f(rank_A), f_1(age), A_P^{t-1})$
10. 通过式(9)更新期刊 Hub 值：
    $H(V)^t \leftarrow GetHubScore(f_1(age), A_P^{t-1})$
11. 通过式(10)更新论文 Hub 值：
    $H(P)^t \leftarrow GetHubScore(f_1(age), A_{P_j}^{t-1})$
12. 更新传播分数：
13. 通过式(11)计算作者传播给论文的权威分数：
    $A_P^{A^t} \leftarrow GetAuthScore(f(rank_A), f_2(age), H_A^{t-1})$
14. 通过式(12)计算期刊传播给论文的权威分数：
    $A_P^{V^t} \leftarrow GetAuthScore(f_2(age), H_V^{t-1})$
15. 通过式(13)计算引文传播给论文的权威分数：
    $A_P^{P^t} \leftarrow GetAuthScore(f_2(age), H_P^{t-1})$
16. 通过式(14)更新论文的权威分数：
    $A_P = \alpha PageRank(P)^t + \beta A_P^{A^t} + \gamma A_P^{V^t} + \delta A_P^{P^t} + \theta Time_P + \eta A_P^T + (1-\alpha-\beta-\gamma-\delta-\theta-\eta)/N_P$
17. 结束

4 实验与分析

为验证所提算法的有效性，在公用数据集上将其与 4 种相关算法进行了比较和分析。

4.1 数据集及其预处理

本文使用的数据集来自 Aminer 数据集[31]，包含 1936－2014 年间的 2092356 篇论文和 8024869 条引文关系，每篇论文给出了标题、作者、发布年份、发布期刊/会议、参考文献以及摘要等数据。

这里对数据进行了以下预处理：首先，删除没有标题、摘要、期刊、发布年份、参考文献的论文；然后，构建引文网络，这里删除了参考文献不足 3 篇的论文，因为这些论文可能不会被正确识别；最后，从数据中提取出期刊，手动将一些名字不一样但应该是同一期刊的标注为同一期刊。最终的数据集中包含 486542 篇论文和 2694289 条引文关系。

4.2 实验设置和评估指标

本节介绍实验比较算法和评估指标等相关内容。

4.2.1 实验设置

本文将提出的 DNT-Rank 算法与 PageRank＋HITS、CTPM 和 BTPR 等基线算法进行了比较。对于基本事实列表，我们采用 Wang 等[13]提出的未来引文量方法。将 2008 年作为时间节点，将论文在 2008 年之后出现的新被引次数进行了统计，并将其作为未来引文量，即为基本事实列表。

同时，本文提出了该算法的几个变体来验证每个部分的功能。3 种变体分别为：删除主题因素的 DNTWT-Rank、删除时间因素的 DNTWD-Rank 和同时删除主题因素与时间因素的 DNTWTD-Rank。各变体模型如下。

1)DNTWT-Rank：删除主题因素 $A_{P_i}^T$，如式(19)所示：

$$A_{P_i}=\alpha PageRank(P_i)+\beta A_{P_i}^A+\gamma A_{P_i}^V+\delta A_{P_i}^P+\theta Time_{P_i}+(1-\alpha-\beta-\gamma-\delta-\theta)\frac{1}{N_P} \tag{19}$$

其中，$\alpha+\beta+\gamma+\delta+\theta=0.85$。

2)DNTWD-Rank：删除 3.3.3 节中所有 $f(age)$函数及 $Time_{P_i}$，如式(20)所示：

$$A_{P_i}=\alpha PageRank(P_i)'+\beta A_{P_i}^{A'}+\gamma A_{P_i}^{V'}+\delta A_{P_i}^{P'}+\eta A_{P_i}^T+(1-\alpha-\beta-\gamma-\delta-\eta)\frac{1}{N_P} \tag{20}$$

其中，$PageRank(P_i)'$，$A_{P_i}^{A'}$，$A_{P_i}^{V'}$，$A_{P_i}^{P'}$ 在式(11)－式(13)、式(15)的基础上删除了 $f(age)$函数；$\alpha+\beta+\gamma+\delta+\eta=0.85$。

3)DNTWTD-Rank：同时删除主题因素 $A_{P_i}^T$ 和 3.3.3 节中的所有 $f(age)$函数及 $Time_{P_i}$，如式(21)所示：

$$A_{P_i}=\alpha PageRank(P_i)'+\beta A_{P_i}^{A'}+\gamma A_{P_i}^{V'}+\delta A_{P_i}^{P'}+(1-\alpha-\beta-\gamma-\delta)\frac{1}{N_P} \tag{21}$$

其中，$\alpha+\beta+\gamma+\delta=0.85$。

4.2.2 评估指标

采用 Spearman 相关系数[33]、归一化折损累积增益(NDCG)[34]和分级平均精度(GAP)[35]对算法效果进行评估。

1)Spearman

Spearman 可用于比较两个排序列表的相似度，如式(22)所示：

$$\rho=\frac{\sum_i(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i(x_i-\bar{x})^2}\sqrt{\sum_i(y_i-\bar{y})^2}} \tag{22}$$

其中，x_i表示论文 i 在 x 列表中的顺位，y_i表示论文 i 在 y 列表中的顺位；$\bar{x}$和$\bar{y}$表示所有论文的平均排名位置。

NDCG 是一种 TOP-K 的排序结果的常用评价指标，它可以体现 TOP-K 个结果的相对排序及其是否真的相关。计算式如下。

2)累积增益 CG

$$CG_k=\sum_{i=1}^{k}rel_i \tag{23}$$

其中，rel_i表示位置 i 的预测结果相关性，k 表示列表大小。

3)折损累积增益 DCG

$$DCG_k=\sum_{i=1}^{k}\frac{2^{rel_i}-1}{\log_2(i+1)} \tag{24}$$

4)归一化折损累积增益 NDCG

$$NDCG_k=\frac{DCG_k}{IDCG} \tag{25}$$

其中，$IDCG$ 表示排列到最好状态的列表所得的 DCG 值。

5）分级平均精度 GAP

GAP 将平均精度（AP）推广到多等级相关性的情况，并继承了 AP 度量的重要属性。计算式如下。

（1）计算第 k 个位置贡献的平均精度（E[PC]）。

$$E[PC_k]=\sum_{j=1}^{g_k}\left(\frac{1}{k}|d_i:i\leqslant k,g_i\geqslant j|\right)\cdot p_j \tag{26}$$

其中：

$$E[PC_k]=\frac{1}{k}\sum_{m=1}^{k}\delta(m,k)$$

$$\delta(m,k)=\begin{cases}\sum_{j=1}^{\min(g_k,g_m)}p_j, & g_m>0\\ 0, & \text{其他}\end{cases} \tag{27}$$

（2）计算排序的文档所贡献的累积平均精度的最大数量。

$$\sum_{i=1}^{c}R_i\sum_{j=1}^{i}p_j \tag{28}$$

其中，R_i 表示等级 i 的相关文档数量。

（3）根据（1）和（2）计算排名前 k 的列表 R 的分级平均精度 GAP。

$$GAP@K=\frac{\sum_{k=1}^{K}\frac{1}{k}\sum_{m=1}^{k}\delta(m,k)}{\sum_{i=1}^{c}R_i\sum_{j=1}^{i}p_j} \tag{29}$$

4.3　比较算法

将 DNT-Rank 与以下基线算法进行了比较。

1）PageRank＋HITS[13]：该方法是由 Wang 等在 2013 年提出的。他们在异构网络中同时利用不同类型的信息，使用 PageRank＋HITS 结合的算法对论文进行评估。但是该方法未涉及论文的文本内容，缺少主题因素对论文评估带来的影响。

2）CTPM[36]：该方法是由 Zhang 等在 2018 年提出的。他们采用了一种新的主题模型 CTM，并使用了改进的 PageRank 算法模型。考虑了主题比例、主题相关性以及期刊威望修正偏向概率和转移概率，以此来对论文影响力进行评估。该方法是一篇考虑了主题内容的论文评估算法，同时也考虑了主题的演变，并计算了期刊随主题的动态学术影响。

3）Min Tao[37]：该方法是由 Tao 等在 2020 年提出的。他们根据论文的 LDA 计算概率分布，利用 PageRank 算法对参考网络中的选项进行重新排序，将排序结果作为论文推荐结果。该方法考虑了主题内容，但缺乏对主题演变的考虑。

4）BTPR[38]：该方法是由 Dayeh 等在 2022 年提出的。他们提出了一种新的指标，称之为引文平均变化率，将其加入 PageRank 提出了一种新的排序算法，解决了 PageRank 中偏向旧论文的问题。该方法是目前新发表的一种论文评估算法，但该方法未考虑论文主题因素。

上述算法均在 Aminer 数据集上使用最优的参数设置与本文算法进行比较。

4.4　实验结果

基线算法所得排序列表的前 10，50，100 和 200 名与基本事实列表对比的 Spearman 相关系数、NDCG、GAP 评价指标显示的评估性能如图 5－图 7 所示。

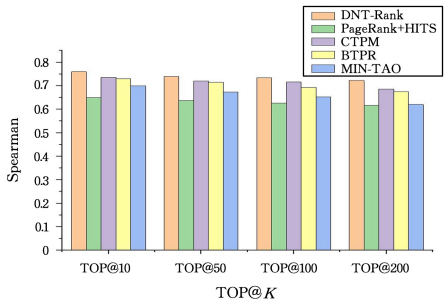


图 5　论文排名方法的准确性（斯皮尔曼）

Fig. 5　Accuracy of paper ranking methods (Spearman)

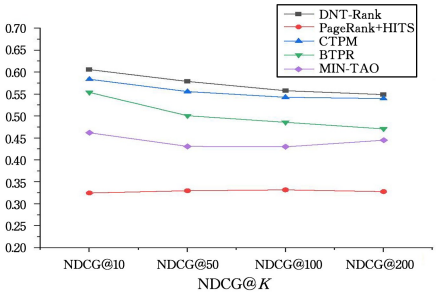


图 6　NDCG 的平均得分实验结果

Fig. 6　Experimental results of average score of NDCG

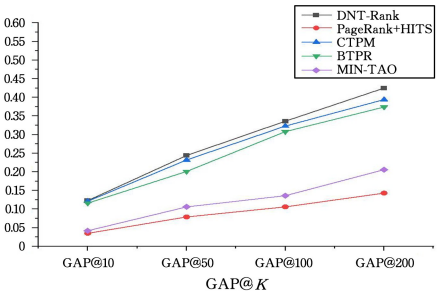


图 7　GAP 的平均得分实验结果

Fig. 7　Experimental results of average score of GAP

由图 5－图 7 可见，DNT-Rank 算法持续且显著地优于基线算法的效果。从图 5 可以看出，本文算法对论文的未来引用量影响最大，这可能是因为目前声望值高的主题在未来几年依然是研究人员重点关注的对象。NDCG 评价指标考虑了不同内容在搜索结果中的相对重要性的影响。更高的 NDCG，意味着更重要的结果将位于搜索结果的前面。因此，本文算法对最具影响力的论文的识别能力也更出众。MAP 值越高，排序结果越接近基本事实列表，同样说明本文算法具有较高的准确性。

4.5　消融实验

为了验证模型中考虑时间因素与主题因素对算法结果的影响，删除了主题因素和时间因素来与原算法及基线算法进行比较。实验结果如图 8－图 10 所示。从图 8－图 10 可见，未删改的 DNT-Rank 效果仍是最好的，而删除了时间因素与主题因素的算法效果有着明显的下降。可以看出，删除时间因素对算法效果的影响最大，这可能是因为研究人员都倾向引用最近发布的论文，所以在排序算法中加入时间函数能更好地识别出有影响力的论文。虽然主题声望对论文声望

的影响不如时间因素，但其结果仍然比 PageRank＋HITS 算法的效果要好，说明主题因素也是论文评价不可缺少的指标。

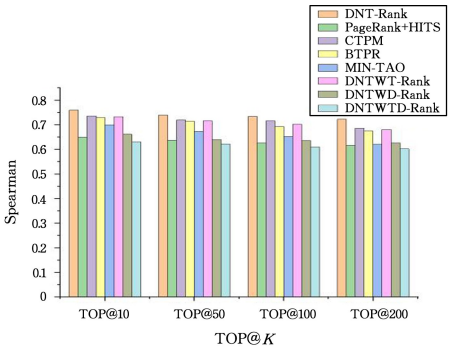


图 8 与 DNT-Rank 变体对比的 Spearman 结果

Fig. 8 Spearman results compared with DNT Rank variants

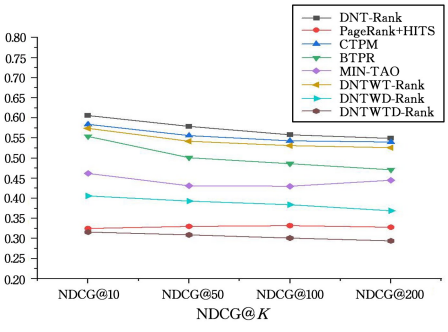


图 9 与 DNT-Rank 变体对比的 NDCG@K 结果

Fig. 9 NDCG results compared with DNT Rank variants

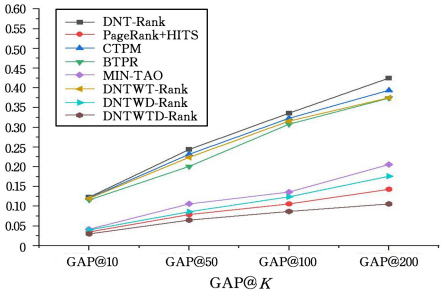


图 10 与 DNT-Rank 变体对比的 GAP@K 结果

Fig. 10 GAP results compared with DNT Rank variants

为了更明显地体现出时间因素与主题因素对算法效果的影响，将 DNT-Rank 算法的变体在 Spearman 评价指标上所占原算法的百分比列出，如图 11 所示。从图 11 可见，3 种变体的效果相比原算法都有着明显的下降。其中，删除主题因素的 DNTWT-Rank 仍有原算法 95％的效果，但删除了时间因素的 DNTWD-Rank 仅占原算法 87％的效果。对此，从式(7)、式(9)－式(13)和式(15)中可看到，在计算各学术实体的权威值、中心值与 PageRank 值时均加入了时间衰减函数，因此删除时间衰减函数对原算法的影响较大。从现实中可以理解为，研究人员更愿意寻找最近的研究成果进行借鉴。而对于主题因素，本算法仅在计算论文权威分数时有所考虑，这可能是造成其影响不如时间因素的原因之一。未来工作中可以考虑将主题因素加入学术异构网络的边权重量化等方面。

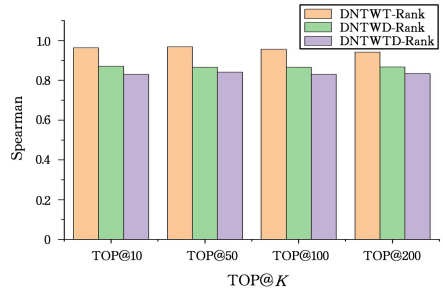


图 11 DNT-Rank 变体与 DNT-Rank 算法对比所占 Spearman 百分比

Fig. 11 Percentage of Spearman between DNT-Rank variants and DNT-Rank algorithm

结束语 学术影响力分析是数据挖掘领域的研究热点。本文旨在探讨对论文的声望产生影响的因素，着重研究了文章的主题因素与时间因素对论文声望带来的影响。我们提出了一个基于结构信息的主题影响模型并将其加入论文学术影响力排序算法中，加入了时间因素来解决算法中缺少动态因素的问题，充分考虑网络中各边的权重值以保证实验结果结果的准确性。实验结果表明，我们的论文排名顺序与实际论文排名相似，验证了算法的有效性。本文着重研究了文章内容对其未来引文行为的影响。实验结果表明，文章的内容特征对其未来的引文行为有影响，但由于文章的内容只是基于其主题，在将来的工作中，我们将继续探索如何更好地量化文章的内容特征，进一步考查文本内容在学术影响力评估中的价值。

参 考 文 献

[1] WU Z, WU J, KHABSA M, et al. Towards building a scholarly big data platform: Challenges, lessons and opportunities[C]// IEEE/ACM Joint Conference on Digital Libraries. IEEE, 2014: 117-126.

[2] GARFIELD E. Citation analysis as a tool injournal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies[J]. Science, 1972, 178(4060): 471-479.

[3] BAI X, ZHANG F, HOU J, et al. Implicitmulti-feature learning for dynamic time series prediction of the impact of institutions [J]. IEEE Access, 2017, 5: 16372-16382.

[4] EGGHE L. An improvement of the h-index: The g-index[J]. ISSI newsletter, 2006, 2(1): 8-9.

[5] HIRSCH J E. An index to quantify an individual's scientific research output[J]. Proceedings of the National academy of Sciences, 2005, 102(46): 16569-16572.

[6] SINGH P K. t-index: entropy based random document and citation analysis using average h-index[J]. Scientometrics, 2022, 127(1): 637-660.

[7] KAPTAY G. The k-index is introduced to replace the h-index to evaluate better the scientific excellence of individuals[J]. Heliyon, 2020, 6(7): e04415.

[8] FAT M O. Mo-Index for multi-authors papers[J]. Annals of Library and Information Studies(ALIS), 2022, 69(4): 323-326.

[9] KLEINBERG J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM(JACM), 1999, 46(5): 604-632.

[10] PAGE L, BRIN S, MOTWANI R, et al. The pagerank citation ranking: Bring order to the web[R]. Technical Report, Stanford University, 1998.
[11] NIEROP E. The introduction of the 5-year impact factor: does it benefit statistics journals? [J]. Statistica Neerlandica, 2010, 64(1): 71-76.
[12] BARTNECK C, KOKKELMANS S. Detecting h-index manipulation through self-citation analysis[J]. Scientometrics, 2011, 87(1): 85-98.
[13] WANG Y, TONG Y, ZENG M. Ranking scientific articles by exploiting citations, authors, journals, and time information [C]// Twenty-seventh AAAI Conference on Artificial Intelligence. 2013.
[14] AMJAD T, DING Y, DAUD A, et al. Topic-based heterogeneous rank[J]. Scientometrics, 2015, 104: 313-334.
[15] YU D, WANG W, ZHANG S, et al. A multiple-link, mutually reinforced journal-ranking model to measure the prestige of journals[J]. Scientometrics, 2017, 111: 521-542.
[16] BAI X, ZHANG F, NI J, et al. Measure the impact of institution and paper via institution-citation network[J]. IEEE Access, 2020, 8: 17548-17555.
[17] LU Y, MA K, DUAN J. Influence Model of Paper Citation Networks with Integrated PageRank and HITS[C]// 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD). IEEE, 2021: 1081-1086.
[18] JIANG X, GAO C, LIANG R. Ranking scientific articles in a dynamically evolving citation network[C]// 2016 12th International Conference on Semantics, Knowledge and Grids (SKG). IEEE, 2016: 154-157.
[19] FRANCESCHET M, COLAVIZZA G. TimeRank: A dynamic approach to rate scholars using citations[J]. Journal of Informetrics, 2017, 11(4): 1128-1141.
[20] RAKOCZY M E, BOUZEGHOUB A, GANCARSKI A L, et al. Time-Dependent Influence Measurement in Citation Networks [J]. Complex Systems Informatics and Modeling Quarterly, 2018(17): 24-43.
[21] ZHOU J, LIU S, FENG L, et al. Weighted P-Rank: a Weighted Article Ranking Algorithm Based on a Heterogeneous Scholarly Network[C]// Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8—12, 2021, Proceedings, Part I. Cham: Springer International Publishing, 2021: 537-548.
[22] ZHANG F, WU S. Predicting future influence of papers, researchers, andvenues in a dynamic academic network[J]. Journal of Informetrics, 2020, 14(2): 101035.
[23] ZHOU X, LIANG W, KEVIN I, et al. Academic influence aware and multidimensional networkanalysis for research collaboration navigation based on scholarly big data[J]. IEEE Transactions on Emerging Topics in Computing, 2018, 9(1): 246-257.
[24] ZHOU Y, LI Q, YANG X, et al. Predicting the popularity of scientific publications by an age-based diffusion model[J]. Journal of Informetrics, 2021, 15(4): 101177.
[25] ZHANG Y, WANG M, GOTTWALT F, et al. Ranking scientific articles based on bibliometric networks with a weighting scheme [J]. Journal of Informetrics, 2019, 13(2): 616-634.
[26] AMJAD T, DING Y, DAUD A, et al. Topic-based heterogeneous rank[J]. Scientometrics, 2015, 104: 313-334.
[27] XU H, MARTIN E, MAHIDADIA A. Contents and time sensitive document ranking of scientific literature[J]. Journal of Informetrics, 2014, 8(3): 546-561.
[28] HUANG X, CHEN C, PENG C, et al. Topic-sensitive influential paper discovery in citation network[C]// Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3—6, 2018, Proceedings, Part II 22. Springer International Publishing, 2018: 16-28.
[29] WANG M, JIAO S, ZHANG J, et al. Identification high influential articles by considering the topic characteristics of articles [J]. IEEE Access, 2020, 8: 107887-107899.
[30] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(Jan): 993-1022.
[31] TANG J, ZHANG J, YAO L, et al. Arnetminer: extraction and mining of academic social networks[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008: 990-998.
[32] FREEMAN L C. Centrality in Social Networks Conceptual Clarification[J]. Social Networks, 1978, 1(79): 215-239.
[33] PIANTADOSI J, HOWLETT P, BOLAND J. Matching the grade correlation coefficient using a copula with maximum disorder[J]. Journal of Industrial and Management Optimization, 2007, 3(2): 305-312.
[34] JÄRVELIN K, KEKÄLÄINEN J. Cumulated gain-based evaluation of IR techniques[J]. ACM Transactions on Information Systems (TOIS), 2002, 20(4): 422-446.
[35] ROBERTSON S E, KANOULAS E, YILMAZ E. Extending average precision to graded relevance judgments[C]// Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2010: 603-610.
[36] ZHANG Y, MA J, WANG Z, et al. Collective topical PageRank: a model to evaluate the topic-dependent academic impact of scientific papers[J]. Scientometrics, 2018, 114: 1345-1372.
[37] TAO M, YANG X, GU G, et al. Paper recommend based on LDA and PageRank[C]// Artificial Intelligence and Security: 6th International Conference (ICAIS 2020). Part III 6. Springer Singapore, 2020: 571-584.
[38] DAYEH M A, SARTAWI B, SALAH S. A Bias-Free Time-Aware PageRank Algorithm for Paper Ranking in Dynamic Citation Networks[J]. Intelligent Information Management, 2022, 14(2): 53-70.

CHEN Pan, born in 1997, postgraduate. His main research interests include data mining and academic big data.

CHEN Hongmei, born in 1971, Ph. D, professor, Ph.D supervisor, is a member of CCF (No. 19214M). Her main research interests include intelligent information processing, pattern recognition, etc.

（责任编辑：柯颖）