云南民族大学学报(自然科学版),2024,33(3):377 - 385　　CN 53 - 1192/N　ISSN 1672 - 8513
doi:10.3969/j.issn.1672 - 8513.2024.03.013　　http://ynmz.cbpt.cnki.net

基于链路预测的潜在专利合作关系研究

于　凯[1],郭煜婕[2]
(1. 新疆财经大学 公共管理学院,新疆 乌鲁木齐 830012;2. 新疆财经大学 信息管理学院,新疆 乌鲁木齐 830012)

摘要: 发明人合作是专利合作的显著表现形式之一,通过构建合作网络,挖掘潜在的合作关系,有助于预测专利合作网络中发明人的未来合作趋势.考虑合作网络节点的位置信息和属性信息,首先,引入社交网络中的链路预测方法计算节点位置的相似性;其次,将发明人的研究方向作为节点的属性信息,分别运用 Doc2vec 和 TF - IDF 建立研究方向的向量耦合矩阵,计算发明人研究方向之间的余弦相似度,衡量发明人研究方向的相近性;最后,构建基于链路预测算法与发明人研究方向的混合算法,从而预测发明人的潜在合作关系.在国内知识图谱领域进行实证研究发现,TF - IDF 在该领域的预测效果较好,并且在链路预测算法的基础上,通过融入研究方向相近性矩阵,预测精度得到了较好的提升.

关键词: 专利合作;链路预测;知识图谱;Doc2vec;TF - IDF
中图分类号:TP391　　**文献标志码:**A　　**文章编号:**1672 - 8513(2024)03 - 0377 - 09

当前,世界已进入知识经济的时代,提升产学研能力已成为必然要求.2021 年是“十四五”开局之年,深入实施国家规划纲要和创新驱动发展战略需要以科技创新工作为首要目标,进一步加强学科交叉、多学科协同攻关.专利反映了最新的技术发明,是科技创新的重要组成部分.通过分析专利文献,可以揭示当前专利的发展状况、专利的技术构成,并可以从合作者的角度去挖掘专利领域的特点.而专利合作是资源共享的过程,在这个过程中由研发人员构成了发明人合作网络.已有研究表明,发明人之间的合作所引起的社会接近性是影响专利合作网络越来越重要的因素,对于提高合作创新有着显著的正向作用,并且研发人员之间的技术相似度越接近,就越容易促进人员间的知识转移和目标实现[1,2].基于专利文献的发明人合作是专利合作的显著表现形式之一,因此,构建基于专利发明人的合作网络,挖掘与预测未来可能的合作关系,有利于推动专利合作、促进学科的创新与发展.

1　相关研究

专利文献承载着大量的技术信息,是知识交流和创新能力的重要产出[3,4].通过研究专利技术发展路径,可以帮助研发人员更高效地选择研发方向,把握知识前沿[5].一些学者从合作主体的角度进行了相关研究.刘颖琦等从社会网络视角对技术热点和专利权人合作进行研究,探究其网络结构特点,进而提出我国智能网联汽车企业的技术研发前景[6].申通远等同样运用社会网络方法构建了专利合作网络,分析节点中心性在企业创新网络的参与程度[7].因此,构建专利合作网络,可以推测未来的专利研究发展趋势,揭示一般规律,提升创新能力.

运用链路预测可以挖掘复杂网络中节点之间的关系.网络中的链路预测是指利用已知的网络节点以及网络结构等信息,预测网络中尚未形成连边的两个节点之间产生链接的可能性[8].自然界中的很多系统都可以用复杂网络来表示,比如生物领域中的蛋白质相互作用网络、交通领域中的航空运输网络等.链路预测也被广泛运用在实际问题中,我们可以把现实中的用户或个体看作网络中的节点,而它们之间的关系可以看作网络中的连边.例如,社交网络的推荐算法,龙增艳等综合考虑了网络结构和用户属性,来推测用户之间建立好友关系的可能性,进而“推荐好友”,提

收稿日期:2022 - 03 - 08.
基金项目:新疆维吾尔自治区自然科学基金项目(2019D01A22);新疆维吾尔自治区社科基金项目(21BTQ162).
作者简介:于凯(1974 -),男,教授,博士,主要研究方向为链路预测、数据挖掘.
通信作者:郭煜婕(1996 -),女,硕士研究生,主要研究方向为链路预测.

高用户体验[9]. 再如,在科学家合作网络的预测方面,汪志兵等以化学领域的作者合作网络为研究对象,使用链路预测算法预测了两位尚未产生合作关系的科学家在未来合作的可能性[10].

专利研究的拓展同时推动了多学科领域交叉,促使了不同研究主体之间的合作,也涌现出了新的合作关系. 在专利合作网络中,网络结构表示合作主体在网络中的节点位置信息,如果节点所处的位置相似,说明他们存在链接的可能性就越大,则更有可能成为合作伙伴. 链路预测方法同样适用于预测专利合作网络中未连接节点之间合作的可能性,Chen 等构建了专利合作组织间的合作网络,通过链路预测的方法探讨了网络中合作伙伴的选择,证明了链路预测方法在专利合作组织选择合作伙伴时的有效性,表明历史合作信息对合作伙伴的选择有积极影响,是组织选择合作伙伴时考虑的重要因素[11]. 尽管单一使用网络的结构信息能够证明链路预测在专利合作预测的有效性,但合作主体作为研究人员或组织,仅仅使用结构信息的关系预测节点的相似性具有一定的局限性,即对合作主体的属性信息掌握不足.

为了更好地刻画节点的相似性,学者们不再局限于单一的链路预测,而是将其与自然语言处理和文本挖掘的研究方法相结合. 任海英等在关键词共现引入了链路预测算法,综合考虑了不同关键词组合的可能性[12]. 刘俊婉等采用了 LDA 主题模型和链路预测相结合的方法,计算主题之间的相似度,对新兴主题的未来关联机会进行了识别[13]. 魏玉梅等通过 TF – IDF 对专利摘要进行了文本挖掘,计算专利之间的语义相似度,通过构建竞争性专利网络,从网络特征、机构与持有人分布、申请人技术竞争对 3D 打印技术领域的专利进行了多个维度的分析[14]. 王菲菲等采用了链路预测和 LDA 主题模型相结合的方法,探测了石墨烯领域产学研机构的研究热点和合作机会[15]. 尽管以上研究使得专利合作研究深入到了文本内容方面,但更多的基于专利研究机构的合作和研究热点的识别,对于专利发明人的合作考虑较少.

发明人作为重要的专利研发主体,已有文献表明,加大专利 R&D 人员投入对于创新产出有着显著的正向影响[16]. 除了网络中节点的位置信息,采用发明人的研究方向刻画节点的属性信息,因为研究方向往往体现于所发表专利的标题或者关键词等等,如果他们的研究方向越接近,则说明他们越相似,在网络中也更容易产生链接. 韩菁等基于知识网络与合作网络,考虑专利发明人的知识属性,将共同邻居指标与所提出的知识属性指标结合构建了混合链路预测算法,并在新能源汽车领域的专利证明了混合指标的有效性[17]. 因此,引入链路预测算法,挖掘和预测未来发明人可能产生的合作关系,有助于探索各专利领域今后的合作趋势. 此外,可以通过结合文本信息提取发明人的研究方向,进行相似性分析,从而把握专利发展脉络. 因此,将链路预测和文本内容两者相结合,能够更全面地挖掘发明人之间的合作关系.

2 研究思路与方法

基于专利合作网络,选取网络拓扑结构的相似性指标,运用链路预测探究节点的位置相似性,在此基础上融入发明人的研究方向作为节点的属性信息,构建基于节点相似性指标与发明人研究方向的专利合作混合预测算法,以此挖掘发明人的潜在合作关系. 研究思路如图 1 所示.

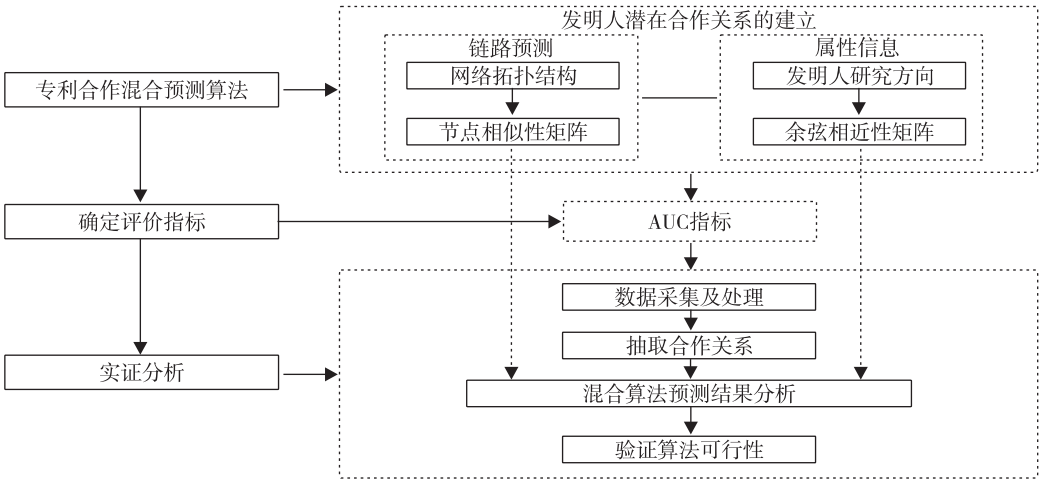

图 1 研究思路

2.1　专利发明人潜在合作关系的构建

2.1.1　专利合作网络的相似性测度

常用的链路预测算法有数十种,并且不同算法在不同的网络上表现也有差异. 考虑网络的拓扑结构,本文从三种维度对比分析 14 种链路预测方法,其中包括基于局部信息的 CN 指标、Salton 指标、Jaccard 指标、Sorensen 指标、HPI 指标、HDI 指标、LHN - Ⅰ指标、PA 指标、AA 指标、RA 指标,基于路径的 LP 指标、Katz 指标和基于随机游走的 ACT 指标、Cos + 指标,这些方法被广泛应用于预测社交网络节点之间的链接关系[8,11]. 具体方法如表 1 所示.

表 1　14 种链路预测指标计算公式

类型	指标	定义
基于局部信息	CN	$s_{xy}=\lvert\Gamma(x)\cap\Gamma(y)\rvert$
	Salton	$s_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\sqrt{k_x k_y}}$
	Jaccard	$s_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$
	Sorensen	$s_{xy}=\frac{2*\lvert\Gamma(x)\cap\Gamma(y)\rvert}{k_x+k_y}$
	HPI	$s_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\min\{k_x,k_y\}}$
	HDI	$s_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\max\{k_x,k_y\}}$
	LHN - Ⅰ	$s_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{k_x k_y}$
	PA	$s_{xy}=k_x k_y$
	AA	$S_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{\log k_z}$
	RA	$s_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{k_z}$
基于路径	LP	$s_{xy}=A^2+\alpha*A^3$
	Katz	$s_{xy}=(I-\alpha*A)^{-1}-I$
基于随机游走	ACT	$s_{xy}=\frac{1}{l^+_{xx}+l^+_{yy}-2l^+_{xy}}$
	Cos +	$s_{xy}=\cos(x,y)^+=\frac{l^+_{xy}}{\sqrt{l^+_{xx}*l^+_{yy}}}$

作为最简单的基于局部信息的相似性指标,共同邻居指标(CN)考虑两个节点的共同邻居节点,共同邻居节点越多,那么两个节点越相似;在 CN 指标的基础上又产生了 Salton 指标、Jaccard 指标、Sorensen 指标、大度节点有利指标(HPI)指标、大度节点不利指标(HDI)、LHN - Ⅰ指标,这些指标都是从不同角度考虑了节点度的影响;偏好连接指标(PA)可以用来构建无标度网络;AA 指标和 RA 指标异曲同工,都是根据度小的共同邻居节点的贡献大于度大的思想,区别是赋予共同邻居节点权重方式的不同. 基于路径的相似性指标是对基于 CN 指标的改进,局部路径指标(LP)考虑了三阶路径,当局部路径扩展到考虑所有路径,就得到了 Katz 指标. 另外,选取了两个基于随机游走的相似性指标平均通勤时间(ACT)和基于随机游走的余弦相似性(Cos +),二者都是基于网络全局的随机游走指标.

运用以上方法,可以从网络结构方面计算节点位置之间的相似性,构建基于链路预测算法的相似性矩阵.

2.1.2　专利发明人研究方向的相近性测度

(1)研究方向的预处理

发明人的研究方向可以体现在专利名称和专利关键词,由于中国知网专利数据库并未明确标识一项专利的关键词,因此将专利名称作为数据集合,从专利名称提取关键词作为发明人的研究方向. 运用 jieba 分词进行文本预处理,将专利名称进行划分,去除停用词,提取主要关键词. 测度发明人研究方向的相近性,需要把各发明人的研究方向转化为向量,这里选取两种常用的方法,分别是 Doc2vec 模型和 TF - IDF.

Doc2vec 模型作为一种无监督的文本聚类方法,可以从文本信息的角度分析内容上的相似程度,用来寻找相似的专利[18]. Doc2vec 模型是在 Word2vec 模改进,增加了段落向量,可以用于可变长度文本的学习[19]. 由于发明人的专利数量不同,专利名称所形成的文档文本长度不一,因此 Doc2vec 模型更适合处理专利文本. Doc2vec 模型有 DM 和 DBOW 两种模型,前者通过上下文预测目标词的概率,后者根据目标词来对上下文的概率做出预测. 本文采用 Doc2vec 模型中的 DM 模型,建立发明人的研究方向耦合的向量矩阵,如图 2 所示.

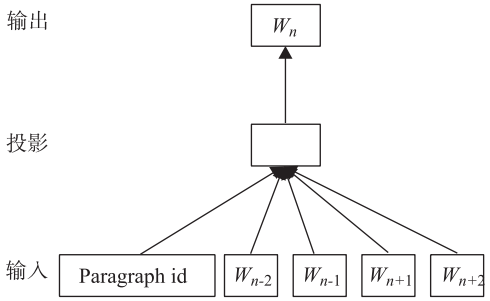

图 2　DM 模型示意图

TF-IDF,即词频—逆文档频率,作为一种统计方法,既可以考虑词语的频率,又可以考察词语在数据集合中的分布情况[20]. 因此,采用词频—逆文档频率同样可以建立发明人的研究方向耦合的向量矩阵. 具体步骤如下:

首先,计算所获得的关键词出现的频率. 其中,关键词频 TF_{ij} 中的元素 f_{ij} 表示关键词 i 在发明人 j 所发表的专利名称中出现的次数,使用发明人 j 中所有 k 个关键词词频之和进行标准化处理. 计算公式为(1):

$$TF_{ij} = \frac{f_{ij}}{\sum_k f_{kj}}. \tag{1}$$

然后,计算 IDF,即逆文档频率,用于衡量关键词在发明人所有专利名称中的分布情况. 其中,N_j 表示发明人 j 发表的专利总数,N_{ij} 表示发明人 j 所发表的专利集中含有关键词 i 的专利数,计算公式为(2):

$$IDF_{ij} = \lg \frac{N_j}{N_{ij}}. \tag{2}$$

最后,将二者相乘得到关键词词频—逆文档频率,获得发明人的研究方向耦合向量矩阵. 计算公式(3):

$$W_{ij} = TF_{ij} \times IDF_{ij}. \tag{3}$$

(2)研究方向的相近性测度

在获得各发明人的向量之后,使用余弦相似度测度发明人研究方向的相近性. 如果数值越大,则说明两位发明人的研究方向越相近. 其中,A_x 和 A_y 分别表示发明人 x 和发明人 y 的文本向量,计算公式(4):

$$\cos\_d(x,y) = \frac{A_x * A_y}{\sqrt{A_x^2} * \sqrt{A_y^2}}. \tag{4}$$

计算不同发明人之间的余弦相似度,建立研究方向的相近性矩阵,将其作为混合算法的一部分.

2.1.3 发明人潜在合作关系的混合算法构建

一般来说,一个节点的属性信息越充分,预测相对就越准确. 在预测过程中,既不能忽视本身存在的合作关系,也不能忽视网络结构的外部信息. 因此,在合作网络结构相似性的基础上,引入发明人的研究方向,将二者结合,构建一种混合预测算法,挖掘发明人之间的潜在合作关系. 其中 A、B 分别为基于链路预测算法的相似性矩阵和基于发明人研究方向的相近性矩阵,w 为权重,取值为[0,1],当 w 为0时,该混合算法退化为链路预测算法的相似性指标,C 表示为公式(5):

$$C = A + wB. \tag{5}$$

w 不同取值产生的具体影响将在下文中讨论.

2.2 链路预测评价指标

为衡量算法整体的精确度,选择 AUC 指标作为链路预测评价指标,即独立比较 n 次,随机选择测试集中存在连边的分数值高于不存在连边的分数值的概率[21]. 如果前者分数较高,则加1分,记为 n' 次;如果二者相等,则加0.5分,记为 n'' 次. 由于网络规模越大,所带来的计算量就越大,因此在保证 AUC 计算的绝对误差不超过千分之一的情况下,设 n = 672 400[22]. 定义如下:

$$AUC = \frac{n' + 0.5n''}{n}. \tag{6}$$

3 实证研究

3.1 数据采集及处理

当前,知识图谱在图书情报界信息处理中起着重要的作用,其作用得到越来越广泛关注. 在此背景下,以中国知网(CNKI)为数据源,以“知识图谱”为主题进行检索,以专利公开日为发表时间,采集国内知识图谱领域专利的有效数据共4154篇,检索日期为2021年1月20日.

2012年Google正式提出知识图谱,因此除去完全未公开发明人的专利,取2012-2020年的专利数据共4 140篇. 由数据分析可知,知识图谱领域的专利合作网络的数据量较大,但稀疏程度较高. 为更好地探究发明人研究方向对合作的影响,在已有的数据中提取专利发表数量5篇及5篇以上的491位发明人作为研究对象,抽取其合作关系,并提取其所发表专利的名称. 去重后,共计1 573个专利名称,在此基础上构建合作网络,该网络共491个节点,1 068条连边,即这491位发明人产生了1 068条合作链接.

3.2 专利合作网络链路预测

3.2.1 链路预测指标对发明人合作关系的预测

选取训练集比例分别为0.9、0.8、0.7、0.6、0.5,随机地对合作网络进行划分,并在此基础上,运用上述的链路预测指标分别对每个比例下的网络进行50次独立实验,取平均数,预测的 AUC 结果如图3所示.

如图3所示,在同一训练集比例下,14种相似性指标得到的 AUC 存在一定的差异,但不同训练集比例的预测结果的变化趋势大致相同:随着训练集比例的升高,各个指标的预测结果都有所升高,基于

路径的 Katz 指标均保持着较高的水准,LP 指标也都处于 0.9 以上;基于局部信息的 CN 指标、Salton 指标、Jaccard 指标、Sorensen 指标、HDI 指标、AA 指标、RA 指标均处于 0.8 以上,HPI 指标、LHN－1 指标、PA 指标表现略低. 而基于随机游走的 ACT 指标和 Cos＋指标平均值介于 0.5 到 0.6 之间,明显低于其他算法. 以上结果表明,基于路径的相似性指标更适合进行专利合作的预测,在一定程度上也符合前文所说的基于路径的相似性指标是对基于 CN 指标的改进.

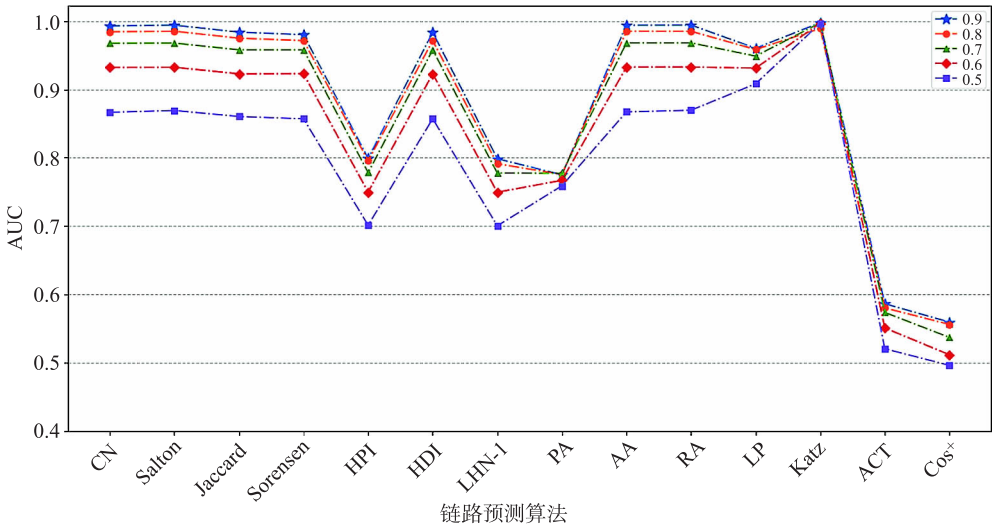

图 3　不同训练集比例的链路预测结果

3.2.2　专利发明人研究方向的相近性预测

在获取 1 573 个专利名称后,提取其中的关键词,去重后共获得关键词 1 337 个,分别运用 Doc2vec 模型和 TF－IDF 建立发明人研究方向的向量耦合矩阵,测度发明人在研究方向上的相近性,并分别进行预测.

(1)研究方向的相近性测度

使用 Python 中的 Gensim 包实现 Doc2vec 模型的计算,得到发明人的研究方向向量耦合矩阵,并运用余弦相似度计算两个发明人在研究方向上的相近性,部分结果如表 2 所示. 其中数值越大,表示发明人之间的研究方向越相近.

表 2　运用 Doc2vec 模型获得的余弦相似性矩阵(部分)

	段玉聪	王振宇	查琳	陈刚	朱勇	赵龙	陈黄	贾巨涛	…
段玉聪	0	0.301 897	0.413 972	0.507 989	－0.046 82	0.500 172	0.542 197	－0.004 24	…
王振宇	0.301 897	0	0.987 992	0.966 658	－0.121 25	0.968 521	0.950 715	－0.083 32	…
查琳	0.413 972	0.987 992	0	0.992 857	－0.120 75	0.991 847	0.981 522	－0.141 34	…
陈刚	0.507 989	0.966 658	0.992 857	0	－0.114 82	0.995 267	0.993 526	－0.122 06	…
朱勇	－0.046 82	－0.121 25	－0.120 75	－0.114 82	0	－0.070 91	－0.012 26	0.466 26	…
赵龙	0.500 172	0.968 521	0.991 847	0.995 267	－0.070 91	0	0.994 545	－0.121 87	…
陈黄	0.542 197	0.950 715	0.981 522	0.993 526	－0.012 26	0.994 545	0	－0.072 89	…
贾巨涛	－0.004 24	－0.083 32	－0.141 34	－0.122 06	0.466 26	－0.121 87	－0.072 89	0	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

运用 TF－IDF 计算得到发明人的研究方向向量耦合矩阵,并运用余弦相似度计算两个发明人在研究方向上的相近性,部分结果如表 3 所示. 其中数值越大,表示发明人之间的研究方向越相近.

表 3 运用 TF – IDF 获得的余弦相似性矩阵(部分)

	段玉聪	王振宇	查琳	陈刚	朱勇	赵龙	陈黄	贾巨涛	…
段玉聪	0	0. 310 945	0. 296 509	0. 297 293	0. 236 264	0. 326 573	0. 296 530	0. 207 342	…
王振宇	0. 310 945	0	0. 975 058	0. 958 963	0. 299 205	0. 929 212	0. 970 973	0. 281 521	…
查琳	0. 296 509	0. 975 058	0	0. 982 720	0. 296 552	0. 950 112	0. 989 465	0. 261 947	…
陈刚	0. 297 293	0. 958 963	0. 982 720	0	0. 313 968	0. 939 682	0. 987 308	0. 271 422	…
朱勇	0. 236 264	0. 299 205	0. 296 552	0. 313 968	0	0. 306 304	0. 303 147	0. 463 955	…
赵龙	0. 326 573	0. 929 212	0. 950 112	0. 939 682	0. 306 304	0	0. 941 525	0. 280 469	…
陈黄	0. 296 530	0. 970 973	0. 989 465	0. 987 308	0. 303 147	0. 941 525	0	0. 269 18	…
贾巨涛	0. 207 342	0. 281 521	0. 261 947	0. 271 422	0. 463 955	0. 280 469	0. 269 180	0	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(2)研究方向的相近性预测

分别运用两种方法获得的余弦相似性矩阵,进行 50 次独立预测实验,获得的 AUC 结果如图 4 所示.

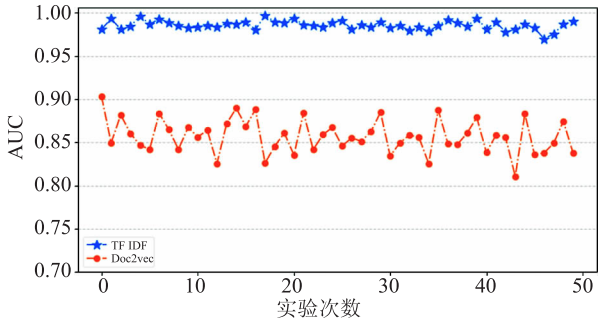

图 4 两种方法的预测结果对比

由图 4 可知,与 Doc2vec 模型相比,使用 TF – IDF 获得的余弦相似性矩阵的预测效果较好,AUC 值主要围绕着 0. 98 上下波动,且方差较小,方法稳定,更适合运用于知识图谱领域的专利研究. 因此,在下一步混合预测中,主要选择 TF – IDF 与链路预测算法结合.

3. 2. 3 发明人潜在合作关系的混合预测结果

(1)发明人研究方向权重对预测结果的影响

根据图 3 的链路预测结果,为了更直观地观察发明人研究方向对算法结果的波动情况,设定训练集比例为 0. 5 进行试验,并进一步观察不同权重对混合算法的影响. 将合作网络的链路预测算法与 TF – IDF 获得的发明人研究方向的余弦相似性矩阵按照公式(5)进行混合,构建混合算法,设置不同的 w 值,计算在不同权重参数下的预测结果. 每种权重随机进行 50 次独立实验,取平均数,结果如图 5 所示.

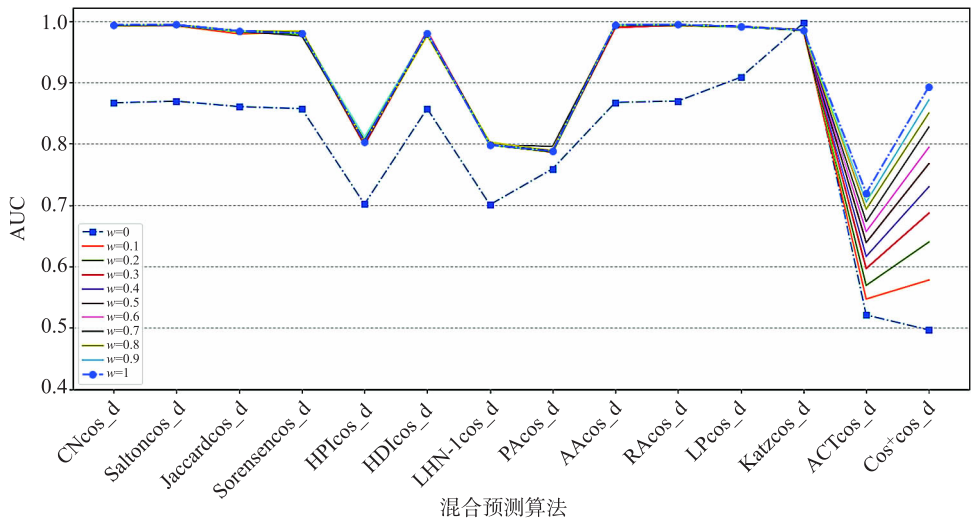

图 5 不同 w 值的混合预测结果

图 5 刻画了 14 种算法在不同 w 值下均值的变化情况,可以看出除了 Katzcos_d 指标,其余考虑发明人研究方向的混合算法的预测结果要明显优于仅考虑链路预测指标的预测结果. 并且随着 w

值的升高,预测结果的 AUC 值虽然相差不大,但有所提高,说明引入发明人研究方向的相近性矩阵对于预测知识图谱领域专利发明人的合作关系是有效的.

(2)不同训练集比例对预测结果的影响

在探究合作关系时,往往存在不同的数据情况,不同的训练集比例对于预测结果也会有一定的影响. 因此依据上述结果,将混合算法中的 w 值设置为 1,观察在发明人研究方向的最优权重的情况下,不同训练集比例对发明人的研究方向对于预测结果的影响. 在不同训练集的比例下,随机进行 50 次独立实验,取平均数,结果如图 6 所示.

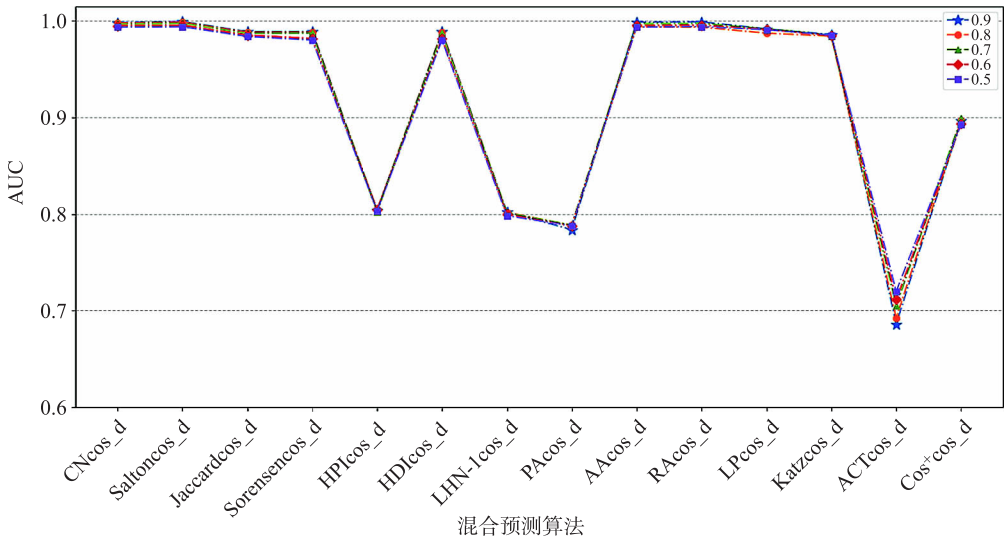

图 6　不同训练集比例的混合预测结果

如图 6 所示,即使在不同训练集比例下,混合预测结果依然很接近. 与图 3 相比,多数算法的预测精度有较为明显的提高. 如表 4 所示,五种比例下除 Katzcos_d 指标的精度有所下降,其余的算法混合相较原算法均有所上升. 尤其是,混合算法的构建可以使预测算法基本不受训练集比例的影响,在训练集比例较低的情况下,仍然能够得到与在训练集比例较高的情况下相比类似的结果,算法较为稳定. 如表 5 所示,除去本身链路预测结果较低的算法,超过半数的混合算法要优于仅使用单一的发明人研究方向的余弦相似性矩阵的预测结果. 以上结果说明专利发明人潜在合作关系的挖掘是可行的,混合算法的构建是合理的并且具有一定的鲁棒性,对于知识图谱领域专利的发明人合作关系具有一定的预测意义.

表 4　混合算法相比原算法的提高精度　　(%)

训练集	训练集比例				
	0.9	0.8	0.7	0.6	0.5
Cn	0.53	1.35	3.03	6.73	14.65
Salton	0.53	1.36	3.08	6.76	14.35
Jaccard	0.51	1.34	3.04	6.76	14.34
Sorensen	0.91	1.73	3.07	6.36	14.35
HPI	0.59	0.96	3.09	7.26	14.49
HDI	0.54	1.74	3.07	6.37	14.35
LHN－1	0.43	1.29	2.96	6.67	13.91
PA	1.09	1.60	1.42	2.64	3.79
AA	0.49	0.89	3.03	6.71	14.59
RA	0.48	0.86	3.02	6.73	14.32
LP	3.29	3.01	4.51	6.44	9.09
Katz	－1.17	－0.50	－1.23	－1.19	－1.15
ACT	16.94	19.26	22.82	29.06	38.16
Cos +	60.23	61.21	67.29	74.47	79.87

表5 混合算法相比仅使用研究方向的余弦相似性矩阵的提高精度 (%)

训练集	训练集比例				
	0.9	0.8	0.7	0.6	0.5
Cn	1.25	1.40	1.19	1.02	0.83
Salton	1.36	1.50	1.27	1.06	0.89
Jaccard	0.30	0.43	0.20	0.03	-0.15
Sorensen	0.32	0.45	0.23	-0.33	-0.52
HDI	0.31	0.44	0.22	-0.34	-0.52
AA	1.31	1.02	1.24	1.05	0.86
RA	1.33	0.98	1.25	1.08	0.90
LP	0.60	0.35	0.66	0.62	0.57
Katz	0.00	0.04	0.00	-0.02	0.00

4 结语

本文从专利合作的角度,构建了专利发明人的合作网络,选取网络拓扑结构的相似性指标,以链路预测探究节点的位置相似性为基础,融入发明人研究方向的相近性矩阵,提出了混合预测算法. 在该算法中既考虑了合作网络中节点局部信息、路径、随机游走等的影响,又考虑了不同权重的发明人研究方向的相近性,并区别了不同训练集比例对不同预测指标的影响. 以2012年到2020年间国内知识图谱领域的专利信息为对象进行实证分析,证实了该算法的可行性. 研究得出,仅考虑网络拓扑结构预测时,基于路径的Katz指标预测更准确;单独使用发明人研究方向进行预测时,TF-IDF比Doc2vec获得的相近性矩阵在该领域的预测效果较好;从混合算法的预测效果看,融入研究方向相近性矩阵的链路预测算法使预测效果随着研究方向相近性矩阵权重的逐渐提高而提高,并且即使在不同训练集比例下,低比例的预测结果与高比例的结果相当接近,混合预测结果均提高到了较高的水平. 因此,本文提出的考虑网络拓扑结构和发明人研究方向相似性指标的混合预测算法是有效的,同时也对于专利发明人合作关系的预测具有一定的理论意义和实践价值. 但不足之处在于,仅以国内知识图谱领域的专利进行研究,混合预测算法的普适性有待进一步检验,之后将做不同领域的深入挖掘.

参考文献:

[1] 王嘉杰,孙建军,石静,等. 技术人员流动视角下的知识转移影响因素分析——技术距离与技术多元化的作用[J]. 信息资源管理学报,2021,11(5):114-123.

[2] 王黎萤,池仁勇. 专利合作网络研究前沿探析与展望[J]. 科学学研究,2015,33(1):55-61.

[3] 温芳芳. 基于专利计量的区域间技术合作网络研究[J]. 情报杂志,2013,32(11):32-36.

[4] 李欣,温阳,黄鲁成,等. 多层网络分析视域下的新兴技术研发合作网络演化特征研究[J]. 情报杂志,2021,40(1):62-70.

[5] 赵蓉英,李新来,李丹阳. 专利引证视角下的核心专利研究——以人工智能领域为例[J]. 情报理论与实践,2019,42(3):78-84.

[6] 刘颖琦,周菲,席锐. 后疫情时期中国智能网联汽车产业技术研究与合作网络:国际专利视角[J]. 中国科技论坛,2021(5):32-45.

[7] 申通远,朱玉杰. 创新合作社会网络中企业中心性特征的影响因素[J]. 技术经济,2018,37(11):19-29.

[8] 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.

[9] 龙增艳,陈志刚,徐成林. 基于用户交互的社交网络好友推荐算法[J]. 计算机工程,2019,45(3):132-137.

[10] 汪志兵,韩文民,孙竹梅,等. 基于网络拓扑结构与节点属性特征融合的科研合作预测研究[J]. 情报理论与实践,2019,42(8):116-120.

[11] CHEN W,QU H,CHI K. Partner selection in China inter-organizational patent cooperation network based on link prediction approaches [J]. Sustainability, 2021, 13(2):1003.

[12] 任海英,于立婷,王菲菲. 国内外技术预见研究的热点和趋势分析[J]. 情报杂志,2016,35(2):81-87.

[13] 刘俊婉,龙志昕,王菲菲. 基于LDA主题模型与链路预测的新兴主题关联机会发现研究[J]. 数据分析与知识发现,2019,3(1):104-117.

[14] 魏玉梅,滕广青,郭思月,等. 基于竞争专利网络的技术竞争多维分析:以3D打印技术为例[J]. 信息资源管理学报,2020,10(4):99-108.

[15] 王菲菲,芦婉昭,贾晨冉,等. 基于论文-专利机构合作网络的产学研潜在合作机会研究[J]. 情报科学,2019,37(9):9-16.

[16] 高华. 专利文献引用对区域创新产出影响的实证研究[J]. 技术经济与管理研究,2017(4): 48 - 51.
[17] 韩菁,唐箫,余乐安. 基于多层网络链路预测的潜在合作关系识别研究[J]. 系统工程理论与实践,2021,41(4): 1049 - 1060.
[18] 张海超,赵良伟. 利用 Doc2Vec 判断中文专利相似性[J]. 情报工程,2018,4(2): 64 - 72.
[19] LE Q V,MIKOLOV T. Distributed representations of sentences and documents[J]. Eprint Arxiv,2014(4):1188 - 1196.
[20] 丁敬达,郭杰. 融合内容相似度和路径相似性的潜在作者合作关系挖掘[J]. 情报理论与实践,2021,44(1): 124 - 128.
[21] FAWCETT T. An introduction to ROC analysis[J]. Pattern Recognition Letters,2006,27(8): 861 - 874.
[22] 吕琳媛,周涛. 链路预测[M]. 北京: 高等教育出版社,2013.

Research on potential patent partnership based on link prediction

YU Kai[1],GUO Yu-jie[2]

(1. School of Public Administration,Xinjiang University of Finance and Economics,Urumqi 830012,China;
2. School of Information Management,Xinjiang University of Finance and Economics,Urumqi 830012,China)

Abstract: Inventor cooperation is one of the obvious manifestations of patent cooperation. By building a cooperation network and digging out potential cooperation relationships,it is helpful to predict the future cooperation trend of inventors in the patent cooperation network. Considering the location information and attribute information of cooperative network nodes,first,the link prediction method in social network is introduced to calculate the similarity of node locations; then,the inventor's research direction is used as the attribute information of nodes,and Doc2vec and TF-IDF are used respectively. Establish the vector coupling matrix of research directions,calculate the cosine similarity between the inventors´research directions,and measure the similarity of the inventor's research directions; finally, construct a hybrid algorithm based on the link prediction algorithm and the inventor's research direction,so as to predict the inventors potential partnership. Empirical research in the field of domestic knowledge graph found that TF-IDF has a better prediction effect in this field,and based on the link prediction algorithm,the prediction accuracy has been improved by incorporating the research direction similarity matrix.

Key words: patent cooperation; link prediction; knowledge graph; Doc2vec; TF-IDF

(**责任编辑**　梁志茂)