科学家相关性测度典型算法比较与评析

吴振新[1,2] 单嵩岩[1,2]
(1. 中国科学院文献情报中心，北京 100190；2. 中国科学院大学图书情报与档案管理系，北京 100049)

摘要：调研科技文献作者相关性研究发展进展，对作者相关度算法进行系统化的分析和对比。从网络拓扑相似度算法入手，梳理和分析面向合作预测领域的作者相关度算法，分析和比较各种常用算法的优劣。对科技文献作者相关度算法进行系统梳理，分析重点方法的基本原理、优缺点并展望未来发展方向。

关键词：作者相关度；网络拓扑相似度；科研合作预测
中图分类号：G250
DOI：10.3772/j.issn.1673-2286.2019.03.002

20世纪末期，学术界发起了一系列旨在克服传统科学弊端的学术运动。这些运动凸显了“自由、开放、合作、共享”的理念，与传统科学文化的封闭性形成鲜明的对比，学术界将之称为开放科学运动[1]。开放科学是一个广义的概念，用于描述科学研究开展的方式，包括运用技术使研究活动更具协作性和开放性。开放科学环境为科研人员提供更多的知识获取途径，更为关键的是开放交流模型为科研人员提供更广泛地寻求潜在的科研合作对象/团体的可能，极大地促进了科研合作共享。因此，科研合作也成为开放科学环境中信息服务的一个重要内容。

为了更好地支持科研、服务科研，很多信息服务机构开始提供科研合作预测分析，并将其作为融入科研一线的智能知识服务的一项重要内容，科研合作关系预测的研究引起了更多的关注。作为科研预测研究的关键技术之一，科学家相关性计算随之得到越来越多的重视，取得不错的进展。但随着新技术方法的不断引入，该研究还在不断地改进和提升。

1 科研合作预测领域的作者相关度研究概述

科研合作预测通常在学术论文构建的科研合作网络中进行，目的是预测从未合作过的作者在未来产生合作的可能性。因此在合作网络中，对科学家相关度计算可转为作者相关度计算。作为社会网络的一种，科研合作网络体现了科学家间存在文章或者研究项目等的合作关系。

作者相关度在很多科研合作预测文章中也被称为相似度，在实际预测中，相比衡量不同作者间属性特征是否相似，更关注不同作者在合作网络中是否近邻、是否属于同一知识社区。如在合作网络中，拥有共同合作者但研究领域不同的两位作者，虽然属性特征相似度不高，但网络结构相似性高，作者相关性大。在科研合作预测领域中的作者相关度应用，主要根据作者节点属性及网络的结构特征等信息（如相关人际关系，研究方向、领域、内容、兴趣等计算作者间的相关度），以相关度表示作者未来合作的可能性。

对于目前的科研合作成因来说，两个作者可能合作是因为同处一个学术机构、互为师生关系、研究领域交叉等。而随着开放科学的发展，科学研究整个过程的开放性和互操作性不断增强，对从未合作过的作者在未来合作的预测会越来越复杂，但合作网络自身的拓扑结构优势能够揭示未来合作的可能性程度。如在合著网络中，两位拥有共同同事的作者；或在作者-关键词网络中，两位拥有共同关键词，研究内容相关的作者，就有合作的可能性。因此，作者相关性计算也就成为科研预测领域的关键技术之一。

科研合作预测在本质上是链路预测问题，通过已知的网络结构信息预测节点间未来产生连接的可能性，其中一类主流算法是基于节点相似性的方法。基于节点相似性的方法是根据已知网络中的作者节点拓扑结构，通过计算每一对未相连作者节点的结构相似度，相似度越高，其存在连边的概率越大，即作者未来合作的可能性更高[2]。

早期科研合作预测研究基于同构网络（合著网络、引文网络等），采用多种节点拓扑相似性指标，如基于共同邻居指标、到达路径指标和随机游走指标计算作者相关性。Liben-Nowell等[3]率先将基于网络拓扑结构的多种节点相似性指数应用于社交网络链接预测，并在合著网络中进行了实验。随后Zhou等[4]在包括合著网络在内的多种现实网络应用多种基于局部信息的指标实施链路预测，并提出资源分配（RA）指标和局部路径（LP）指标。近年来，越来越多的研究者采用相似度指标在合著网络中通过计算作者相关度来预测合作的可能性。张斌等[5]在7门学科的合作网络中应用多种相似性指标进行链路预测。张金柱等[6]运用多种相似度指标在合著网络中研究合作演化规律。

现实中，科研合作网络往往是异构的，同构网络节点相似性虽然易于计算，但丢失了很多语义信息。传统的节点相似性指标，根据同构信息网络设计，无法直接应用到异构信息网络中。为了计算异构网络中的节点相似性，Sun等[7]于2011年提出元路径的概念，并在异构书目网络中研究了合作关系预测问题，将基于路径指标、随机游走指标扩展到异构网络中。随后多种基于元路径的网络拓扑相似度指标相继被提出，伍转华[8]利用PathSim算法在DBLP文献数据集构成的“论文-作者-术语-会议”异构网络中寻找相关作者。Shi等[9]提出的HeteSim算法度量异质网络中任意节点对的相关性，在ACM（“机构-作者-论文-术语-学科-会议-出版物”异构网络）和DBLP数据集上计算作者节点相关度。孟晓峰[10]提出了一种基于元路径的新型相似性度量算法AvgSim，并在ACM数据集和DBLP数据集上计算作者节点相关度。张舒虹[11]在APS（“论文-作者-机构-术语-学科-期刊-年刊”异构网络）和DBLP数据集上，基于时间动态的路径数、传递相似性的归一化路径数和作者属性的对称随机游走计算作者节点间的相关性。

由于传统链路预测方法使用的网络拓扑相似性指标普遍存在计算效率较低和数据稀疏造成的维数过高问题，很难应用于大规模数据集的科研合作网络的合作预测。随着表示学习的不断发展，新兴的网络表示学习方法能够将图中的节点表示成向量，通过计算向量相似度获得节点相似度。该方法可以高效地计算网络中节点间的语义联系，也能够解决数据稀疏下的语义关联抽取和计算复杂问题[12]，因此学者们尝试将新方法应用于合作预测。Tang等[13]提出了LINE算法并在合著网络中进行了实验，在识别相关作者中取得了良好的效果。张金柱等[12]利用LINE网络表示学习方法得到作者的向量表示；通过向量夹角余弦值计算作者间的语义相似度。姚锐[14]构建“论文-期刊-作者”的异构网络，以作者为中心，结合元路径应用Node2vec模型得到作者的向量表示，根据明可夫斯基距离、余弦值计算作者间的向量相似度。Dong等[15]提出了metapath2vec表示学习方法，并在“作者-论文-会议”异构网络中进行了相关作者聚类实验。

2 面向科研合作预测的作者相关度算法分析和比较

利用学术论文构建的科研合作网络主要有同构网络（如合著网络[3]）和异构网络（如“作者-关键词”网络[16]、“作者-文献”网络[17]、“作者-文献-术语-会议”网络[18]）。基于节点相似性的方法在科研合作网络中进行合作预测，根据作者节点的拓扑信息，利用合著、引用、同属一个机构等连边的语义信息计算作者间的相关性，即利用拓扑相似度算法计算作者网络信息的相似程度。

2.1 基于同构网络节点相似性指标的作者相关度计算

基于网络拓扑结构相似度衡量作者间的相关度，是将作者实体间的关系连结起来构成网络图，利用图中节点间的连接属性，来判定两个作者的相关性。

衡量同构网络（合著网络）中作者节点的相关性，一般采用节点拓扑相似性指标来计算。相似性指标包含基于邻居的度量（网络局部结构的相似性）、基于路径的度量（准局部结构的相似性）、基于随机游走的度量（网络全局结构的相似性）。这里的“相似性”是相关文献已成习惯的术语，实际上很多相似性指标衡量的并非是节点对是否具有相似的特征，而是衡量节点对在几何或者拓扑空间是否邻近，或者在功能上是否具有较大的关联[19]，因此也被称为“接近性”或“相关

性”。其中最简单的相似性指标是共同邻居（Common Neighbors，CN），两个节点如果有更多的共同邻居就可能更相似。基于路径度量的相似性算法考虑到使用共同邻居指标进行计算时，相似性分数可能分布过于集中，使得预测结果没有区分度。因此，将两个节点的共同邻居扩展到“n阶共同邻居”[5]。基于随机游走的度量是利用一个节点到其邻居的转移概率来描述当前节点随机游走的目的地，可以根据整个网络图的信息来计算节点相似度，即使两个节点之间没有公共邻居节点也能计算（见表1）。

表1 代表性节点拓扑相似度指标

类 别	指 标	内 容	优 点	缺 点
基于邻居的度量	CN指标	直接计算两个实体相同邻居节点的个数来获得实体对的结构相似性	简单、直接	不考虑邻居的不同权重；参数估计问题
	Jaccard相关系数	实体对共同邻居集合的交集与并集的比	简单	不考虑邻居的不同权重
	Adamic/Adar指标	存在关联关系越多的实体其作为邻居在计算中所分配权重越低	考虑到不同的邻居权重，以获得更准确的结果	更高的计算复杂性
	RA指标	网络中没有直接相连的两个节点（x和y），x通过共同邻居传递资源到y，y可以接收到的资源数为节点x和节点y的相似度	在平均度大的网络表现良好；考虑到了三阶邻居	计算复杂性高
	优先链接（PA）指标	网络中度数大的两个节点更容易产生连接	简单	精确度较低
基于路径的度量	LP指标	利用具有长度为2和3的不同路径数量的信息，来表征节点之间的相似性	简单	在平均路径大于三阶路径的网络中不够精确
	Katz指标	考虑实体对之间的最短连接距离，如果两个实体之间由更多更短的关系路径所连接，则它们更相似	考虑实体之间的各种关系；有效的结构相似性匹配方法	参数估计问题；较高的计算复杂性
基于随机游走的度量	SimRank	使用图的拓扑信息来度量两个对象之间的相似度；如果两个对象被类似对象引用，则它们是相似的	考虑对象之间相互作用对结构相似度计算的影响	大数据集效率低，可扩展性差
	到达时间（可拓展为往返时间）	从节点a随机游走到节点b需要步数的期望值（计算从节点a到b，以及从b到a的期望步数）	简单	受终止节点和力大小的影响；对远离源点的拓扑噪声敏感
	RootedPageRank	从节点a随机游走到节点b，当到达b时，以概率α跳回a，以1~α继续随机游走，并记录下经过b的次数	不受节点影响力大小影响	计算复杂性高

拓扑相似性指标只涉及网络的结构信息，相似性指标计算比较简单，但不同指标在不同网络中的预测能力不一致，其预测的精确度取决于对网络结构特征刻画的好坏[20]。在高凝聚性的网络中，基于邻居和路径的相似性指标表现良好；在稀疏网络中，基于随机游走的度量预测效果比较好。在合著网络中识别作者相关度，基于邻居和路径的相似性指标表现良好，尤其是CN指标、Adamic/Adar指标、RA指标和Katz指标。

合作关系所形成的合著网络是一个熟人网络，日常生活中往往通过他人介绍或者更间接推荐来认识某个人进而与其合作。CN指标能很好地衡量两位作者的直接合作者，Katz指标和LP指标能很好地衡量两位作者的间接合作者。但是随着路径的增加，越间接的合著者对产生合著关系的影响越小，因此随机游走指标在合著网络中表现不理想。Adamic/Adar指标、RA指标是改进指标，赋予度数小的共同邻居节点更大的权重，比共同邻居指标取得了更好的效果，因为度数小的作者选择的合作者与其相关性更高。而Jaccard相关系数不考虑邻居权重因此表现一般。PA指标表示度数大的节点更容易产生连接，在合著网络中往往取得的效果不好，因为两位度数大的作者即影响力大的作者通常合作概率小[3,5,6,20]。

2.2 基于异构网络的元路径拓扑相似度指标的作者相关度计算

科研合作网络通常是异构的，即网络中存在多种类型的节点或连边。同构网络只是异构网络的投影，如

合著网络是由“文献-作者”网络投影形成的，虽然合著网络易于计算分析，但失去了原异构科研合作网络中丰富的语义信息。近年来，学者通过异构网络来解决科研合作预测问题，常见的方法包括基于元路径。

元路径是定义在网络模式上，用于描述异构网络中组合关系的路径。不同的元路径具有不同的语义来描述节点之间的相似程度。通过考虑依据不同元路径的路径，可以将同构网络中基于邻居和路径的属性拓展到异构信息网络中。例如，如果区别看待不同类型的邻居节点，并且把一阶邻居扩展为n阶邻居（某一节点和它的邻居之间的距离为n），则两个作者间共同邻居属性就变成两个作者之间依据不同元路径的路径数目[14]。

基于元路径的相似性计算首先用元路径定义两个节点之间的拓扑结构，然后在具体的拓扑上定义不同的度量标准。该方法考虑异构信息网络中不同拓扑结构的丰富语义信息和形成原因来进行计算。如包含作者（A）、论文（P）、出版物（V）3种节点的合作异构网络，两个作者节点间的元路径有：A1-P1-V1-P2-A2代表A1和A2在同一出版物上发表过文章；A1-P1→P2-A2代表A1的论文P1引用了作者A2的论文P2。

在元路径相似度指标中（见表2），以路径数和随机游走为基础的相似性度量适用于具有高出入度的对象，基于成对的随机游走的相似性度量适用于集中的对象（即大部分的链接属于小部分节点）[8]。

表2 代表性元路径相似度指标

类 别	指 标	内 容	优 点	缺 点
基于路径的度量	路径数	基于某一给定的元路径，计算两个节点之间的路径实例数量	计算复杂性较低	度量结果未标准化
	归一化路径数	对网络中两个节点之间存在的路径数标准化，分子为节点a到节点b的路径实例数量与节点b到节点a的逆关系路径实例数量之和，分母为以a为起始节点的所有路径总数与以b为终止节点的所有路径总数之和	具有较高的准确率	计算有一定复杂性
	PathSim	对路径数进行规则化的方法，分子为沿着元路径P，节点a到节点b的路径实例数量的2倍；分母为连接节点a、节点b自身的路径实例数量的和	具有对称性；考虑路径权重	两个节点对象必须属于同一类型；无法应用于非对称元路径；计算复杂性较高
基于随机游走的度量	随机游走	沿着元路径上的一个出发节点，随机地选择一个邻居节点，移动到邻居节点上，然后把当前节点作为出发点，重复以上过程	可度量不同类型节点的相似性	不具有对称性；计算复杂性较高
	对称随机游走	沿着元路径的两个方向进行随机游走	具有对称性	计算复杂性较高
	成对随机游走	元路径分解为2条相同长度的短元路径，从节点a和节点b开始，随机游走直到同样的中间节点的概率	具有对称性	无法应用于奇数元路径；计算复杂性较高
	HeteSim	将元路径P分成2条等长的路径P1、P2，之后从节点a和节点b出发分别沿着路径P1、P2进行随机游走，最后将两者游走到同一中间节点的概率作为a和b的相似性	可度量任意相同或不同类型的节点；具有对称特性	计算复杂性高；无法应用于大规模的网络
	AvgSim	节点对间的相似度是源节点在给定元路径下到目标节点的可达概率和目标节点在逆向元路径下到源节点的可达概率的平均值	可度量任意相同或不同类型的节点；具有对称特性；具有较高的效率和准确率	计算复杂性较高

在科研合作异构网络中，连接两个作者之间的元路径越多，两者越相关，元路径相似度指标均能取得不错的效果，其中归一化路径数指标表现更突出。PathSim指标更倾向于发现对等作者，如领域和声誉类似的作者。对称随机游走更倾向于高出入度的作者节点，表示在网络中越容易相互到达的作者更相关，如合著论文数越多的两位作者越相关。HeteSim指标思想为与相关对象相连的对象是相关的，如相关的作者会在相关会议中发表论文，它能够有效地度量作者相关度，但计算复杂性高也无法处理大规模网络。AvgSim指标

以HeteSim指标为切入点，能够有效地度量作者相关度，同时降低了计算复杂度[7,8,10]。

表示两位作者拥有共同合作者、在同一出版物上发表论文、研究相关领域和引用相同论文的元路径，这些都在识别作者相关度中发挥了重要作用。虽然越长的元路径携带信息越多，但随着元路径长度的增加，算法的复杂性也在增长，其精度增长幅度不大，因此长度一般控制在6个节点以内。

2.3 基于新兴网络表示学习方法的作者相关度计算

除在科研合作网络中采用结构相似性指标计算作者节点相关度外，随着表示学习的发展，基于深度学习的网络表示学习方法也得到了广泛的应用。网络表示学习方法将图中的节点表示成低维、实值、稠密的向量形式，通过计算向量间的距离判断节点的相关度。

基于神经语言模型的网络表示学习是目前的研究热点（见表3），其基本原理和思路来源于代表性的词向量生成工具Word2Vec[21]。Word2Vec工具包含CBOW模型和Skip-gram模型，通过选取输入词的前后n个词作为上下文，学习到包含语义信息的输入词的向量表示。针对网络结构和神经语言模型的特点，网络表示学习把节点类比为词，把在网络中获得的节点序列类比为句子，将节点序列作为Word2Vec的输入，根据每个节点的上下文信息，得到节点的向量表示。根据节点序列获取方式的不同形成了以DeepWalk[22]、LINE[13]、Node2vec[23]、Metapath2Vec[15]等为代表的基于神经语言模型的网络表示学习方法。

在科研合作网络中利用网络表示学习方法预测科研合作，学习作者在网络中的上下文语境信息，得到每位作者的向量表示，将合作预测变为作者向量相似度计算问题，相似度越高，尚未合作过的作者越有可能进行合作。

表3 基于神经语言模型的网络表示学习代表性算法

类 别	算 法	内 容	优 点	缺 点
基于随机游走	DeepWalk	通过构造节点在网络上的随机游走路径，模仿文本生成的过程，提供一个节点序列作为Skip-gram模型的输入从而得到节点的向量表示	具有可扩展性和可并行性；在信息缺失、标签数据稀疏及训练数据较小的时候也有良好的表现	只考虑一阶近邻；随机游走策略完全随机
	Node2vec	在DeepWalk的基础上引入偏向的随机游走策略，结合宽度优先搜索与广度优先搜索风格的邻域探索，生成节点序列作为Skip-gram模型的输入，从而得到节点的向量表示	考虑网络结构中的结构等价性与同质性；具有可扩展性和抗干扰性	具有参数敏感性
	Metapath2Vec	DeepWalk的扩展，使用基于元路径的随机游走来捕获不同类型节点之间的关系，获得的节点序列输入Skip-gram模型得到节点向量	能够捕获不同节点和关系的语义相关性；具有可扩展性	只能按照给定元路径模式游走；具有参数敏感性
基于非随机游走	LINE	考虑二阶邻居，采用广度优先搜索策略获得节点序列，输入Skip-gram模型生成节点向量	适用于大规模网络；计算速度较快；具有可扩展性	具有参数敏感性

网络表示学习为复杂网络分析提供了新的视角，部分研究者开始初步探索将其应用到科研合作网络。在合著网络中，DeepWalk、LINE、Node2vec都能取得不错的效果，其中Node2vec因为更灵活地选取邻居节点，同时考虑了合著网络结构中的结构等价性与同质性，在计算作者相关性方面表现得更好，但是能解决的网络规模不如LINE。LINE更适合稠密大规模网络，能够在具有高度数节点的合著网络中有效识别相关作者。DeepWalk更适合稀疏的网络，但提出时间早，完全随机的随机游走策略在竞争力方面不如之后提出来的改进算法。Metapath2Vec能够考虑不同类型节点间的不同语义，在科研合作异构网络中计算作者相关度方面取得良好的效果[13,15,23]。网络表示学习能在大规模数据集中自动提取合作网络中作者关联语义，在计算作者相关度方面有广阔的研究应用空间。

3 结语

在科研合作预测领域，作者相关度计算方法的研究发展紧跟新兴技术发展步伐。通过科研合作网络结构信

息判断作者相关度，经历了从同构网络到异构网络的发展，在越来越复杂的研究中不断的精细化、精准化。

从上述研究不难发现网络表示学习方法将在作者相关度计算中得到进一步应用。随着词向量在文本相似度计算上的成功，涌现出一批借鉴语言模型完成的网络图表示学习的方法已在合作网络中尝试应用，那么其他基于深度学习的网络表示学习方法能否有更好的表现，以及网络中其他结构的表示（如子图向量、图向量）能否应用到作者相关度计算将成为今后探索的方向。此外，构建科技知识图谱能够为作者相关度计算提供更多支持。与简单的科研合作网络（如合著网络、二分网络、三种节点网络等）相比，构建拥有更全面的作者及相关实体节点、更丰富的作者语义信息的科技知识图谱，能够更全面地比较作者间相关度，因此在知识图谱中寻找相关作者也将有更多应用场景。

开放科学给科研合作领域带来了挑战，也带来了机遇。作者相关度计算作为基础研究问题已经取得诸多成果，随着新兴技术与作者相关度研究不断交叉融合，该研究成果势必会进一步推动科研合作预测领域的发展。

参考文献

[1] 彭媛媛，陈雪飞. 认识“开放科学”[EB/OL].（2016-12-01）[2019-01-01]. http://blog.sciencenet.cn/blog-1035376-1018085.html.

[2] 吕琳媛. 复杂网络链路预测 [J]. 电子科技大学学报，2010，39（5）：651-661.

[3] LIBEN-NOWELL D，KLEINBERG J. The link-prediction problem for social networks [J]. Journal of the American Society for Information Science and Technology，2007，58（7）：1019-1031.

[4] ZHOU T，LV L Y，ZHANG Y C. Predicting missing links via local information [J]. The European Physical Journal B，2009，71（4）：623-630.

[5] 张斌，李亚婷，戴怡清. 学科合作网络的链路挖掘与应用分析 [J]. 情报理论与实践，2018，41（9）：108-113.

[6] 张金柱，胡一鸣. 利用链路预测揭示合著网络演化机制 [J]. 情报科学，2017，35（7）：75-81.

[7] SUN Y Z，BARBER R，GUPTA M，et al. Co-author Relationship Prediction in Heterogeneous Bibliographic Networks [C] // International Conference on Advances in Social Networks Analysis and Mining：ASONAM 2011. Washington DC：IEEE，2011：121-128.

[8] 伍转华. 异构信息网络的相似性度量方法 [J]. 计算机与现代化，2016（3）：78-84.

[9] SHI C，KONG X，HUANG Y，et al. HeteSim：A general framework for relevance measure in heterogeneous networks [J]. IEEE Transactions on Knowledge and Data Engineering，2014，26（10）：2479-2492.

[10] 孟晓峰. 基于异质信息网络的相似性度量研究 [D]. 北京：北京邮电大学，2015.

[11] 张舒虹. 学术异构信息网络中的作者合作关系预测 [D]. 大连：大连理工大学，2016.

[12] 张金柱，于文倩，刘菁婕，等. 基于网络表示学习的科研合作预测研究 [J]. 情报学报，2018，37（2）：132-139.

[13] TANG J，QU M，WANG M Z，et al. LINE：Large-scale information network embedding [C] //Proceedings of the 24th International Conference on World Wide Web. Florence：ACM，2015.

[14] 姚锐. 采用Node2Vec模型对网络特征表示方法研究 [D]. 南京：南京大学，2018.

[15] DONG Y，CHAWLA N V，SWAMI A，et al. metapath2vec：Scalable Representation Learning for Heterogeneous Networks [C] //Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. Halifax，Canada：ACM，2017.

[16] 张金柱，韩涛，王小梅. 作者-关键词二分网络中的合著关系预测研究 [J]. 图书情报工作，2016，60（21）：74-80.

[17] 张金柱，王小梅，韩涛. 文献-作者二分网络中基于路径组合的合著关系预测研究 [J]. 现代图书情报技术，2016（10）：42-49.

[18] LUONG N T，NGUYEN T T，JUNG J J，et al. Discovering Co-author Relationship in Bibliographic Data Using Similarity Measures and Random Walk Model [C] //7th Asian Conference on Intelligent Information and Database Systems. Cham：Springer，2015，9011：127-136.

[19] 刘宏鲲，吕琳媛，周涛. 利用链路预测推断网络演化机制 [J]. 中国科学：物理学 力学 天文学，2011，41（7）：816-823.

[20] 张斌，马费成. 科学知识网络中的链路预测研究述评 [J]. 中国图书馆学报，2015，41（3）：99-113.

[21] MIKOLOV T，SUTSKEVER I，CHEN K，et al. Distributed Representations of Words and Phrases and Their Compositionality [C] //Proceedings of the 26th International Conference on Neural Information Processing Systems，

Nevada：Curran Associates Inc，2013（2）：3111-3119.

[22] PEROZZI B，AL-RFOU R，SKIENA S. Deepwalk：Online learning of social representations [C] //In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York：ACM Press，2014.

[23] GROVER A，LESKOVEC J. node2vec：Scalable feature learning for networks [C] //In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York：ACM Press，2016.

作者简介

吴振新，女，1968年生，研究馆员，博士生导师，研究方向：数字资源的组织、管理、长期保存以及重用，E-mail：wuzx@mail.las.ac.cn。
单嵩岩，女，1993年生，硕士研究生，研究方向：数字图书馆技术，E-mail：shansongyan@mail.las.ac.cn。

Comparison and Analysis of Typical Algorithms for Correlation Measurement of Scientists

WU ZhenXin[1,2] SHAN SongYan[1,2]
(1. National Science Library, Chinese Academy of Sciences, Beijing 100190, China; 2. Department of Library Information and Archive Sciences, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: This paper attempts to investigate the development of author relevance research, as well as systematically analyze and compare the author relevance algorithm. From the perspective of network topological similarity algorithm, this paper combs and analyzes the author relevance algorithm in the field of cooperative prediction, analyzes and compares the advantages and disadvantages of various commonly used algorithms. The study summarizes the author relevance algorithm, with analyzing the basic principles, advantages and disadvantages of the key methods and look forwards to the future development direction.

Keywords: Author Relevance; Topological Similarity; Research Collaboration Prediction

（收稿日期：2019-02-18）

书 讯

《汉语主题词表》

《汉语主题词表》自1980年问世以后，经1991年进行自然科学版修订，在我国图书情报界发挥了应有作用，曾经获得国家科学技术进步二等奖。为适应网络环境下知识组织与数据处理的需要，由中国科学技术信息研究所主持，并联合全国图书情报界相关机构，自2009年开始进行重新编制工作，拟分为工程技术卷、自然科学卷、生命科学卷、社会科学卷四大部分逐步完成。目前工程技术卷和自然科学卷已出版。

《汉语主题词表（工程技术卷）》共收录优选词19.6万条，非优选词16.4万条，等同率0.84，在体系结构、词汇术语、词间关系等方面进行了改进创新。《汉语主题词表（自然科学卷）》共收录专业术语12.4万条，包含数学、物理学、化学、天文学、测绘学、地球物理学、大气科学、地质学、海洋学、自然地理学等学科领域，收词系统、完整，语义关系丰富、严谨，每条词汇都有相应的学科分类号表现其专业属性，并与同义英文术语对应。同时，建立《汉语主题词表》网络服务系统，提供术语查询、文本主题分析、知识树辅助构建等服务。《汉语主题词表》可用于汉语文本分词、主题标引、语义关联、学科分类、知识导航和数据挖掘，是文本信息处理及检索系统开发人员不可或缺的工具。

《汉语主题词表（工程技术卷）》已于2014年由科学技术文献出版社出版，分为13个分册，总定价3 880元。《汉语主题词表（自然科学卷）》已于2018年5月由科学技术文献出版社出版，分为5个分册，总定价1 247元。两卷均可分册购买。