[文章编号]1000-1832(2019)02-0071-10
[DOI]10.16163/j.cnki.22-1123/n.2019.02.013

面向作者消歧和合作预测领域的作者相似度算法述评

单嵩岩[1,2],吴振新[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] 从文本相似度和结构相似度算法入手,对面向作者消歧和科研合作预测领域的作者相似度算法进行了研究。分析和比较了各种常用算法的优劣,以及目前的应用情况,并对作者相似度算法进行系统梳理与展望.

[关键词] 作者相似度;文本相似度;结构相似度;作者消歧;科研合作预测
[中图分类号] G 250 [学科代码] 870·10 [文献标志码] A

随着互联网的发展和大数据时代的到来,面对海量的文献信息,用户从广泛获取转为个人需求选择,向用户提供精准、个性化的智能知识服务成为图书情报领域的发展方向.

精准知识服务是面向用户个体,提供基于个体特征和个体兴趣的完全个性化信息服务.实施精准信息的精准发现和投递,一方面可依据用户自定义或自描述的特征及信息;另一方面可通过挖掘用户潜在需求主动提供所需信息.因此,能够定位用户的需求与兴趣的用户画像、兴趣图谱等相关技术成为了发展精准知识服务的重要方法.可以通过用户画像和兴趣图谱获取用户不同角度上的相似性,找到与目标用户兴趣、喜好、需求相似的用户群体,将群体中用户喜欢的信息精准推荐给目标用户.这些应用场景的关键技术之一就是作者相似度计算.尽管作者相似度研究目前已经取得了丰硕成果,但仍有很大的改进和提升空间.

中国科学院文献情报中心(NSLC)面向中国科学院"十三五"规划和"四个率先"的需求,努力建设智能知识服务体系来支撑中国科学院快速创新需求.作为智能知识服务体系中的关键内容,NSLC从多个角度开展了与精准服务相关的技术和方法研究.本文通过对作者相似性研究的调研分析,力图为后期的作者相似性计算模型设计实现打下一个比较好的基础.

1 作者相似度算法的研究现状与进展

我们于2018年初,在CNKI数据库、SpringerLINK电子期刊和Webof Science等核心数据库中,以"作者相似度(Author Similarity)""学者相似度(Scholar Similarity)"等作为检索词进行搜索并获取了相关文献,在梳理作者相似度算法研究中发现,大部分文献都可以归并到两个应用领域:作者消歧和科研合作预测领域.因此本文选择从这两个领域入手,梳理总结常用的作者相似度计算方法.

作者相似度是很多领域的基础研究问题,如作者消歧、合作预测、相关学者推荐[1]、学科知识结构探测[2]等领域,由于应用场景不同,其内涵有所区别,并没有统一的定义.目前对作者间的相似比较研究,不仅要关注属性信息的字符相似度,而且要比较作者间信息在深层语义上的相似度,即两者的接近程

[收稿日期] 2019-05-01
[基金项目] 国家自然科学基金青年基金资助项目(41801320);中国科学院文献情报能力建设专项项目(院1754-1).
[作者简介] 单嵩岩(1993—),女,硕士研究生,主要从事数字图书馆技术研究;吴振新(1968—),女,研究馆员,博士研究生导师,主要从事数字资源的组织、管理等研究.

度.例如,同名但研究不同领域的作者间的相似度就小于研究领域相同但名字不同的作者.

1.1 作者消歧领域的作者相似度计算

在文献知识库中,作者姓名往往存在歧义性,主要体现形式有两种:一是不同名字指代同一个人,产生这种情况的原因有姓名拼写的变形(如唐琰琪的外文拼写方式有 Tang Yanqi、Yanqi Tang、Tang Y Q、Y Q Tang 等)、拼写或者印刷错误、作者有笔名及曾用名等(如周树人有包括鲁迅在内的多个笔名);二是同一个名字指代不同的人,即重名问题,可能几千人使用同一个大众化的名字.作者消歧就是确定一个作者名指代的真实世界中的具体人物,旨在找到属于特定作者的所有出版成果.

现有的作者消歧方法大多通过机器学习的分类和聚类算法实现.基于分类的作者消歧方法是有监督学习方法,在对每个目标消歧前,需要包括对已标注的作者数据进行训练和学习,建立分类模型,再利用模型计算作者相似度,判断新出现的作者与已标注的作者是否是同一作者.常用的有监督消歧分类模型包括朴素贝叶斯模型和支持向量机模型.文献[3]采用贝叶斯概率模型和支持向量机两种监督学习方法,选取合作者名称、标题、出版物名称作为特征,度量作者姓名的相似度以及论文的相似度解决作者消歧的问题.文献[4]采用标题、机构名称、出版物名称、主题词作为特征,利用贝叶斯概率模型判断 MEDLINE 中不同文献记录之间的相似度对作者进行消歧.对于有监督的消歧方法,分类结果的准确度相对更高一些,但该方法只适用于小型数据库.面对大量的文献数据,很难人工标注足够多的训练数据,可伸缩性差,因此未能广泛应用.

基于聚类的作者消歧方法是无监督学习方法,通过提取作者属性特征,采用相似度衡量方法和聚类算法,将所有相似的、可能指向同一作者实体的作者聚为一类,得到的聚类簇就是消歧结果.无监督的聚类方法包括基于文本的方法和基于网络图的方法.文献[5]使用合作者、标题和出版物这 3 个特征,提出 K-way 谱聚类的无监督学习方法,在引用中消除作者歧义,通过实验验证了这些特征能实现作者消歧.文献[6]设计的 GHOST 算法通过建立合作者关系图,计算作者节点之间的路径长度和条数来判断作者相似度,最后采用仿射传播(affinity propagation)进行聚类.文献[7]提出了一种称为 GCLUSIM 的方法,该方法构建作者合著网络,使用图形结构聚类和文本相似性度量来解决模糊作者消歧问题.无监督的消歧方法不需要训练数据,适用于大型数据库,其伸缩性较好,应用范围更广.

在作者消歧领域中,无论采取分类还是聚类方法,都需要计算作者相似度,将相似的作者归为一类.

1.2 科研合作预测领域的作者相似度计算

随着研究问题的多样化和复杂化,多学科交叉融合解决问题的情况越来越普遍,而作者研究方向的细致化也使得不同领域间的作者合作日益增多.为了向作者在合作者的选择上提供建议和参考,合作关系预测的研究变得越来越重要.科研合作预测旨在预测尚未合作过的作者将来具有合作的可能性,主要根据作者间的社会关系,相关研究领域、主题、兴趣等计算作者间的相似度,用相似度衡量未来作者潜在的合作机会[8].

科研合作预测一般在科研合作网络中进行.科研合作网络是由科技文献的元数据关联构建而成的,根据网络中存在的实体和连边类型,可分为同构网络(如合著网络[9])和异构网络(如作者-论文网络[10]、作者-关键词网络[8]、作者-论文-术语和会议网络[11]),通过学者间的连边表现其在文章、研究项目中的合作关系.以应用范围较广的合著网络为例,节点是作者,边是合著关系,合著网络中的合作关系预测就是计算尚未产生连边的作者节点对之间产生连边的可能性.

科研合作预测在本质上是链路预测问题,主要采用基于相似性的方法和基于学习的方法.基于相似性的方法是根据作者节点属性信息和网络结构信息,通过文本相似性算法和结构相似性算法比较每一对无连边的作者节点间的相似度,越相似的 2 个节点越有可能产生连边,即两位作者未来更有可能合作[12];基于学习的方法是将合作预测看作二分类问题,即 2 个节点有连边(正类)或没有连边(负类).该方法也是根据已知网络中的作者节点属性和节点拓扑结构,通过无监督或有监督的机器学习算法(如分类器、概率模型等)来预测新作者节点对的连边属于正类或负类的概率[13].特别是采用分类器进行链路预测时,需要从网络中提取合适的特征,由节点、拓扑的相似性度量提供的特征属性在分类学习算法中得到广泛应用.

文献[9]率先利用多种节点相似性指标解决社交网络中链接预测问题,并在合作者网络中取得良好

结果. 文献[14]提出局部概率图模型，利用节点共现概率属性、拓扑属性和语义特征在学术网络中预测合作者关系. 文献[15]将链路预测问题看作二分类问题，在合著网络中使用有监督模型学习，将结构相似性指标作为特征，学习训练集中的链接信息，从而预测测试集中可能产生的链接.

基于相似性方法所采用的传统结构相似性指标，多应用于同构信息网络中，但这些基于邻居集合和节点之间路径的相似性指标并不能直接应用到异构信息网络中进行计算. 基于学习的方法则可以在多种网络中应用，但在特征选择和模型训练过程中也会引起高额的计算成本.

2　面向作者消歧和合作预测的作者相似度算法的分析与比较

作者相似度计算方法在作者消歧和合作预测领域都得到了不错的研究应用. 作者消歧领域多从文献中提取特征，将作者的所属机构、专业、研究领域等文本信息直接用于计算作者之间的相似性. 近年来也通过构建社会网络，利用图结构信息计算作者相似性. 合作预测通常在科研合作网络上开展研究，少量文献选取作者节点的属性信息，通过文本信息比较作者相似度；大多数文献选取作者节点的拓扑信息，利用包含合著、同属一个机构、同一出版物上发表论文等语义信息的连边比较作者间的相似性. 不难看出，两个领域在算法上有交叉重用，也各有侧重. 但总的来说，这两个领域所采用的方法基本上可以归为两大类：一类是通过属性信息利用文本相似度算法比较作者的相似度；另一类是在社会网络中利用结构相似度算法比较作者的相似度.

2.1　基于文本相似性的作者相似度计算方法

作者相似度计算通常依赖于作者的相关属性信息来判断作者的相似性. 如合著者、电子邮箱、从属机构等强特征，可以有效地计算出作者的相似程度，而标题、关键词、摘要、研究方向、出版物等弱特征的计算效果较弱. 这些属性信息一般采用文本相似性计算，主要分为基于字符串的方法和基于语料库的方法.

2.1.1　基于字符串的文本相似度计算

基于字符串方法从字符串匹配度出发，以字符串共现和重复程度为相似度的衡量标准[16]. 第一类方法单纯从字符或词语的组成考虑相似度算法，如 Jaccard 相似系数、余弦相似度、Tanimoto 系数、汉明距离；第二类方法衡量编辑操作，即一个字符串最少需要多少次编辑才能变成另一个字符串，如 Levenshtein 距离、Smith-Waterman 距离、affine gap 距离、Jaro-Winkler 相似度函数. 常用的基于字符串的文本相似性函数总结见表 1.

基于字符串的文本相似性算法，在计算作者姓名、机构、会议、期刊、关键词等信息的相似度上有着广泛的应用. 文献[17]采用 Jaccard 相似系数和 Levenshtein 距离计算所属机构、出版地的字符相似性；使用 Jaccard 相似系数和余弦值计算标题、摘要的相似性. 文献[11]采用 Jaccard 相似系数、Soergel 相似系数、Lorentzian 相似系数、汉明距离计算论文/会议/关键词的相似性. 文献[18]使用 Tanimoto 系数计算标题、摘要、合著者的姓氏和首字母缩写、参考文献、规范化作者关键词、规范化索引关键词、规范化研究地址、期刊名称的相似度.

基于字符串的方法实现起来简单、易于操作，但只比较了文本的拼写相似性，并未考虑文本的词义和语义. 以同义词为例，尽管词语写法不同，但意义相同，基于字符串的方法并不能识别出这类词语间存在着的相似性.

2.1.2　基于语料库的文本相似度计算

基于语料库的方法利用从语料库中获取的信息计算文本相似度[16]，进一步考虑了词语的语义. 基于语料库的方法主要分为基于词袋模型和基于神经网络 2 种方法. 词袋模型(Bag of Words Model，BOW)建立在分布模型(Distributional models)的基础上，即“相似的词会出现在同一文本区域”. 基本思想是将文档表示成词的集合，每个词用词频表示，构成文档向量，缺点是没有考虑文本序列. 通过神经网络模型生成词向量则是建立在分布式模型(Distributed models)的基础上，即“相似的词会出现在相似的语境里，但可能不会同时出现”. 其基本思想是考虑词语的上下文，将每个高维空间的词映射到低维空间形成一个固定长度的短向量.

表1 代表性基于字符串的文本相似性函数

类别	分类	内容	优点	缺点
基于字符的相似性函数	Jaccard相似系数	等于2个字符集合的交集个数与并集个数的比值	集合相交操作与顺序无关	具有错误敏感性
	余弦相似度	把字符串表示成n维的向量,通过计算向量夹角的余弦值来比较字符串,夹角越小越相似	与顺序无关;加入权重,准确性更高	具有错误敏感性
	Tanimoto系数	一种广义Jaccard相似度,常通过词语-词频向量比较文档相似度. Tonimoto系数将向量的长度也纳入考虑,长度差异越大相似性越小	与顺序无关	计算复杂度高
	汉明距离	2个等长字符串之间对应位置的不同字符的个数	简化长文本计算;效率高	不能比较不同长度的字符串
基于编辑距离的相似性函数	Levenshtein距离	将一个字符串转换成另一个所需的最少编辑操作(插入、删除和替换)次数	降低错误敏感性	不考虑不同字符或子串的重要性
	Smith-Waterman距离	执行局部序列比对,比较所有可能长度的片段,寻找最长公共子序列,其他短序列作为前后缀,在匹配时对公共子序列和前后缀给予不同权重	有效减少不同前缀和后缀的影响	更高的复杂性;有限的应用范围:只考虑前缀和后缀的影响
	Affine Gap距离	如待匹配的字符串中间存在连续多个空位,对空位加入惩罚,赋予较低的权重	有效减少字符串缩写的影响	计算复杂性更高;应用范围较少
	Jaro-Winkler相似性函数	通过比较2个字符串的公共部分来计算相似程度	容忍少量的拼写错误	主要用于英文姓名相似度;不适用于长文本

基于语料库计算文本相似度是一种与向量相似度计算相混合的方法,基于词袋模型和基于神经网络的方法均先将文本处理成向量表示,通过计算向量相似度衡量文本相似度. 梳理基于语料库的文本向量算法见表2.

表2 代表性基于语料库的文本向量算法

类别	分类	内容	优点	缺点
基于词袋模型	向量空间模型(VSM)	将文献表示成一个基于词频或者词频-逆文档频率(TF-IDF)权重的特征向量,通过计算特征向量相似度衡量文献相似度	有效地对文本进行表示	语义信息较少;特征向量维度较高,很高的稀疏性;计算效果一般
	潜在语义分析(LSA)	通过将文档的空间向量表示通过奇异值分解(SVD)进行降维,转换为文档的潜在语义表示	语义信息较丰富	计算复杂度比较高;可移植性差
	概率潜在语义分析(PLSA)	在LSA基础上引入主题层,采用期望最大化算法训练主题,得到改进的算法	加入了文档更深层的语义信息,结果更加准确	不适用于大规模文本
	潜在狄利克雷分布(LDA)	对文本进行主题建模,基于基普斯采样,得到文本的主题分布表示	语义程度较高;适用于大规模文本;具有鲁棒性	计算复杂度比较高
基于神经网络	词向量(Word Vector, Word Embeddings, Distributed Representation)	将未标记的非结构文本,通过训练把每个词映射成K维实数向量,可通过词之间的距离来判断它们之间的语义相似度	向量维数低;语义信息较丰富	计算复杂度比较高

基于语料库的文本相似度算法的引入,使得在比较作者属性信息时可以挖掘文本深层语义方面的信息,与字符串相似度算法相比,结果的准确性有所提高,在实践中已有一定的应用. 文献[19]提出了一种跨文档的人名对齐方法,利用VSM模型计算摘要间的相似度,将人名共指的文档聚类在一起. 文献[5]采用TF-IDF和标准词项频率(NTF)对文献引用中的合作者、题目和出版物进行特征表示,利用特征向量的余弦相似度表示文献相似性.

为了进一步提高作者相关词语义的利用程度,文献[20]提出基于名为LDAcosin的衡量内容相似性指标,论文之间的相似性越高,作者越相似. 使用论文的标题和摘要信息通过LDA模型生成每篇论文的表示向量. 通过计算论文向量相似度得到作者相似度. 文献[21]提出了使用PLSA以及LDA方法,利用文献内容生成作者主题向量,并采用欧几里得距离计算主题向量间的相似度.

通过神经网络模型生成词向量计算文本相似度的广泛研究也使得不少产生词向量的模型和工具也被提出，Word2Vec 工具[22]就是其中的典型代表. 文献[23]通过抽取文献的标题、关键词、摘要信息作为学者属性文本集合，综合 Word2Vec 词向量与词的 TF-IDF 值计算出学者向量，学者向量间相似度计算采用 Jensen-Shannon 距离进行衡量.

2.2　基于结构相似性的作者相似度计算

应用属性信息的确可以很好地比较作者间相似度，但是在很多情况下，无法轻易地获取这些信息，而且有些时候并不能保证获取信息的准确性. 与属性信息相比，获取作者及其他相关实体间关系，构建作者社会网络（如合著网络、作者-论文网络、作者-关键词网络等）更加容易，也更加可靠. 同时利用图中节点间的拓扑信息，来判定两个作者相似性的方法，对于结构相似的网络具有普适性. 因此，在合作预测领域中应用更广的是基于网络结构信息的结构相似性比较，近年来在作者消歧领域中基于社会网络的结构相似性研究也越来越多.

2.2.1　同构网络中节点拓扑相似度计算

早期在网络中衡量作者相似度的研究，大多选取在合著网络、引文网络等同构网络上计算节点的结构相似性. 结构相似性指比较信息网络中节点间连接属性的相似性，同构网络中的相似性指标可分为基于网络局部结构的相似性（基于邻居的度量）、准局部结构的相似性（基于路径的度量）、网络全局结构的相似性（基于随机游走的度量）[24]，见表 3.

表 3　代表性节点拓扑相似度指标

类别	函数	内容	优点	缺点
基于邻居的度量	共同邻居指标	2 个实体节点共同的邻居节点数量，共同邻居越多越相似	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	实体节点间共同邻居数与总邻居数的比，比值越高越相似	简单	不考虑邻居的不同权重
	Adamic-Adar 指标（AA）	出入度越少的共同邻居节点在计算中所分配权重越高	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
	资源分配指标（RA）	间接相连的 2 个节点 a 和 b，a 将资源通过共同邻居传递到 b，通过 b 接收到的资源数比较 2 个节点相似性	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
基于路径的度量	局部路径指标	利用节点间长度为 2 和 3 的路径数量，来表示节点之间的相似性	简单；考虑三阶邻居的贡献	在平均路径大于三阶路径的网络中不够精确
	Katz 指标	考虑实体对之间的所有路径，并赋予短路径较大的权重，如果 2 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；较高的计算复杂度
基于随机游走的度量	SimRank	基本假设为 2 个节点间的相似度正比于其入链邻节点的相似度	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；结构信息考虑不完整
	P-Rank	在 SimRank 的基础上，考虑了出链信息. 假设被相似节点引用或引用了相似节点的 2 个实体节点相似	同时考虑入链和出链的语义信息	大数据集效率低
	到达时间（HT）	从节点 a 随机游走到节点 b 需要步数的期望值	计算简单	受终点影响力大小的影响，对远离源点的拓扑噪声敏感
	Rooted PageRank	从节点 a 出发以概率 $1-\alpha$ 选择邻居随机游走，到达目标节点 b 后以概率 α 返回 a	不受节点影响力大小影响	计算复杂性高
	PropFlow（PF）	与从节点 a 随机游走（长度固定为 l）到 b 的概率成正比	计算速度快，可扩展性强，对远离源点的拓扑噪声不敏感	计算复杂性较高

最基础的相似性指标是共同邻居,2个节点的共同邻居越多就越可能相似,即在合著网络中有更多共同合作者的2个作者更相似.这种方法在集聚系数较高的网络中表现非常好,有时甚至超过一些更复杂的算法[12].基于路径思想的相似性算法考虑到使用共同邻居指标进行计算时,所获得的值很可能局限在0,1,2,相似性分数的分布过于集中,从而造成预测结果没有区分度.因此,将2个节点的共同邻居扩展到"n阶共同邻居",即考虑到2个节点间的n阶路径的数量为准局部路径指标[9].基于随机游走的思想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2个节点之间没有公共邻居节点也能计算[24].拓扑相似性指标只考虑了网络的结构信息,因此计算结果的准确性取决于指标的定义是否符合网络结构特征.如在集聚系数高的网络中,基于邻居和路径的度量方法能更准确地表示节点相似性;而在集聚系数低的网络中,更适合采用基于随机游走的度量方法.

Liben-Nowell和Kleinberg率先在社交网络的链接预测问题中应用基于结构的节点相似性指标,在合著网络中进行了实验.此后,通过构建同构网络,应用节点相似性指标衡量作者相似度的研究也越来越多.文献[9]在合著网络中系统地比较了几种节点拓扑相似性指数,包括图最短距离、共同邻居、优先连接(PA)、Adamic/Adar、Jaccard、SimRank、到达时间(HT)、rooted PageRank和Katz.其中,基于邻居度量的共同邻居指标和Adamic/Adar指标计算作者相似度表现良好;基于路径度量的Katz指标的表现良好.文献[25]选取7门学科构建合作网络,采用AUC评测指标,对多种相似性指标效果进行了比较,发现AA指标和Katz指标都是很有效的指标,并在图书馆情报文献学合作网络中应用AA指标和Katz指标计算作者相似度.文献[26-27]分别提出了在作者-关键词二分网络中运用SimRank和P-Rank指标的作者相似度计算方法,考虑了网络整体结构,得到了作者间以及词汇间的潜在关联关系,该算法的指导思想是关键词相似度越高,与其相连的作者相似度也越高.

2.2.2 异构网络中元路径拓扑相似度计算

现实世界中作者的社会网络往往是异构的,即网络中包含的节点或连边是不同种类的.合著网络等同构网络是将异构网络中的一种实体提取出来构建的网络,虽然计算简单但丢失了丰富的语义信息.近年来,学者们转而在异构网络中研究作者相似度.文献[28]提出的一种基于元路径的解决办法,能将基于同构网络的节点相似度指标扩展到异构网络,是目前应用较广的方法之一.

元路径是一条包含关系序列的路径,这些关系定义在不同类型的实体之间.根据不同元路径包含语义的不同,比较实体节点间的相似度.通过区分不同种类的邻居节点、依据不同的元路径,把一阶邻居扩展为n阶邻居,将2个节点间共同邻居属性转变为2个节点之间依据不同元路径的路径数目[29],这样就可以将同构网络中的节点相似度指标扩展到异构网络中.

基于元路径的节点相似度计算,首先根据需求指定2个节点间的元路径,然后在具体的路径上使用不同的相似度指标,代表性算法见表4.此算法通过计算异构网络中相关节点不同连边的丰富语义来比较相似性.与节点拓扑相似度指标一样,不同的元路径相似度指标适合不同结构特征的网络.具有高出入度节点的网络适合用以路径数和随机游走为基础的相似性指标,集中网络(即多数链接属于少数节点)适合用基于成对的随机游走的相似性度量[30].比起同构信息网络,异构信息网络中不同拓扑结构有着更丰富的语义信息,基于元路径相似性指标的作者相似度计算的实践研究也越来越多.文献[28]在异构书目网络中通过研究合作预测问题,验证提出的元路径概念,采用PC、NPC、RW、SRW指标计算元路径相似度,度量异构信息网络中节点的同级相似性.文献[31]提出了一种基于元路径的新型相似性算法HeteSim指标用来比较异构网络中任意2个节点间的相似性,在ACM和DBLP数据集上进行了验证.文献[32]在HeteSim指标的基础上提出了新型相似性指标AvgSim,与HeteSim指标相比降低算法的复杂性,并在ACM和DBLP数据集上进行了实验.文献[33]在APS和DBLP数据集上,采用路径数指标衡量具有时间动态的元路径相似度、标准化路径数衡量元路径的传递相似性以及对称随机游走衡量具有作者属性的元路径相似度,从而比较作者节点间的相似性.

表 4　代表性元路径相似度指标

类别	算法	内容	优点	缺点
基于路径的指标	路径数(PC)	给定元路径，计算连接 2 个节点的元路径数目	计算复杂性较低	度量结果未规则化
	标准化路径数(NPC)	给定元路径，考虑 2 个节点在网络中与其他节点的元路径数对路径数标准化. 即节点 a 到 b 的元路径数与 b 到 a 的路径数之和，比上从 a 开始的所有元路径数与以 b 结束的所有元路径数之和	具有较高的准确率	度量结果未规则化
	PathSim	给定对称的元路径 P，比较同类型节点间的相似度. 2 倍的节点 a 到 b 的元路径数，比上分别以节点 a 和 b 为开始和结束节点的元路径数之和	具有对称性；考虑路径权重	源和终节点只能属于同一种类；不适用于非对称元路径；计算复杂性较高
基于随机游走的指标	随机游走(RW)	给定元路径，以节点 a 为起始点，随机游走到邻居节点 b，然后把节点 b 作为起始点，重复以上过程	可度量不同类型的结点的相似性	不具有对称性；计算复杂性较高
	对称随机游走(SRW)	沿着元路径的 2 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(PRW)	将元路径 P 拆分成长度相同的短元路径 P_1、P_2，从节点 a 和 b 开始分别按照路径 P_1、P_2 随机游走，将两者游走到同样中间节点的概率作为 a 和 b 的相似性	具有对称性	无法应用于奇数长度元路径；计算复杂性较高
	HeteSim	在成对随机游走的基础上进行改进，能将奇数长度的元路径转变成偶数长度的元路径，使成对随机游走方法适用于任意长度元路径	可以度量不同类型的对象；具有对称特性	计算复杂性较高；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终止节点的可达概率和终止结点在反向元路径下到源结点的可达概率的平均值	可度量任意相同或不同类型的结点；具有对称特性；具有较高的效率和准确率	计算复杂性较高

2.2.3　基于新型网络表示学习的网络结构相似度计算

除了采用结构相似性指标计算网络拓扑相似度以外，随着表示学习的兴起，网络表示学习方法也逐渐应用于节点相似度计算. 网络表示学习方法是把网络中的节点语义信息映射成低维、稠密、实值向量，通过计算向量间的距离比较节点的相似性. 随着 Word2Vec 工具的成功，基于神经网络的网络表示学习方法应用更加广泛. Word2Vec 工具本质上是一种神经语言模型，包含了 CBOW 和 Skip-gram 模型，通过考虑当前词的上下文，学习包含语义信息的词向量[34]. 针对网络结构特点，借鉴 Word2Vec 工具的网络表示学习方法把节点看成自然语言中的单词，把在网络中随机游走生成的节点序列当作自然语言中句子. 依据获取节点序列的不同方式，形成了以 DeepWalk[35]、LINE[36] 和 Node2vec[37] 等为代表的基于节点位置信息的网络表示学习方法. 网络表示学习方法可以在有效地保证网络中节点的特征与相似性的基础上对网络进行有效的结构特征提取分析，可以解决目前网络研究中高度非线性、保留网络结构、网络高度稀疏的三大难点，从而达到更好的数据抽象效果，更加真实的还原模型. 由于网络表示学习得到的向量是多维连续的，因此梳理的相似度计算方法更偏重于连续向量(见表 5).

网络表示学习为在复杂网络中分析节点结构相似度提供了新的方法，科研人员开始尝试将其运用到作者社会网络，通过获得作者节点的向量表示，计算作者间的相似度. 文献[34]首先构建作者合著网络，利用 LINE 模型学习作者节点在合著网络中的上下文语境信息，得到作者的向量表示，采取余弦相似度计算作者向量间的相似度. 文献[29]构建包含期刊-论文-作者实体的学术异构网络，利用 Node2vec 模型获得作者的向量表示，根据余弦相似度计算他们之间的向量相似度.

表5 代表性向量相似度量函数

类别	算法	内容	优点	缺点
距离度量函数	曼哈顿距离	将多个维度上的距离进行求和后的结果,距离值越小,相似度越大	计算量较少,性能相对高;能衡量维度的数值差异	将向量各个分量的纲量同等看待;没有考虑各个分量的数字特征可能不同
	欧几里得距离	衡量的是多维空间中各个点之间的绝对距离,距离越小,相似度越大	对连续稠密的数据表现良好;计算简单;能衡量维度的数值差异	将向量各个分量的纲量同等看待;没有考虑各个分量的数字特征可能不同
	马哈拉诺比斯距离	表示数据的协方差距离	与量纲无关;排除变量之间的相关性干扰	夸大了变化微小的变量的作用
相似度度量函数	余弦相似度	多维空间两点与所设定的点形成夹角的余弦值;夹角越小,两点相距就越近,相似度就越大	适合高维度向量相似度计算;计算量较少,性能相对高;与量纲无关	只能分辨个体在方向上的差异,无法衡量每个维数值的差异
	Jaccard 相似度	常用于集合间计算相似度,衡量的是2个集合公共元素所占的比例	简单;计算效率高	无法衡量差异具体值的大小,只适用于二元数据的集合
	Tanimoto 系数	一种广义 Jaccard 相似度. Tonimoto 系数考虑了2个向量的长度差异,长度差异越大相似性越小	能衡量维度的数值差异	计算复杂度高;余弦相似度受到向量的平移影响
	皮尔森相关系数	将数据归一化(数据减去其对应均值)后进行余弦相似度计算;系数的绝对值越大表明相似性越强	适用于2个具有线性关系的连续向量,具有平移不变性和尺度不变性	计算复杂度高;结果受样本容量大小影响

3 结语

作者相似度计算方法的研究发展紧跟新兴技术发展步伐,在基于文本相似性的作者相似度算法方面,经历了从拼写比较到语义比较的发展;在基于结构相似性的作者相似度算法方面,经历了从同构网络到异构网络的发展. 随着研究的不断深入,作者相似度算法逐渐走向精细化、精准化. 从上述总结分析发现:

(1) 作者相似度研究将进一步应用表示学习方法. 在作者消歧领域广泛应用的文本相似性计算中,比起基于字符串和词袋模型的方法,词向量包含更丰富的语义信息,能更准确地比较文本相似度,因此将会被继续探索使用. 随着词向量的成功和应用,科研合作预测领域也把网络表示学习应用其中,将节点表示成向量计算节点结构相似性的方法已有了一定的实践,后续研究尝试使用考虑更全面的结构语义信息的方法,把元路径、子图、图等其他网络结构表示成向量应用于作者相似度计算.

(2) 学术知识图谱为作者相似度研究提供多方面的支持. 学术知识图谱是一种语义网,包含了丰富的作者及相关实体属性信息与结构信息,能够支持从属性信息和网络结构两方面比较作者间相似性,无论是在作者消歧还是在科研合作预测领域都有广阔的应用前景. 因此利用知识图谱比较作者相似度值得深入研究.

(3) 大数据给图书情报领域带来了挑战,也带来了机遇. 精准知识服务是图书情报领域面向未来的转折切入点. 作者相似性计算作为精准服务的基础关键技术方法,在很大程度上影响了精准服务的发展. 一个有效的作者相似度计算模型常常不能依赖对于相似性算法的简单评判,还需要根据应用数据集的具体特性. 作为多个领域的基础研究问题,作者相似度研究已取得诸多进展,而且不断引入的新兴技术持续改进着现有研究方法,这也将进一步推动精准服务的发展.

[参　考　文　献]

[1] MAKAROV I, BULANOV O, ZHUKOV L E. (2017) Co-author recommender system[C]. Models, Algorithms, and Technologies for Network Analysis. Cham:Springer,2016,197:251-257.

[2] ZHAO D,STROTMANN A. Intellectual structure of stem cell research:a comprehensive author co-citation analysis of a highly collaborative and multidisciplinary field [J]. Scientometrics,2011,87(1):115-131.

[3] HAN H, GILES L, ZHA H, et al. Two supervised learning approaches for name disambiguation in author citations[C]. Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries. New York:ACM,2004:296-305.

[4] TORVIK V I,SMALHEISER N R. Author name disambiguation in medline[J]. Acm Transactions on Knowledge Discovery from Data,2009,3(11):1-29.

[5] HAN H,ZHA H,GILES C L. Name disambiguation in author citations using a *k*-way spectral clustering method[C]//Joint Conference in Digital Libraries. Newyork:ACM,2005:334-343.

[6] **蒲旭，王建勇，范小明**. GHOST:**作者名字排歧系统**[J]. **计算机研究与发展**,2010,47(z1):512-515.

[7] HUSSAIN I, ASGHAR S. Author name disambiguation by exploiting graph structural clustering and hybrid similarity[J]. Arabian Journal for Science and Engineering,2018,43(12):7421-7437.

[8] **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.

[9] 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.

[10] **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.

[11] LUONG N T, NGUYEN TT, 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 International Publishing,2015,9011:127-136..

[12] **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.

[13] **王卫，李晓娜，闫帅**. **基于链路分析的作者合著关系预测研究：以图情领域为例**[J]. **现代情报**,2018,38(11):109-115.

[14] WANG C,SATULURI V,PARTHASARATHY S. Local probabilistic models for link prediction[C]//Data Mining,2007. ICDM2007. Seventh IEEE International Conference On. Washington DC:IEEE,2007:322-331.

[15] YU Q, LONG C, LV Y, et al. Predicting co-author relationship in medical co-authorship networks[J]. Plos One, 2014, 9(7):e101214.

[16] **陈二静，姜恩波**. **文本相似度计算方法研究综述**[J]. **数据分析与知识发现**,2017(6):1-11.

[17] WU H, LI B, PEI Y, et al. Unsupervised author disambiguation using Dempster-Shafer theory[J]. Scientometrics, 2014, 101(3):1955-1972.

[18] GURNEY T, HORLINGS E, BESSELAAR P V D. Author disambiguation using multi-aspect similarity indicators[J]. Scientometrics,2012,91(2):435-449.

[19] BAGGA A. Coreference, cross-document coreference, and information extraction methodologies [M]. Durham: Duke University,1998:4234.

[20] CHUAN P M,SON L H,ALI M,et al. Link prediction in co-authorship networks based on hybrid content similarity metric [J]. Applied Intelligence,2018,48(8):2470-2486.

[21] COUNCILL I,GILES C,HUANG J,et al. Efficient topic-based unsupervised name disambiguation[J]. Jcdl,2007,39:342-351.

[22] 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..

[23] **刘俊婉，杨波，王菲菲**. **基于引证行为与学术相似度的学者影响力领域排名方法研究**[J]. **数据分析与知识发现**,2018,4:59-70.

[24] **吴振新，单嵩岩**. **科学家相关性测度典型算法比较与评析**[J]. **数字图书馆论坛**,2019,3:11-17.

[25] **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.

[26] **刘萍，黄纯万**. **基于** SimRank **的作者相似度计算**[J]. **情报理论与实践**,2015,38(6):109-114.

[27] **刘萍，郭月培，郭怡婷**. **利用作者关键词网络探测作者相似性**[J]. **数据分析与知识发现**,2013(12):62-69.

[28] 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 & Mining. Washington DC:IEEE,2011:121-128.

[29] **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.

[30] **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.

[31] 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.

[32] 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.
[33] 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.
[34] 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.
[35] 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,2014:701-710.
[36] TANG J,QU M,WANG M Z,et al. Line:largescale information network embedding[C]//In Proceedings of the 24th International Conference on World Wide Web. Florence:ACM,2015:1067-1077.
[37] 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,2016:855-864.

Review on the author similarity algorithm in the field of author name disambiguation and research collaboration prediction

SHAN Song-yan[1,2],WU Zhen-xin[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:From the perspective of text similarity and network topological similarity algorithm,this paper studies the author similarity algorithm in the field of author disambiguation and cooperative prediction. The paper analyzes and compares the advantages and disadvantages of various commonly used algorithms,as well as the current application. The study summarizes the author similarity algorithm,and look forwards to the future development direction.

Keywords:author similarity;text similarity;network topological similarity;author name disambiguation;research collaboration prediction

(责任编辑:石绍庆)