Journal of Computer Applications
计算机应用,2021,41(5):1356 - 1360
ISSN 1001-9081
CODEN JYIIDU
2021-05-10
http://www.joca.cn

文章编号:1001-9081(2021)05-1356-05
DOI:10.11772/j.issn.1001-9081.2020081206

基于Web of Science的PageRank人才挖掘算法

李　翀[1],王宇宸[1,2*],杜伟静[1,2],何晓涛[1],刘学敏[1],张士波[1],李树仁[1]
(1. 中国科学院 计算机网络信息中心,北京 100190; 2. 中国科学院大学,北京 100049)
(*通信作者电子邮箱 wangyuchen@cnic.cn)

摘　要:高水平论文是优秀科技人才的标志性成果之一。聚焦“Web Of Science(WOS)”热点研究学科,在构建学术论文语义Neo4j网络图和挖掘出活跃科研社区基础上,利用PageRank人才挖掘算法实现对科研社区中优秀科研人才的挖掘。首先,对现有的人才挖掘算法进行详细研究和分析;其次,结合WOS论文数据对PageRank人才挖掘算法进行了优化设计和实现,加入了论文发表的时间因子、作者署名排序递减模型、周围作者节点对当前节点的影响、论文被引用量等多维度考量因素。最后,基于热点学科计算机科学某社区近五年的论文数据进行了实验和验证。结果表明,基于社区的挖掘更具有针对性,能够快速定位各学科代表性优秀和潜在人才,且改进后的算法对人才的发现更加客观有效。

关键词: Web Of Science;Neo4j图数据库;PageRank算法;人才挖掘
中图分类号:TP391　　**文献标志码:**A

PageRank-based talent mining algorithm based on Web of Science

LI Chong[1], WANG Yuchen[1,2*], DU Weijing[1,2], HE Xiaotao[1], LIU Xuemin[1], ZHANG Shibo[1], LI Shuren[1]
(1. *Computer Network Information Center, Chinese Academy of Sciences, Beijing* 100190, *China*;
2. *University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: The high-level paper is one of the symbolic achievements of excellent scientific talents. Focusing on the “Web of Science (WOS)” hot research disciplines, on the basis of constructing the Neo4j semantic network graph of academic papers and mining active scientific research communities, the PageRank-based talent mining algorithm was used to realize the mining of outstanding scientific research talents in the scientific research communities. Firstly, the existing talent mining algorithms were studied and analyzed in detail. Secondly, combined with the WOS data, the PageRank-based talent mining algorithm was optimized and implemented by adding consideration factors such as the paper publication time factor, the author's order descending model, the influence of surrounding author nodes on this node, the number of citations of the paper. Finally, experiments and verifications were carried out based on the paper data of the communities of the hot discipline computer science in the past five years. The results show that community-based mining is more targeted, and can quickly find representative excellent and potential talents in various disciplines, and the improved algorithm is more effective and objective.

Key words: Web Of Science (WOS); Neo4j graph database; PageRank algorithm; talent mining

0　引言

科研论文是科研人员重要成果之一,高水平科研论文既可以反映作者的科研水平,一定程度也能反映出研究热点变化及国家科研投入变化情况。因此,基于时间序列对科研论文进行热点学科、科研社区、合著网络、人才发现研究非常有意义。人才作为重大科技成果、科技发展和社会进步的主体和源动力,挖掘优秀人才、培养和发现潜在人才尤为重要。

目前有较多对优秀科研人才挖掘的研究,并取得了一定的成效,不论是整体数据挖掘范围、挖掘精度方面,还是对科研人员学术能力评价方面,都取得了不错的效果。如冯岭等[1]从专利数据中抽取发明人的各个特征构建多层感知机模型,从而发现技术创新人才。江艳萍等[2]基于文献计量方法对全球潜力华人青年学者进行发现与评价,通过制定相应的检索策略获取数据集,从数据集中提炼出学者信息,利用筛选指标体系和综合评价指标体系确定潜力候选人,最后与同学科领域的标杆人物进行比较分析,明确潜力候选人的科研水平和学术定位。王孟頔等[3]利用Hadoop计算平台,通过网页数据提取分析关键词,根据关联规则算法挖掘出关联关键词,采用基于相似项的策略推荐人才。

上述人才挖掘分析算法,在人才发现和学者评价角度都取得了较好的进展,但也存在一定的不足之处。首先在科研

收稿日期:2020-08-10;**修回日期:**2020-10-30;**录用日期:**2020-11-13。　**基金项目:**中国科学院“十三五”信息化专项(XXH13504-03)。
作者简介:李翀(1978—),男,安徽霍邱人,高级工程师,博士,CCF高级会员,主要研究方向:大数据、推荐系统; 王宇宸(1996—),男,安徽怀远人,硕士研究生,主要研究方向:大数据管理; 杜伟静(1993—),女,河北廊坊人,硕士研究生,主要研究方向:大数据管理; 何晓涛(1971—),女,河北衡水人,高级工程师,硕士,主要研究方向:数据挖掘; 刘学敏(1975—),男,山东烟台人,高级工程师,硕士,主要研究方向:大数据、云计算; 张士波(1986—),男,山东聊城人,硕士,主要研究方向:大数据; 李树仁(1972—),男,安徽亳州人,高级工程师,博士,主要研究方向:数据挖掘。

成果数据的选取上缺乏权威性，同时数据较为杂乱；其次在人才学术评价上需要与标杆学者进行对比，具有评价的片面性；最后在人才挖掘上多数算法都属于广泛挖掘，缺乏针对性，并且在计算上过于复杂，对计算能力要求较高。除此以外还存在学术评价上不具有时间序列特性、不能根据学者自身特点进行公平化评价等。

本文聚焦全球最大、覆盖学科最多的综合性学术资源 WOS（Web Of Science）中收录的中国科学院学术论文，在前期工作中，完成对热点学科的学术论文语义图谱构建，并采用 Louvain 社区发现算法（Community Detection）[4]对研究热点背后相近研究领域的活跃学术圈进行挖掘，使人才挖掘研究更具有针对性。本文主要工作基于前期研究成果，深入研究了相关人才挖掘算法，结合学术论文语义网络属性和优化后的 PageRank 人才发现算法进行了设计和实现。实验表明，基于科研社区使得人才发现更有针对性，能够快速定位不同学科方向代表性人才，改进后算法使得在对优秀人才挖掘、潜在人才发现更加精准。

1　相关工作

本章首先介绍关于人才挖掘领域的一些研究成果，然后介绍基于科研社区的人才挖掘算法研究并分析比较。

1.1　人才挖掘算法相关研究

在目前的人才发现算法研究中，大致可以分为两类：一类为利用学者相关特征进行模型训练的监督学习方法，另一类为通过合著网络形式进行预测的无监督学习方法。以冯岭等[1]研究成果为例，其工作主要是抽取了反映各个发明人技术创新实力的专利特征。抽取的发明人特征包括专利申请量、专利总被引用量、合作发明人数量、合作发明人的平均专利申请量、申请人维持的专利数量以及所申请专利的文本特征等；然后再通过神经网络模型进行训练与预测，并且在其实验中将神经网络模型与传统机器学习模型进行了对比，结果表明该实验取得了不错的效果。除此之外，随着近几年图神经网络与知识图谱领域的发展，也出现了一些新的思路。比如 Park 等[5-6]提出的基于图神经网络分析知识图谱中节点重要性的方法，利用网络拓扑结构信息与节点间谓词关系，结合每个节点的自身特征，通过图神经网络模型进行节点重要性的预测。这个思路可以应用到人才挖掘研究当中，但需要合适且权威的数据集用于模型训练。

通过合著网络方法进行人才发现的研究也有很多，比如谢瑞霞等[7]的研究是基于合著网络构建学者影响力评价指标。在其评价指标中，不仅考虑了学者自身论文的影响力，还通过合著网络中节点的介数中心度计算了学者的网络影响力，也就是该学者在网络中的重要性体现。在实验中，通过将两种影响力结合计算，也取得了不错的效果。

本文充分吸取前面提到的相关研究的成功经验，在合著网络的基础上，首先通过学者论文相关特征计算学者的初始评分，再结合 PageRank 在合著网络上的传递性计算最终的评分，从而综合考虑学者个人特征与合著网络特征的影响，达到人才挖掘的目的。

1.2　人才挖掘算法比较

在已挖掘的科研社区基础上，后续工作将利用社区网络中心性对科研社区中的优秀科研人才进行挖掘推荐。本节将对与此相关的 Degree Centrality、Closeness Centrality、PageRank 三个图算法进行深入研究，其关系及区别如图 1 所示。

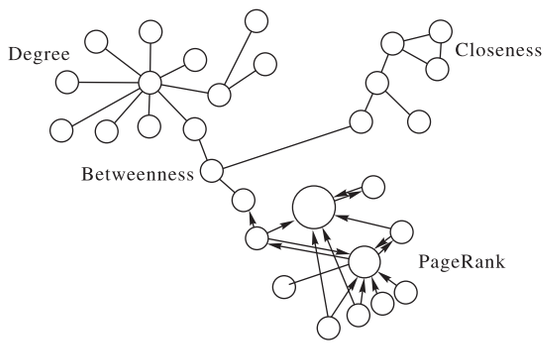

图 1　基于中心性的人才挖掘算法之间的比较

Fig. 1　Comparison between centrality-based talent mining algorithms

1.2.1　Degree Centrality 算法

Degree Centrality 算法可用于在没有方向的图谱中，利用度中心性去测量网络中节点间的相互关联关系程度，类似于关联关系矩阵，即表示当前节点与其他所有节点的直接联系总数[8]。但该种计算方式存在一定的弊端，如果社区中节点规模增大，则测量值均会增大，各节点的度中心性也会逐步增高。1994 年，Stanley Wasserman 和 Katherine Faust 针对该问题提出一个新的标准化测量公式，如式（1）所示：

$$C'_D(N_i)=\frac{C_D(N_i)}{g-1} \tag{1}$$

在对节点的度中心性进行衡量过程中，首先以本身节点 i 为初始阶段，测量出自身度中心性；其次测量出除本身节点外，其他 $g-1$ 个节点相连接的可能连接数，从而计算出与本身节点 i 相关联的其他节点的占比。最终比例范围为 0~1，0 表示节点 i 不与任何节点相关联，1 表示与所有节点都有关系。

Degree Centrality 用于计算来自节点的传入和传出关系的数量，并用于在图中查找流行节点[9]。基于以上分析，在适用性方面，如果试图通过查看传入和传出关系的数量来分析影响力，或者找到各个节点的“流行度”，可以使用 Degree Centrality 算法。

1.2.2　Closeness Centrality 算法

Closeness Centrality 依靠节点之间的距离判断节点间的近邻程度。首先计算本身节点 i 与网络中其他所有节点之间的距离，并进行相加求和，总值越小说明节点间可达且路径越短，即在空间上与其他各节点越接近，最终发现处于有利位置的节点，从而控制和获取组织内的重要信息和资源，具体应用如文献[10]。

为更明晰地表达该距离程度，Bavelas 于 1950 年将计算的近邻程度进行归一化定义，定义为近邻距离计算的倒数，最终的计算值取值范围限定在（0，1），越接近于 1 则节点的中心度越大，每个节点的具体计算公式如式（2）所示：

$$C(u)=\frac{1}{\sum_{v=1}^{n-1}d(u,v)} \tag{2}$$

其中：u 代表当前节点；n 代表图中节点的数量；$d(u,v)$ 代表节点 u 到节点 v 之间的最短距离。

Closeness Centrality 适用于筛选以最快速度传播信息的节点，其中使用加权关系对评估交流和行为分析中的交互速

度效果展示较为明显。该算法适用于连接图中的节点中心性计算，但当图中两个节点间没有路径时，计算该节点的所有距离之和会出现偏差，紧密度趋向于无限，最终影响整个图的中心性计算。

1.2.3 PageRank算法

PageRank算法初始用途是对网站网页重要性进行排序，以此来评判网页产生的影响力，具体计算如式(3)所示：

$$PR(u) = \sum_{v \in B_u} \frac{PR(V)}{L(v)} \tag{3}$$

其中：u为待评估页面。B_u为页面u的链入集合。对于页面u来说，每个入链页面自身影响力$PR(V)$与V页面的所有出链页面数量之比，作为页面V给页面u带来的影响力。这样可以将页面自身影响力平均分配至其每个出链上，再计算所有带给u页面的影响之和，便是网页u的影响力。

但式(3)存在一些问题，如一个节点没有出链或者入链，会出现等级泄漏或等级沉没现象，故提出了一种新的优化方式，加入阻尼系数d，如式(4)所示，这个阻尼系数代表用户通过跳转链接进入的概率，通常取值0.85。

$$PR(u) = \frac{1-d}{N} + d * \sum_{v \in B_u} \frac{PR(V)}{L(v)} \tag{4}$$

PageRank算法通过关联关系间的紧密程度来量化彼此间的影响力，通过出链入链的影响程度，最终确定最优影响能力的节点。PageRank算法更加适用于关系较多，且彼此影响力不均匀的关联状况。这与论文之间引用等关联关系相似，适用于挖掘关系复杂的图信息。PageRank算法还存在一些缺点，PageRank算法在使用过程中，过于注重当前数据特征，周围关联的节点会直接影响当前节点的影响力；除此以外，PageRank算法考量维度单一，对于出现较早的页面会因链接度较高而提升影响力，没有时间序列性。

综上几种对人才挖掘算法的分析，可以看出Degree Centrality主要是度量节点的出度与入度，说明当前节点的权威只受周围关联节点影响，应用于优秀科研人才挖掘上会具有单一性；另外，出入度计算上也存在大量重复计算，会导致计算效率较低。Closeness Centrality算法主要利用节点间的距离来计算中心性，如果存在没有相互关联的节点，会导致计算结果偏离正常值，应用于优秀人才挖掘上会导致挖掘结果不准确。PageRank算法是计算网页重要性排名的算法，主要利用链接关联性进行分析，在计算上将节点影响力进行均分，后进行统计分析来确定节点的重要性，这在一定程度上突出了重要节点的影响力，达到了较为公平的计算效果，应用于优秀人才挖掘上能对优秀人才赋予较大的影响力，从而突出其贡献度。综合比较分析，本文人才挖掘算法最终选择为PageRank算法。

2 PageRank算法优化与实现

PageRank算法的使用前提是需要有每位学者学术能力的初始评分，这能在一定程度上突出优秀人才的贡献度，但应用在学术论文的人才挖掘上也会存在一定的不足。首先不能根据时间连续性对人才进行筛选，随着时间的变化，优秀人才的科研方向和成果会发生变化，但PageRank算法不能动态地对科研能力进行调整；其次，PageRank算法评价维度单一，只是单一地考虑了关联节点的影响力，没有多维度评价因素，如论文被引用量、作者发文量等维度可以在一定程度上体现作者学术能力的强弱，提升优秀人才挖掘的准确性。为了解决该问题，达到更加准确的人才挖掘效果，有必要对PageRank算法进行了多维度优化。

经过调研，本文在实验中采用了Prathap于2010年提出的一种综合性评价学术成果指标，对学者的学术能力从学术论文数量以及引用次数进行评价。并通过结合常雨萧[11]的研究成果，为学术指标的计算加入时间因素、作者署名排序因素；在PageRank算法中加入了作者间余弦相似度作为影响系数。将优化后的算法应用在科研社区中，进行人才发现。

时间因素，作者署名排序因素以及学术指标$P(i)$的计算如式(5)~(7)所示。其中作者署名排序是采用了贡献度等级分配法[12]，并参考了科研成果评价研究成果[13]。论文发表的时间越早，在学术成果指标中的影响就越小；作者署名次序越靠后，该论文对于作者的影响力也越小。通过计算策略调整，使得近期活跃的学者可以得到更高的学术指标值，更有利于活跃人才的挖掘。

$$T = \exp(-\alpha(T_c - T_k)^2) \tag{5}$$

$$W(i,k) = \frac{a_k - i_k + 1}{\sum_{i_k=1}^{a_k} i_k} \tag{6}$$

$$P(i) = \left[\frac{C(i)^2}{N(i)}\right]^{\frac{1}{3}} = \left[\frac{\left[\sum_{k=1}^{n}\left(W(i,k)*c_k*T\right)\right]^2}{\sum_{k=1}^{n}(W(i,k)*T)}\right]^{1/3} \tag{7}$$

其中：α为尺度系数；T_c为当前时间，T_k为论文发表时间；a_k为论文k的作者总数，i_k为作者i在论文k中的位次，c_k为论文k的引用次数；$C(i)$为作者i的论文引用得分，$N(i)$为作者i的论文数目得分。

学者自身学术指标值的计算，见算法1。

算法1 Calculate Initial Score。

输入 待消歧作者的全部相关论文数据。其中：i表示作者；n表示论文篇数；a_k为论文k的作者总数；c_k为论文k的引用次数；i_k为作者i在论文k中的位次；T_c为当前时间；T_k为论文发表时间。

输出 学者i的自身学术指标值。

```
C(i) = 0, N(i) = 0
for each k ∈ [1, n] do
    index_sum = 0
    for each j ∈ [1, a_k] do
        index_sum = index_sum + j
    end for
    W_{i,k} = (a_k - i_k + 1)/index_sum
    T = exp( - α*(T_c - T_k)^2)
    C_i = C_i + W_{i,k}*c_k*T
    N_i = N_i + W_{i,k}*T
end for
P_i = (C_i^2/N_i)^{1/3}
return P_i
```

对于PageRank影响力传递过程，通过余弦相似度的方式计算作者节点间的关系。具体计算如式(8)、(9)所示，分别为作者间贡献影响程度和作者影响力得分。

$$Attr(i,j) = \cos[W(i,k), W(j,k)] = \frac{\sum_{k=1}^{n} W(i,k)*W(j,k)}{\sqrt{\sum_{k=1}^{n}(W(i,k))^2} * \sqrt{\sum_{k=1}^{n}(W(j,k))^2}} \tag{8}$$

$$Imp(i) = (1 - d) * P(i) + d * Imp(j) * Attr(i,j) \quad (9)$$

其中 d 为 PageRank 中的阻尼系数，一般取值为 0.85。最终的学者影响力评分由多轮迭代后的 $Imp(i)$ 得出。

PageRank 算法的 *Imp* 值计算，见算法 2。

算法 2　Modified PageRank Algorithm。

输入　所有作者的自身学术指标值为 *Imp*，所有作者间的贡献影响度为 *Attr*，每个作者的邻居节点为 *neighbors*，迭代轮次为 *n*。

输出　所有作者的最终评分列表。

```
1)  Initialize array Imp of authors
2)  for epoch ∈ [1, n] do
3)      for each author i ∈ authors do
4)          Imp[i] = P_i*(1 - d)
5)          neighbor_imp_sum = 0
6)          for each neighbor j ∈ neighbors[i] do
7)              neighbor_imp_sum += Imp[j]*Attr(i,j)
8)          end for
9)          Imp[i] = Imp[i] + d*neighbor_imp_sum
10)     end for
11) end for
12) return Imp
```

3　实验验证与分析

3.1　基础环境

操作系统为 CentOS 7 64 位，Kernel Linux 3.10.0。开发环境为 python3.7.3 + Neo4j 3.5.13；CPU 为 Intel Xeon Silver 4114 @2.20 GHz 40 核心；内存为 128 GB。

3.2　实验数据

实验数据为 1949 — 2019 年的 WOS 核心合集数据库中国科学院发表的 4 199 篇计算机科学学术论文数据，通过 Neo4j 创建论文语义网络图[14]，其中有作者 19 200 位，机构 26 232 个，生成 Workwith 关系数 15 799 个，其中实体类型为 Author（作者）、Paper（论文）、Org（作者所属机构）；实体间关系为 Belong to、Write、Workwith（Workwith 中包含属性 Weight）。如图 2 所示。

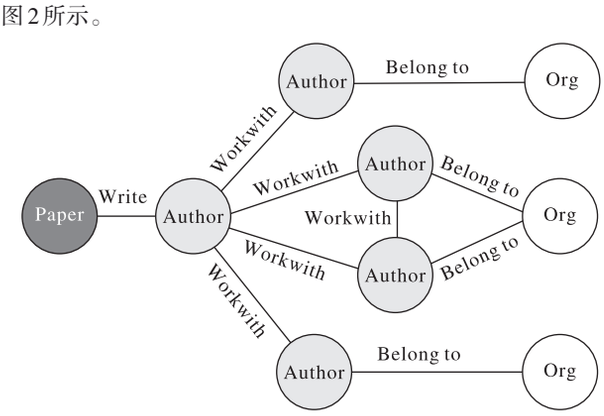

图 2　论文语义实体关系示意图

Fig. 2　Paper entity relationship diagram

在学术语义网络图基础上，应用 Louvain 社区发现算法对活跃科研社区进行挖掘[15]。通过使用模块度和模块度收益进行评价[16]，成功挖掘出模块度收益较高的前 10 个活跃科研社区，其分布如表 1 所示。

3.3　验证过程

本实验是在计算机科学领域挖掘出活跃度前 10 个科研社区基础上（见表 1）对活跃科研人才进行挖掘。

实验分为两个部分：一是根据式（4）采用优化前的 PageRank 算法对社区人才进行挖掘。在优化前的算法中，得分值计算只利用了语义图谱中作者节点间关系，而没有考虑作者节点自身特征。二是根据式（9）采用优化后的 PageRank 算法进行计算，综合考虑了作者自身节点的多个特征因素，并且作者间的关系也使用作者间贡献影响程度值进行了改进，使得不同邻居节点对中心节点的影响程度具有独特性。

表 1　社区人数及社区中论文数量表

Tab. 1　Number of communities and the number of papers in communities

社区 id	作者数量	合著论文数量	社区 id	作者数量	合著论文数量
141	151	30	529	104	20
229	155	26	67	79	19
17	109	23	29	64	16
128	86	23	18	52	13
28	98	22	300	56	13

本文以活跃度排名第一的 141 号社区进行的人才挖掘为例，优化前后的挖掘结果对比如表 2 和表 3 所示。

表 2　活跃人才排名表（优化前）

Tab. 2　Excellent talent ranking table (before optimization)

学者	得分	学者	得分
Yan	0.965 24	Du	0.581 52
Ma	0.537 61	Qu	0.512 48
Shi	0.498 24	Song	0.436 98
He	0.446 59	Ling	0.426 27

表 3　活跃人才排名表（优化后）

Tab. 3　Excellent talent ranking table (after optimization)

姓名	论文数	总引用数	作者位次均值	平均发表时间	得分	分数变化
Yan	55	459	2.32	2017	1.045 15	+0.080
Du	48	147	1.15	2017	0.661 85	+0.080
Ma	21	22	1.05	2018	0.638 96	+0.101
Shi	3	64	1.33	2018	0.605 24	+0.107
Qu	17	53	1.47	2018	0.582 46	+0.070
Song	25	104	1.04	2017	0.566 62	+0.130
He	39	212	1.23	2017	0.564 63	+0.118
Ling	28	77	1.07	2017	0.559 77	+0.134

3.4　结果分析

对于优化前后的两张表中的优秀人才挖掘结果，本文利用自然科学基金委项目数据以及人才个人信息对挖掘结果进行了验证分析，同时也对优化的效果进行了分析。

首先对挖掘结果的准确性进行分析，使用了较为权威的国家自然科学基金委员会项目数据对结果进行佐证。八位学者在自然科学基金委中的项目数据如图 3 所示。八位学者中有七位都在国家自然科学基金委中都承担有项目，其中有一位学者博士刚毕业尚无基金项目。另外，经查证八位均为领域内国家级或地方优秀人才，说明了优化改进后的学术成果指标和 PageRank 算法可以在人才挖掘方面较为准确。

其次对算法优化有效性进行分析，通过表 3 中的分数变化，可以看到受多个特征因素以及周边关联作者的得分变化的影响，八位学者的得分变化幅度不均。其中署名位次越靠前，论文发表时间越晚的学者得分增加幅度越大。以第四位与第五位学者为例，因为加入了署名顺序因素，在优化后排序发生了变化。这说明多个特征因素的加入会对学者的得分有

着不同幅度的影响，进而能使学者的最终得分更客观、科学。

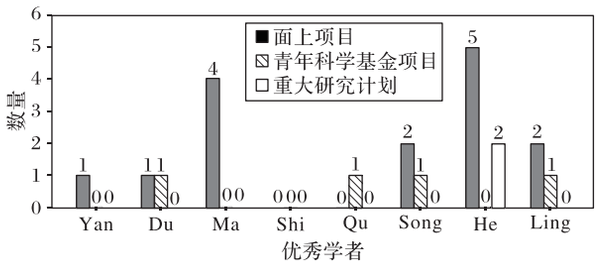

图3 国家自然科学基金委员会项目数据统计

Fig. 3 Statistics of projects of the National Natural Science Foundation of China

4 结语

本文基于WOS中收录的中国科学院学术论文数据，在构建学术论文语义网络图和Louvain科研社区发现结果的基础上，将人才挖掘范围聚焦于活跃科研学术圈，对PageRank人才挖掘算法加入论文发表时间因子、作者署名排序递减模型、周围作者节点对当前节点的影响因素、论文被引用量等指标进行算法优化，使得人才挖掘更加客观有效。实验结果表明，该算法具有一定的准确性和有效性，对优秀人才和潜在人才发现有一定的参考意义；同时也在一定程度证明了从高水平学术论文成果发现人才的可能性。

参考文献（References）

[1] 冯岭，谢世博，刘斌. 基于多层感知机的技术创新人才发现方法[J]. 计算机应用与软件，2019，36(7)：26-31，42.（FENG L，XIE S B，LIU B. A technical innovation talents discovery method based on multi-layer perceptron[J]. Computer Applications and Software，2019，36(7)：26-31，42.）

[2] 江艳萍，夏琬钧，赵颖梅，等. 基于文献计量方法的全球潜力华人青年学者发现与评价策略研究[J]. 情报杂志，2019，38(7)：178-183.（JIANG Y P，XIA W J，ZHAO Y M，et al. Discovery and evaluation strategy for the global potential Chinese young scholars：a research based on bibliometric method[J]. Journal of Intelligence，2019，38(7)：178-183.）

[3] 王孟頔，郜泳，薛安荣. 基于Hadoop平台的人才发现与推荐系统研究[J]. 软件导刊，2014，13(1)：4-6.（WANG M D，TAI Y，XUE A R. A discovery and recommendation system for talents based on Hadoop[J]. Software Guide，2014，13(1)：4-6.）

[4] ZENG J，YU H. A scalable distributed Louvain algorithm for large-scale graph community detection[C]// Proceedings of the 2018 IEEE International Conference on Cluster Computing. Piscataway：IEEE，2018：268-278.

[5] PARK N，KAN A，DONG X L，et al. Estimating node importance in knowledge graphs using graph neural networks[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York：ACM，2019：596-606.

[6] PARK N，KAN A，DONG X L，et al. MultiImport：inferring node importance in a knowledge graph from multiple input signals[C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York：ACM，2020：503-512.

[7] 谢瑞霞，李秀霞. 基于合著网络的作者影响力评价指标[J]. 情报理论与实践，2019，42(1)：100-104.（XIE R X，LI X X. Evaluation indices of author influence based on collaboration network[J]. Information Studies：Theory and Application，2019，42(1)：100-104.）

[8] FREEMAN L C. Centrality in social networks conceptual clarification[J]. Social Networks，1979，1(3)：215-239.

[9] BANGCHAROENSAP P，KOBAYASHI H，SHIMIZU N，et al. Two step graph-based semi-supervised learning for online auction fraud detection[C]// Proceedings of the 2015 Joint European Conference on Machine Learning and Knowledge Discovery in Databases，LNCS 9286. Cham：Springer，2015：165-179.

[10] BOUDIN F. A comparison of centrality measures for graph-based keyphrase extraction[C]// Proceedings of the 6th International Joint Conference on Natural Language Processing. Stroudsburg，PA：Association for Computational Linguistics，2013：834-838.

[11] 常雨萧. 复杂合著网络分析及可视化[D]. 重庆：重庆邮电大学，2017：31-40.（CHANG Y X. Complex co-authorship network analysis and visualization[D]. Chongqing：Chongqing University of Posts and Telecommunications，2017：31-40.）

[12] 娄策群. 社会科学评价的文献计量理论与方法[M]. 武汉：华中师范大学出版社，1999：10-14.（LOU C Q. Bibliometric Theory and Method of Social Science Evaluation[M]. Wuhan：Central China Normal University Press，1999：10-14.）

[13] 高伟. 基于InCites数据库的国家重点实验室科研成果评价研究[J]. 图书情报导刊，2018，3(2)：57-65，73.（GAO W. Analysis on scientific research achievements evaluation of state key laboratory based on InCites database[J]. Journal of Library and Information Science，2018，3(2)：57-65，73.）

[14] 张琳，熊斯攀. 基于Neo4j的社交网络平台设计与实现[J]. 情报探索，2018(8)：77-82.（ZHANG L，XIONG S P. Design and implementation of social network platform based on Neo4j[J]. Information Research，2018(8)：77-82.）

[15] DE MEO P，FERRARA E，FIUMARA G，et al. Generalized Louvain method for community detection in large networks[C]// Proceedings of the 11th International Conference on Intelligent Systems Design and Applications. Piscataway：IEEE，2011：88-93.

[16] CLAUSET A，NEWMAN M E J，MOORE C. Finding community structure in very large networks[J]. Physical Review，E，Statistical，Nonlinear，and Soft Matter Physics，2004，70(6 Pt 2)：No. 066111.

This work is partially supported by the CAS Informatization Special Project in the 13th Five-Year Plan (XXH13504-03).

LI Chong，born in 1978，Ph. D.，senior engineer. His research interests include big data，recommendation system.

WANG Yuchen，born in 1996，M. S. candidate. His research interests include big data management.

DU Weijing，born in 1993，M. S. candidate. Her research interests include big data management.

HE Xiaotao，born in 1971，M. S.，senior engineer. Her research interests include data mining.

LIU Xuemin，born in 1975，M. S.，senior engineer. His research interests include big data，cloud computing.

ZHANG Shibo，born in 1986，M. S. His research interests include big data.

LI Shuren，born in 1972，Ph. D.，senior engineer. His research interests include data mining.