符号网络中融合聚集系数与符号影响力的链路预测算法

刘苗苗[1,3]，扈庆翠[1]，郭景峰[2]，陈　晶[2]
(1. 东北石油大学计算机与信息技术学院，大庆 163318；2. 燕山大学信息科学与工程学院，秦皇岛 066004；
3. 黑龙江省石油大数据与智能分析重点实验室，大庆 163318)

摘　要：为快速、准确地实现符号社会网络中的链接预测与符号预测双重目标，提出一种融合共同邻居节点的聚集系数与连边符号影响力的链路预测算法. 基于结构平衡理论，有效利用节点的度、聚集系数、路径上的中间传输节点、连边符号及其影响力等信息，分别定义了两节点基于一阶共同邻居和二阶共同邻居的相似性，最终得到两节点的总相似性得分，用其绝对值度量两节点建立链接的可能性，通过其符号获得链接的符号预测结果，从而实现符号网络中的链路预测. 在 6 个有代表性的符号网络数据集上进行了实验，以 AUC、调整的 Precision'、Accuracy 等为评价指标，对比了多个符号网络链接预测算法，并进行了可调步长参数的敏感性分析. 实验结果表明，所提算法在符号网络链接预测与符号预测两方面均达到了较好的性能，无论是稀疏网络还是负链接预测，准确性均高于其他算法.

关键词：符号社会网络；链接预测；符号预测；聚集系数；结构平衡理论；相似性
中图分类号：TP301.6　　**文献标识码**：A　　DOI：10.19907/j.0490-6756.2021.052003

Link prediction algorithm in signed networks based on clustering coefficient and sign influence

LIU Miao-Miao[1,3], *HU Qing-Cui*[1], *GUO Jing-Feng*[2], *CHEN Jing*[2]
(1. College of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China;
2. College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China;
3. The Key Laboratory for Oil Big Data and Intelligent Analysis of Heilongjiang Province, Daqing 163318, China)

Abstract: In order to achieve the dual goals of link prediction and sign prediction in signed social networks quickly and accurately, a link prediction algorithm is proposed based on the clustering coefficient of common neighbor nodes and the influence of the sign of edges. With the structural balance theory, the similarity of the two nodes based on their first-order common neighbors and the second-order common neighbors is defined respectively by using the degree, clustering coefficient, intermediate transitive nodes, and the influence of the sign of the edge, the total similarity score of the two nodes is finally obtained and its absolute value is used to measure the possibility to establish a link of the two nodes, then its sign is the sign prediction result of the link. Accordingly, the link prediction and sign prediction are realized in signed networks. Experiments have been carried out on six representative signed network

收稿日期：2020-10-11
基金项目：国家自然科学基金(42002138)；黑龙江省自然科学基金(LH2019F042)；东北石油大学青年基金(2018QNQ-01)；东北石油大学优秀中青年科研创新团队培育基金(KYCXTDQ202101)
作者简介：刘苗苗(1982－)，女，教授，博士，硕士生导师，研究方向为社会网络分析. E-mail：liumiaomiao82@163.com
通讯作者：扈庆翠. E-mail：hqc70681@163.com

datasets, with evaluation indicators such as AUC, adjusted precision' and accuracy. The experiment results are compared with several link prediction algorithms in signed networks the sensitivity of adjustable step size parameters is also analyzed. Experimental results show that the proposed algorithm can achieve good performance in both link prediction and sign prediction, and its accuracy is higher than other algorithms for both sparse networks and the prediction of negative links.

Keywords: Signed social networks; Link prediction; Sign prediction; Clustering coefficient; Structural balance theory; Similarity

1 引　言

随着人工智能和机器学习技术的快速发展,涌现了许多社会媒体网络平台,随之也产生了大量复杂多样的网络大数据,对这些数据的表征和预测逐渐成为社会网络分析领域的研究热点. 在现实网络中,实体间通常具有正负两方面的关系. 例如,社会领域的人与人之间存在朋友和敌人关系,信息领域的用户间在观点上存在支持和反对关系,生物领域的细胞之间存在促进和抑制关系. 这种同时具有正负关系的网络称为符号社会网络,简称符号网络[1]. 它与传统无符号网络的区别在于节点间是否存在链接的正负符号属性. 符号网络是一种包含正负对立关系的二维网络,这种对立关系包括朋友、支持、喜欢等积极关系和敌人、反对、厌恶等消极关系.

符号网络是社会网络的重要分支,因其更贴近现实世界的特性,从而受到学术界的广泛关注. 有关符号网络的研究主要集中在其结构分析,其中一个研究热点便是链路预测[2]. 它通过对观察到的网络结构进行分析来预测未知的链接,包括对未知链接建立的可能性预测、未知链接的符号预测以及对已观测到的链接缺失的符号类型的预测[3]. 符号网络链路预测在社会与信息领域的推荐系统以及知识图谱中实体间关系的学习中有着广泛的应用,在生物领域蛋白质相互作用关系的发现方面更是有着实际的意义和价值,可帮助指导实验过程,在节省时间和成本的同时提高预测准确率.

目前,针对符号网络的链路预测研究主要有基于节点相似性、基于概率统计的矩阵分解或填充以及基于机器学习等方法. 第一类方法主要结合结构平衡理论,利用符号网络的局部或全局信息如节点的度、共同邻居数量、路径特征等设计相似性指标,代表性研究成果有:文献[4]提出基于结构平衡理论与网络局部特征的符号预测算法. 文献[5]利用路径上传输节点相似度以及拉普拉斯聚类算法实现了符号网络的链路预测及推荐. 文献[6]在基于节点共同邻居的符号预测算法 CN-Predict 的基础上,融合符号密度提出改进后的 ICN-Predict 算法,能够较好地实现符号预测,但该方法针对负链接的预测效果欠佳. 文献[7]提取结构平衡环的局部特征以及频繁子图出现的次数构建特征来进行符号预测,但时间复杂度较高. 文献[8]结合局部路径指标以及结构平衡理论对符号网络链路预测方法进行了研究. 文献[9]提出一种符合结构平衡理论的高度对称四边形结构,基于局部结构的统计特性提取节点对的相似性、相异性以及反映节点对正负态度倾向的构造特征,在此基础上完成了符号预测. 文献[10]融合结构平衡理论与节点的局部和路径结构相似性提出了一种符号网络边值预测算法 PSNBS. 文献[11]以结构平衡理论为基础,将 Katz 指标与网络拓扑相融合,提出一种符号预测算法. 文献[12]考虑到负链接在符号网络中的重要性,融合结构平衡理论和地位理论提出基于隐空间映射矩阵的符号网络链接预测算法,在 Epinions 和 Slashdot 数据集上获得了较好的效果. 针对符号网络拓扑结构中的非确定性因素,文献[13]利用集对理论,融合网络中的确定和不确定关系以及局部和全局信息实现了符号预测. 总体而言,基于节点相似性的链路预测算法简单快捷,且通常能达到较高的预测准确率,但针对稀疏网络及负链接的预测效果往往欠佳. 第二类算法主要通过将符号网络转化为矩阵,利用信任传播模型、矩阵分解或填充来完成符号预测,代表性研究成果有:文献[14]首次提出将符号预测问题转化为低秩矩阵分解和填充问题,先将 $n\times n$ 矩阵分解为 $n\times k$ 和 $k\times n$ 矩阵的乘积,再通过逐点误差来测量结果矩阵和原矩阵间的误差,最终达到了较好的符号预测效果. 文献[15]在矩阵分解损失函数中运用了成对经验误差并引入了拉格朗日乘子,通过随机梯度下降算法求解符号预测结果. 文献[16]提出带偏置的低秩矩阵分解模型,将邻居节点的出边和入边符号作为偏置信息

引入模型，提高了符号预测精度. 文献[17]提出一种基于投影非负矩阵分解的框架，通过嵌入网络结构和用户属性的无监督学习实现了负链接预测. 针对大型符号网络的链路预测，文献[18]提出基于异步的分布式随机梯度下降算法的矩阵分解模型，在大大降低参数空间大小的同时提高了计算效率. 总体而言，基于矩阵处理的符号预测方法计算复杂度较高，且模型评价难度大，因此限制了此类方法在大型网络中的实际应用. 近几年，相关学者利用深度学习机制对基于卷积神经网络、循环神经网络等的链接表示与预测方法也进行了研究[19]，利用节点间的局部拓扑结构构建有序节点序列，并使用节点向量表达生成潜在链接的矩阵表示[20]，最后基于神经网络运算提取节点序列中节点对的多层隐含关系，实现链路预测[21]. 但此类算法大多关注的是传统社会网络中链接建立的可能性研究，针对符号网络中的链接与符号预测的研究相对较少.

综上所述，针对符号网络中的链接预测与符号预测双重目标，如何有效挖掘网络图的局部与全局特征，在保证算法效率的前提下提高预测的正确性，尤其是负链接以及拓扑结构特殊的符号网络的预测，是一个值得思考的问题. 基于此，本文在考虑连接两节点的路径(包括路径长度及数量)、路径上的中间节点(包括一阶和二阶邻居节点的数量、度数)以及连边符号等信息的基础上，综合引入共同邻居节点的聚集系数以及基于结构平衡环的符号影响力的概念定义两节点的相似性. 该方法能更全面地捕获符号网络的拓扑结构特征对于两节点间的链接建立的可能性以及符号类型的影响程度，既能保证算法执行效率，又能提高预测的准确性. 本文在多个经典符号网络数据集上对所提方法进行了验证，实验结果也表明了该方法对于常规和稀疏的大型符号网络，以及拓扑结构特殊的小型网络的链接预测以及符号预测的有效性和更高的预测准确性.

2　预备知识

根据节点间的链接是否带方向，可将符号网络分为有向和无向符号网络，本文关注无向符号网络中的链路预测研究. 一个无向符号网络通常被形式化表示为 $G=(V,E,S)$，其中，$V=\{v_1,v_2,\cdots,v_n\}$ 表示节点集，$E=\{e(i,j)\mid v_i,v_j\in V,i\neq j\}$ 表示边集，$S=\{\mathrm{sign}(i,j)\mid v_i,v_j\in V,i\neq j\}$ 表示符号集合，取值如下.

$$e(i,j)=\begin{cases}1, & <v_i,v_j>\in E\\ 0, & <v_i,v_j>\notin E\end{cases}$$

$$\mathrm{sign}(i,j)=\begin{cases}+1, & e(i,j)=1\text{ 且为正链接}\\ -1, & e(i,j)=1\text{ 且为负链接}\\ 0, & e(i,j)=1\text{ 但符号未知或}\\ & e(i,j)=0\end{cases}$$

2.1　结构平衡理论

Heider[22]源于社会心理学提出的用户间结构关系的平衡模型为无向符号网络的结构分析提供了理论基础，它最初是针对三角形的平衡性分析开始，如图 1 所示. 根据该理论，若无向符号网络中一个闭合环上所有边的符号之积为正，则该环为结构平衡环，否则为非平衡环. 众多研究者在社会媒体网站中的实证研究表明，真实网络中平衡的三元环数目远大于不平衡的三元环数目，且随时间推移平衡三元环所占的比例日益增加[23]，像 Epinions 和 Slashdot 这类大型符号网络的平衡指数分别达到了 89.6%和 86.2%[24]. 2010 年，Leskovec 等[4]首次将结构平衡理论应用于符号预测问题. 目前，该理论的一些基本规律已被广泛应用于符号网络链路预测算法研究[25]. 针对符号网络的预测研究，一方面要分析未知链接或缺失符号的已有链接的符号属性，即符号预测. 通常依据结构平衡理论，力图使两个目标节点所在的环能最大限度地增强网络的结构平衡性.

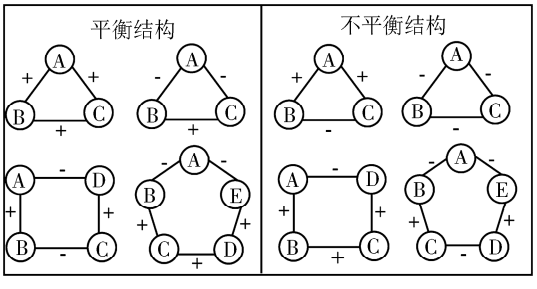

图 1　结构平衡理论示意图

Fig. 1　Sketch of structural balance theory

2.2　经典的相似性指标

针对符号网络的预测研究，除了完成符号预测之外，还要分析两个尚未相连的节点间存在或建立链接的可能性，即链接预测. 通常认为两节点间相似性越高，两者存在或建立链接的可能性越大. 总体而言，经典的相似性指标有 CN、Jaccard、AA、LP、Katz 等，如下式所示.

$$S_{xy}^{\mathrm{CN}}=|\Gamma(x)\cap\Gamma(y)| \tag{1}$$

$$S_{xy}^{\mathrm{Jaccard}}=\frac{|\Gamma(x)\cap\Gamma(y)|}{|\Gamma(x)\cup\Gamma(y)|} \tag{2}$$

$$S_{xy}^{AA}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{\log k(z)} \tag{3}$$

$$S_{xy}^{LP}=(A^2)_{xy}+\alpha(A^3)_{xy} \tag{4}$$

$$S_{xy}^{Katz}=\sum_{l=1}^{\infty}\beta^l\times|paths_{xy}^{<l>}|=\beta A_{xy}+\beta^2(A^2)_{xy}+\beta^3(A^3)_{xy}+\cdots \tag{5}$$

上式中,S_{xy}代表节点 x 与节点 y 的相似度;$\Gamma(x)$代表节点 x 的邻居集合;$k(x)$代表节点 x 的度;符号"| |"代表集合的大小;α为调节三阶路径对相似性影响程度的参数;A 为 G 的邻接矩阵;$path_{xy}^{<l>}$代表所有连接节点 x 与节点 y 的长度为 l 的路径的集合;β^l代表长度为 l 的路径的阻尼因子.

2.3　预测准确度评价方法

衡量链路预测算法准确性的基本方法是,给网络图对应的全集 U 中所有没有连边的节点对$<x,y>$按其建立链接的可能性赋予一个分数值 S_{xy},然后根据实际网络的演化情况观察哪些节点对之间出现了新的连边,并与预测结果进行比较以判断预测的准确性.然而,社会网络的动态性使得实验中无法预知或观察到连边何时会出现,故而无法将预测结果与网络真实演化结果进行比较.因此,许多链路预测研究均以网络在某个时刻的瞬时快照为研究对象.针对传统无符号网络的静态快照的链路预测方法,常用的预测准确度的评价指标有 AUC 和 Precision[26].实验中,为衡量算法预测结果的准确性,需将已知的边集 E 划分为训练集 E^{tr} 和测试集 E^{te}.通常采用 K 折交叉验证法[27]进行划分,每次取其中一个子集作为 E^{tr},其余 $K-1$ 个子集形成 E^{te}.大量实验表明,十折交叉验证在计算量和性能间达到了最好的折中[28]即,针对每个数据集独立进行 10 次划分,保证每次划分中 $E^{tr}\cup E^{te}=E$ 且 $E^{tr}\cap E^{te}=\Phi$ 且 $|E^{tr}|=9|E^{te}|$.

2.3.1　链接预测准确率评价指标 AUC′　针对符号网络中的链接预测,本文所提算法计算所得两节点间总相似度有正有负,其绝对值代表了两节点存在或建立链接的概率,其正负代表了被预测链接的符号类型.故而,本文对 AUC[26]进行调整得到新的指标 AUC′,如式(6)所示.

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

实验中,计算每次随机从 E^{te} 中选择的边和从实际不存在的边集 $E^{tn}(E^{tn}=U-E)$ 中选取的边对应的节点对的相似性得分,只有在两者符号相同时才进行比较.若 E^{te} 中所取边对应的节点间相似度绝对值大于 E^{tn} 中所取边对应的节点间相似度,则 $\bar{n}'$ 加 1;若两者绝对值相等,则 $\bar{n}''$ 加 1;若两者符号不同,则放弃本次计算,重新选取边;$\bar{n}$ 为 10 次独立实验中每组实验的总次数,本文取值 2 万次.

2.3.2　符号预测准确率评价指标 Precision′　针对符号网络中的符号预测,每次随机从测试集中取一条边作为待测边,假定其不存在,之后基于算法计算得到待测边的符号预测结果,并与真实的符号类型进行比较,以此评价符号预测准确性,相应的指标有 TP、FP、TN、FN、TPR(又称 Recall)、TNR(又称 specificity)、Precision、Accuracy、F_1-score 等[29].符号网络的符号预测需评价正负符号预测准确性的综合指数,相关研究显示[12,17,30],绝大多数真实符号网络中正链接数与负链接数的比值超过 4∶1,也即实验中选取的待测边是正链接的概率远高于负链接.故而,本文实验中融合上述指标进行调整,为正链接的符号预测结果赋予权重 1,为负链接的符号预测结果赋予权重 0.5,得到调整后的指标 Precision′,用于综合评价符号预测准确性,其定义如式(7)所示.

$$Precision'=\frac{TP+0.5\times TN}{(TP+FN)+0.5\times(TN+FP)} \tag{7}$$

3　融合聚集系数与符号影响力的链路预测方法

在考虑符号网络的局部拓扑信息时,CN、RA、AA 指标没有考虑到待测节点对的共同邻居节点的聚集系数对于两者的相似性影响.如图 2,对于节点对$<X,Y>$而言,图 2(a)和(b)中节点 X 与 Y 的度数、两者的共同邻居数目以及共同邻居的度数都相同,且共同邻居节点 B 的聚集系数也相同,但共同邻居节点 A 的聚集系数不同.显然,图 2(b)中共同邻居节点 A 对于节点对$<X,Y>$的相似性贡献更大.针对该情况,综合引入共同邻居聚集系数 CNCC(Common Neighbor Clustering Coefficient)全面衡量两节点的共同邻居的属性特征对于两者的相似性贡献.此外,对真实符号网络拓扑特征的相关研究显示,符号网络中正负链接的分布不均衡,且负链接具有更重要的作用.因此,在进行符号预测时,连接两节点的路径上正链接数目也远超过负链接数目,也即由于正负链接数目比例的不同,正负符号在结构平衡环中对于目标节点对的符号影响也不同.为此,引入符号影响力 SI(Sign Influ-

ence)的概念，对多步长路径上的符号类型赋予不同的权重，以更精确地衡量多条路径对于两节点间所建链接的符号类型的影响程度.

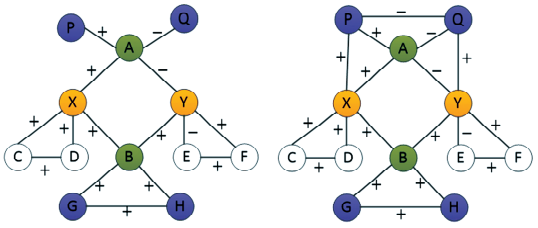

图 2　基于共同邻居聚集系数的相似性定义示意图

Fig. 2　Sketch map of similarity definition based on common neighbor clustering coefficient

基于以上思考，我们提出 CNCC_SI 算法. 在考虑基于局部路径信息的三元环对节点的相似性影响时，引入共同邻居聚集系数，综合考虑两节点的度、共同邻居数目、共同邻居的度及聚集系数的贡献；在考虑基于全局路径信息的平衡环对节点的相似性影响时，引入符号影响力，综合考虑连接两节点的路径上中间传输节点的度数以及过渡链接的符号类型的贡献；考虑到高阶步长的路径信息计算复杂度较高，根据文献[31]和文献[32]的研究结果，本文研究中利用了连接两节点的步长为 2 和 3 的路径信息分别定义了节点对的二阶相似性和三阶相似性，以达到预测准确性和计算效率上较好的均衡.

3.1　相关定义

为描述方便，对变量及符号表示如表 1.

表 1　CNCC_SI 算法相关变量定义及符号说明

Tab. 1　Variables definition of CNCC_SI algorithm

符号表示	描述
$G=(V,E,S)$	无向无权符号网络图 G
V	节点集. $V=\{v_1,v_2,\cdots,v_n\}$, $\|V\|=n$
E	边集. $E=\{e(i,j)\|v_i,v_j\in V,i\neq j\}$, $\|E\|=m$
S	符号集. $S=\{\mathrm{sign}(i,j)\|v_i,v_j\in V,i\neq j\}$
$k(x)$	节点 v_x 的度数
$N_1(x)$	节点 v_x 的一阶邻居节点集合
$N_2(x)$	节点 v_x 的二阶邻居节点集合
$CNCC^z_{<x,y>}$	节点对$<v_x,v_y>$的共同邻居节点 v_z 的聚集系数
$T^z_{<x,y>}$	v_z 邻居节点间的连边数且 $v_z\in N_1(x)\cap N_1(y)$
$\mathrm{SCN}_1(x,y)$	$<v_x,v_y>$基于一阶共同邻居的相似性得分
$\mathrm{SCN}_2(x,y)$	$<v_x,v_y>$基于二阶共同邻居的相似性得分
$\mathrm{SCN}(x,y)$	$<v_x,v_y>$的总相似性得分
$SimPath^l_{<x,y>}$	$<v_x,v_y>$基于路径 l 的相似性得分

CN、AA 和 Jaccard 指标在计算节点相似性时只考虑了共同邻居节点的数量或者度数，当这两个特征都相同时无法区分聚集系数不同的邻居节点对于两节点相似性贡献的差异. 基于此，引入节点对$<v_x,v_y>$的共同邻居 v_z 的聚集系数，记作 $CNCC^z_{<x,y>}$，如式(8)所示，用来综合衡量共同邻居节点的属性对于两节点的相似性贡献.

定义 1　共同邻居的聚集系数.

$$CNCC^z_{<x,y>}=\frac{2\times T^z_{<x,y>}}{k(z)\times[k(z)-1]} \tag{8}$$

为提高预测准确率，CNCC_SI 算法综合考虑两节点的度数、一阶共同邻居的度数和聚集系数、连边符号等局部结构特征对于两者的相似性影响. 设 $G=<V,E,S>$, $\forall v_x,v_y\in V$ 且 $\mathrm{sign}(x,y)=0$，基于结构平衡理论，定义节点对基于一阶共同邻居节点的相似性得分，记作 $\mathrm{SCN}_1<x,y>$，如式(9)所示.

定义 2　两节点基于一阶共同邻居的相似性得分.

$$\mathrm{SCN}_1(x,y)=\sum_{|l|=2}\mathrm{SimPath}^l_{<x,y>}=\sum_{z\in N_1(x)\cap N_1(y)}\frac{\mathrm{CNCC}^z_{<x,y>}\times\mathrm{sign}(x,z)\times\mathrm{sign}(z,y)}{k(z)} \tag{9}$$

针对符号网络中正负链接数不均衡的特性，在高阶路径中引入符号影响力的概念，为负链接赋予较小的权重，为正链接赋予较大的权重，来综合度量不同路径上基于结构平衡环的符号预测结果对目标链接最终符号类型的影响，记作$\mathrm{SIPath}^{|l|=3}_{<x,y>}$，如式(10)所示.

$$\mathrm{SIPath}^{|l|=3}_{<x,y>}=\begin{cases}3\alpha, \mathrm{sign}(x,p)+\mathrm{sign}(p,q)+\mathrm{sign}(q,y)=3\\ 3\beta, \mathrm{sign}(x,p)+\mathrm{sign}(p,q)+\mathrm{sign}(q,y)=-3\\ 2\alpha+\beta, \mathrm{sign}(x,p)+\mathrm{sign}(p,q)+\mathrm{sign}(q,y)=1\\ \alpha+2\beta, \mathrm{sign}(x,p)+\mathrm{sign}(p,q)+\mathrm{sign}(q,y)=-1\end{cases} \tag{10}$$

式(10)中，$l=v_xe(v_x,v_p)v_pe(v_p,v_q)v_y$ 为连接 v_x 与 v_y 的长度为 3 的路径；v_p 和 v_q 为路径 l 上的两个中间节点，即 $v_p\in N_1(x)\cap N_2(y)$, $v_q\in N_1(y)\cap N_2(x)$；α 代表路径 l 上正链接的权重，设为 1；β 代表路径 l 上负链接的权重，设为 0.5.

定义 3　路径 l 基于平衡环的符号影响力.

基于上述定义,利用连接两节点的步长为 3 的路径信息定义两节点基于二阶共同邻居的相似性得分,记作 $SCN_2<x,y>$,如式(11)所示.

定义 4　两节点基于二阶共同邻居的相似性得分.

$$SCN_2(x,y)=\sum_{|l|=3}SimPath^{l}_{<x,y>}=\sum_{|l|=3}\frac{SIPath^{|l|=3}_{<x,y>}\times sign(x,p)\times sign(p,q)\times sign(q,y)}{k(p)+k(q)-1} \tag{11}$$

将不相连的两节点间的总相似度定义为两节点基于一阶共同邻居和二阶共同邻居的相似性得分之和,记作 $SCN(x,y)$,如式(12)所示.$|SCN(x,y)|$代表节点 v_x 与 v_y 建立链接的可能性大小,链接的符号类型与 $SCN(x,y)$的符号类型相同.

定义 5　节点对总相似性得分.

$$SCN(x,y)=\sum_{2\leqslant|l|\leqslant3}SimPath^{l}_{<x,y>}=SCN_1(x,y)+SCN_2(x,y) \tag{12}$$

3.2　算法实现

Algorithm: CNCC_SI
Input: $G=(V,E,S)$
Output: $SCN(x,y)$ and $sign(x,y)$
Begin
1) Read Dataset File
2) For each v_x, $v_y\in V$ do
3) IF $e(x,y)=0$ or $e(x,y)=1\wedge sign(x,y)=0$
4) {Find all paths where $|l|=2$, Calculate $SCN_1(x,y)$
5) Find all paths where $|l|=3$, Calculate $SCN_2(x,y)$
6) Calculate $SCN(x,y)$}
7) If {$SCN(x,y)>0$ then $sign(x,y)=+1$
8) Else $sign(x,y)=-1$}
9) Output $sign(x,y)$
10) Sort $|SCN(x,y)|$ and get Top k $<v_x,v_y>$
End

4　实验与分析

4.1　数据集

采用符号网络研究中常用的 3 个经典大规模真实数据集 Epinions、Slashdot 和 Wikipedia,以及 3 个小型数据集进行实验,基本信息见表 2. 所选 3 个小型数据集拓扑结构(正负链接数的比例、节点的度分布特征等)都较为特殊,其中 CRA 和 FEC 是仿真数据集,Gahuku Gama Subtribes(记作 GGS)是真实的符号网络数据集.

表 2　数据集基本特征

Tab. 2　Basic characteristics of datasets

数据集名称	Epinions	Slashdot	Wikipedia	CRA	FEC	GGS
节点数$\|V\|$	131 828	791 20	138 592	36	28	16
边数$\|E\|$	840 799	515 397	740 106	74	42	58
正边占比/%	78.7	77.4	85	93.2	71.4	50
负边占比/%	21.3	22.6	15	6.8	28.6	50
节点平均度数	12.76	13.02	10.78	4	3	7.25
平均最短路径	3.16	3.57	4.01	3.53	3.16	1.54
平均聚集系数	0.19	0.08	0.07	0.47	0	0.54

4.2　实验结果及分析

针对符号网络中的符号预测,文献[6]中所述 CN-Predict 和改进后的 ICN-Predict 是经典的符号预测算法. 针对符号网络中的边值预测(包括链接预测与符号预测),PSNBS 算法[10]是较为经典的算法. 为衡量本文所提算法对符号网络链路预测的准确性,采用十折交叉法划分实验数据集,并以前文所述 AUC'、Precision'等为评价指标,与上述 3 个经典算法分别进行了链接预测与符号预测准确性的实验对比.

4.2.1　基于 AUC′的链接预测准确率实验结果

以 AUC′为评价标准,将所提算法与 PSNBS 算法进行了链接预测准确性的对比分析,结果如图 3 所示,图中显示的是 10 次独立实验的平均值. 且针对前 5 个数据集,图中显示的 PSNBS 实验结果为该算法中步长影响因子 λ 分别取 0.6、0.9、0.8、0.9 和 0.8 时所得到的算法的最高预测准确率.

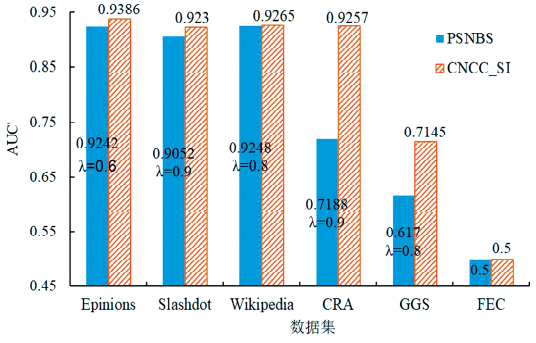

图 3　基于 AUC′指标的链接预测实验结果

Fig. 3　Link prediction results based on AUC′

从图 3 可清晰看出,本文所提算法在 3 个大型

经典符号网络以及小型仿真网络 CRA 中均得到了较好的预测效果，链接预测准确率均高于 PSNBS 算法. 尤其针对正负链接数目分布不均衡的小型网络 CRA(接近 14∶1)，所提算法链接预测准确率较 PSNBS 有较大幅度提升.

对于 GGS 网络，算法预测准确率相比前四个数据集相对较低. 该网络描述了新几内亚高地 16 个子部落(节点)之间真实的政治联盟和对立关系，其拓扑结构较为特殊，如图 4 所示. 16 个子部落根据同盟与敌对关系形成了 3 个小的社区(团体)，同一社区内节点间都为正向的同盟关系，不同社区的节点间均为负向的对立关系，16 个节点的度数以及聚集系数的分布情况分别如图 5 和图 6 所示，58 条边中正负链接比例为 1∶1. 针对正负链接数量完全相同的小型数据集，CNCC_SI 算法链接预测准确率仍可达到 71%，具有较强的健壮性.

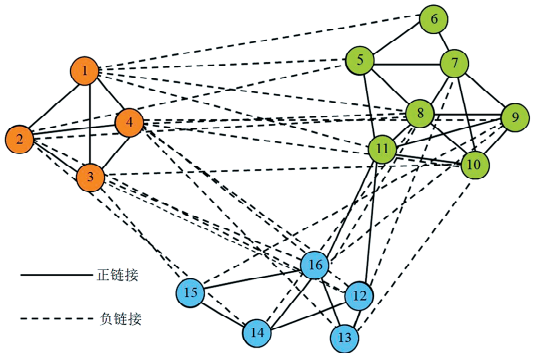

图 4　GGS 网络拓扑结构示意图
Fig. 4　Topology of GGS network

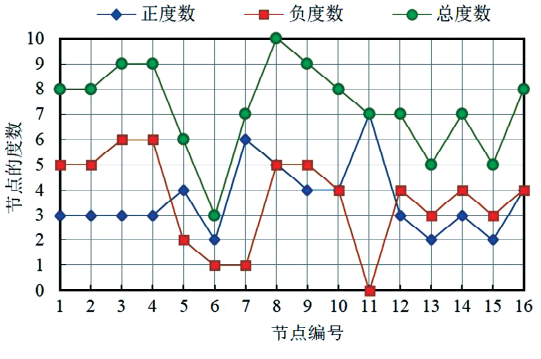

图 5　GGS 网络节点度数分布示意图
Fig. 5　Degree distribution of GGS network

由图 3 可知，针对 FEC 网络，两个算法 AUC′ 指标均为 0.5. 该仿真数据集拓扑结构也极为特殊，如图 7 和图 8 所示. 可以看出，28 个节点聚集系数都为 0 且被分为 2 类度分布情况完全相同的集合，在图 7 中用不同的颜色表示，其中有 24 个节点正度为 2、负度为 1，其余 4 个节点正度为 3、负度为 0. 在计算 AUC′ 时，绝大多数情况下从 E^{te} 中取得的链接和从 E^{tn} 中取得的链接对应的节点对的拓扑结构几乎相同，两者不相同的概率为 $C_{24}^{2}C_{4}^{2}/C_{28}^{2}C_{26}^{2}=0.0135$，也即 $\overline{n}'\approx 0$，$\overline{n}''\approx \overline{n}$，故而计算所得 AUC′ 值也应为 0.5，实验结果进一步验证了所提算法的正确性.

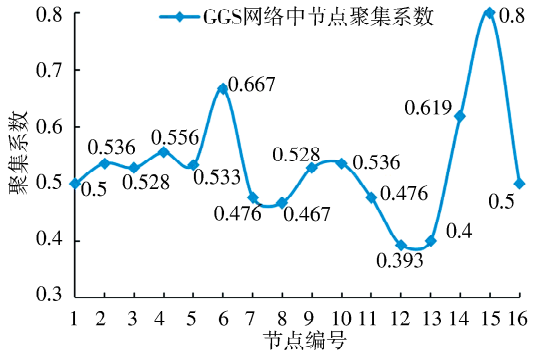

图 6　GGS 网络节点聚集系数分布示意图
Fig. 6　Clustering coefficient distribution of GGS

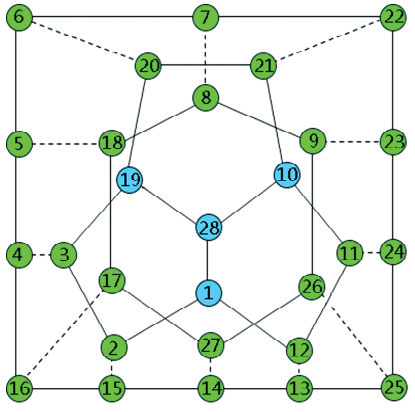

图 7　FEC 网络拓扑结构示意图
Fig. 7　Topology of FEC network

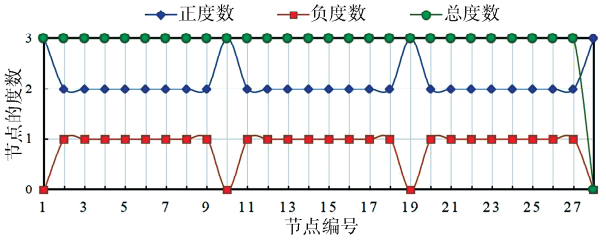

图 8　FEC 网络节点度数分布示意图
Fig. 8　Degree distribution of FEC network

4.2.2　基于 Precision′ 的符号预测准确率实验结果　以 Recall、Precison、F_1-score、Accuracy 等为评价指标，对 CNCC_SI 算法的符号预测准确性进行了验证，结果见表 3 所示. 从表 3 可以看出，所提

算法无论对于具有常规拓扑结构分布的大型真实符号网络,还是拓扑结构特殊的小型仿真及真实数据集,均达到了较好的符号预测性能,且针对负链接的预测准确率较高,具有良好的稳健性.

表 3　CNCC_SI 算法符号预测准确性实验结果

Tab. 3　Experimental results of sign prediction of CNCC_SI

	Epinions	Slashdot	Wikipedia	CRA	FEC	GGS
TP (+/+)	0.821 5	0.689 5	0.858 1	0.6	0.875	0.5
FP (+/−)	0.055 5	0.107 1	0.063 5	0.4	0.125	0.166 7
TN (−/−)	0.104	0.144 5	0.063 5	0	0	0.333 3
FN (−/+)	0.019	0.058 9	0.014 9	0	0	0
Recall	0.977 4	0.921 3	0.982 9	1	1	1
Precision	0.936 7	0.865 5	0.931 1	0.6	0.875	0.75
F_1-score	0.956 6	0.892 6	0.956 3	0.75	0.933 3	0.857 1
Accuracy	0.925 5	0.834	0.921 6	0.6	0.875	0.833 3

与此同时,以 Precision′为评价标准,将所提算法与 PSNBS 算法进行了符号预测准确性的对比实验,结果如图 9 所示,图中显示的依然是 10 次独立实验的平均值. 从图中可以看出,CNCC_SI 算法在 6 个数据集上的符号预测准确性均高于 PSNBS 算法,总体获得了较好的符号预测性能. 尤其针对 3 个拓扑结构特殊的小型符号网络,所提算法符号预测准确性均有较大幅度提升,进一步显示了 CNCC_SI 算法融合共同邻居聚集系数和符号影响力进行符号预测的正确性和有效性.

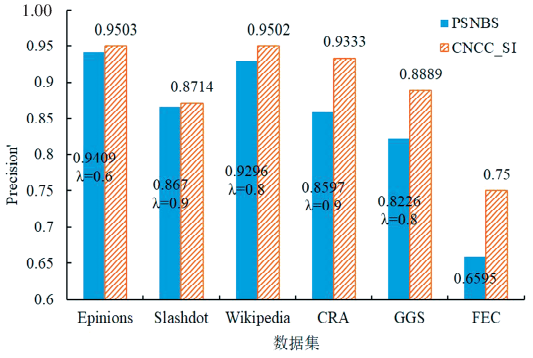

图 9　基于 Precision′的符号预测准确性实验结果

Fig. 9　Sign prediction results based on Precision′

4.2.3　可调步长参数敏感性分析　为达到预测准确性与计算复杂度上较好的均衡,本文算法将两节点基于一阶共同邻居的二步相似性得分和基于二阶共同邻居的三步相似性得分之和作为两节点的总相似度. 然而,相关研究也已表明,相对于低阶路径而言,高阶路径对节点的相似性贡献相对较低,且针对不同拓扑结构的数据集,网络的平均最短路径也不尽相同. 为此,许多基于路径结构信息的相似性计算方法为不同步长的路径赋予了可调步长参数,以区分高阶路径与低阶路径的相似性贡献程度. 本文实验中,为连接两节点的步长为 2 和 3 的路径分别赋予了可调步长参数 $\varepsilon(0.5\leqslant\varepsilon\leqslant1)$ 和 $1-\varepsilon$,并进行了预测准确率的分析,将式(12)中两节点的总相似性得分修改为式(13)所示,记作 $SCN(x,y)^{\varepsilon}$,修改后的算法记作CNCC_SI$^{\varepsilon}$.

定义 6　融合步长影响因子的相似性得分.

$$SCN(x,y)^{\varepsilon}=\varepsilon\times SCN_1(x,y)+(1-\varepsilon)\times SCN_2(x,y)\qquad(13)$$

基于式(13),在相同的条件下进行了实验,可调步长参数 ε 分别取 0.5、0.6、0.7、0.8、0.9 和 1,所提算法基于 AUC′和 Precision′评价指标的实验结果分别如图 10 和图 11 所示.

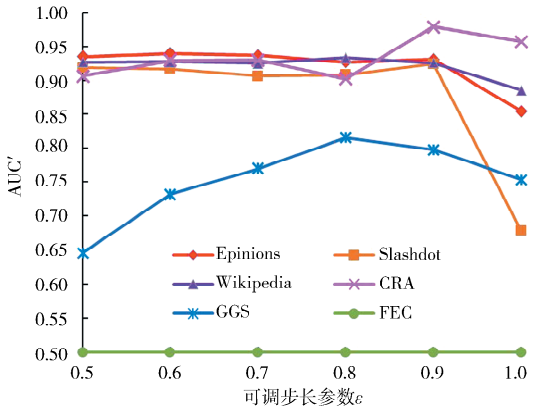

图 10　CNCC_SI$^{\varepsilon}$ 算法基于 AUC′的链接预测准确率

Fig. 10　Link prediction results of CNCC_SI$^{\varepsilon}$ based on AUC′

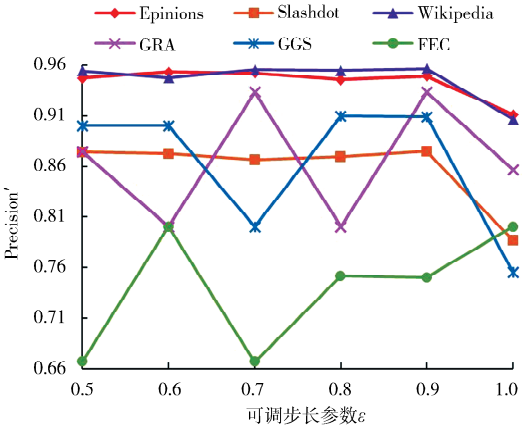

图 11　CNCC_SI$^{\varepsilon}$ 算法基于 Precision′的符号预测准确率

Fig. 11　Sign prediction results of CNCC_SI$^{\varepsilon}$ based on Precision′

从图 10 和图 11 可知，对于同一个网络，所提算法链接预测和符号预测准确率随 ε 的变化趋势是一致的. 针对 Epinions、Slashdot、Wikipedia、CRA 和 GGS 网络，链接预测准确率和符号预测准确率都是在 ε 分别取 0.6、0.9、0.8、0.9 和 0.8 时达到了最高值，又一次验证了所提算法的正确性.

4.2.4　与其他算法预测准确率对比

(1) 基于 AUC 的符号预测准确率对比. 为进一步验证本文所提算法的正确性和有效性，以文献[6]中的 AUC 为符号网络链路预测准确性的评价指标，将 CN-Predict、ICN-Predict、PSNBS(λ)、CNCC_SI、CNCC_SI$^{\varepsilon}$ 算法进行了预测结果的对比，见图 12. 在此说明，文献[6]中 AUC 评价指标定义见式(14)所示，其中 n 代表被预测的链接对数，取值为 10 000；n' 代表符号预测结果中正链接预测正确的数量，权重为 1；n'' 代表符号预测结果中负链接预测正确的数量，权重为 0.5.

$$\mathrm{AUC}^{[31]}=\frac{n'+0.5*n''}{n} \tag{14}$$

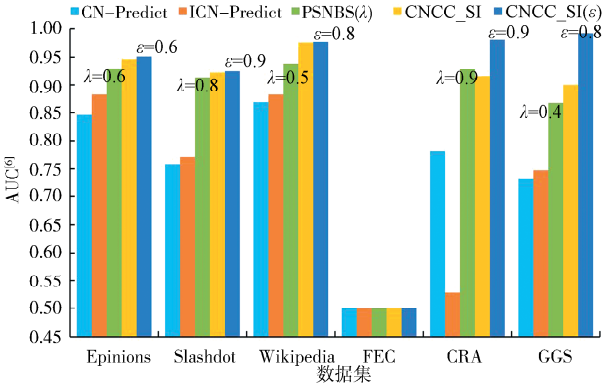

图 12　基于 AUC[6] 指标的符号网络链路预测准确率对比

Fig. 12　Link prediction results comparison based on AUC[6]

实验结果显示，针对 6 个符号网络数据集，基于可调步长参数敏感性分析的 CNCC_SI$^{\varepsilon}$ 算法预测准确率均高于其他算法，表明基于共同邻居聚类系数和符号影响力的相似性计算方法能有效解决其他算法(基于共同邻居的度、节点符号密度等)对某些拓扑结构特殊的网络存在的预测准确率较低的问题，具有更好的稳健性.

(2) 基于 Accuracy 的符号预测准确率对比. 同样地，我们以 Accuracy[29] 作为符号预测准确率的评价指标，在 3 个大型数据集上进行了实验，将所提算法与已有的经典的符号预测算法进行了对比. 例如，文献[33]所提基于谱分析的不平衡度量的符号预测方法 MOI、文献[34]所提符号网络中基于高阶环监督学习的链路预测算法 HOC、文献[35]所提将聚类之间的相对相似性定义应用于协同过滤算法的符号预测方法 CF，文献[36]所提基于谱分析聚类的矩阵分解方法 MF，以及文献[37]所述基于封闭三角结构的符号预测算法 CTMS. 以上 7 个算法在 3 个大型数据集上的实验结果如图 13 所示.

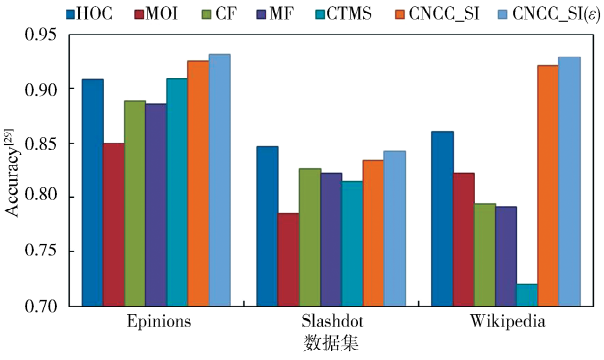

图 13　基于 Accuracy[29] 的符号网络链路预测准确率对比

Fig. 13　Link prediction results comparison based on Accuracy[29]

从图 13 可知，MOI 算法虽度量了符号网络中长度小于等于 10 的环的不平衡程度，但其符号预测准确率依然低于其他方法. CF 算法较低的符号预测准确率也说明，进行符号预测时除考虑网络的结构平衡性以外，节点的局部和全局结构特征对符号预测结果的影响更大. HOC 算法不仅学习了三元环的结构特征，同时还融入了四元环和五元环的结构特征，但符号预测效果总体上依然差于本文所提仅使用三元环和四元环的结构特征的 CNCC_SI 算法. 上述实验结果再次显示，影响符号网络中连边符号的主要因素为边的两个端点的属性特征，其次是连边所处的局部结构特征和全局结构特征. 此外，本文所提算法虽然在 Epinions 和 Slashdot 两个数据集上 Accuracy 指标(分别为 93.2% 和 84.3%)略低于文献[9]所述 SPR 模型(分别为 94.7% 和 93.8%)，但在 Wikipedia 数据集上预测正确率(92.9%)明显优于 SPR 算法(86.6%). 且 Accuracy 指标并没有区分预测正确的样本是正例还是负例，因此针对拓扑结构特殊的数据集，相关算法预测效果欠佳. 而本文所提算法能同时实现链接预测与符号预测双重目标，且关注了符号网络中正负链接的比例问题，针对各种拓扑结构的符号网络，总体而言均具有较好的预测性能.

4.3　算法复杂度分析

CNCC_SI 算法使用邻接表储存图边关系,针对无向符号网络图 $G=(V,E,S)$,节点数和边数分别为 n 和 m,算法空间复杂度是 $O(m+n)$. 算法在计算网络图中任意两节点基于一阶共同邻居的二步相似性得分时,计算复杂度为 $O(m^2n)$;算法在计算任意两节点基于二阶共同邻居的三步相似性得分时,时间复杂度为 $O(mn^2)$. 因而,算法总的时间复杂度为 $O(m^2n)$. 与其它几种算法相比,所提算法计算复杂度略微提高. 例如,文献[9]所述 SPR 模型需遍历网络中的所有边,获取任意节点对相应的邻居节点并计算节点对的相似性-相异性,进行符号预测,算法的计算复杂度为 $O(m<d>^2)$,其中,m 为网络中的边数,$<d>$为网络中节点的平均度数. 针对大规模符号网络,本文所提算法计算复杂度虽略微增加,但在达到较高预测准确率的前提下,仍可保证时间上的可行性和有效性.

5　结　论

提出 CNCC_SI 算法,结合共同邻居节点聚类系数和符号影响力分别定义了两节点基于一阶共同邻居的二步相似性和基于二阶共同邻居的三步相似性,并通过可调步长影响因子的敏感性分析对算法做进一步改进,在多个数据集上的实验结果验证了所提算法对于符号网络链接预测和符号预测的正确性、较高的预测准确率和良好的稳健性. 然而,针对具有复杂结构的符号网络(例如含时网络、多层网络、超网络等)中的链路预测,仍存在较多挑战. 针对超大规模符号网络的动态性、网络中节点及其链接关系的不确定性等,如何有效利用多维的丰富信息进行快速、准确的链路预测,设计局域化或并行化算法等,都将是下一步的研究内容.

参考文献:

[1]　程苏琦, 沈华伟, 张国清, 等. 符号网络研究综述[J]. 软件学报, 2014, 25: 1.

[2]　Liben-Nowelly D, Kleinberg J. The link prediction problem for social networks [J]. J Am Soc Inf Sci Tec, 2007, 58: 1019.

[3]　刘苗苗, 扈庆翠, 郭景峰, 等. 符号网络链接预测算法研究综述[J]. 计算机科学, 2020, 47: 21.

[4]　Leskovec J, Huttenlocher D, Kleinberg J. Predicting positive and negative links in online social networks [C]//Proceedings of the 19th International Conference on World Wide Web. New York: ACM, 2010.

[5]　Symeonidis P, Tiakas E. Transitive node similarity: prediction and recommending links in signed social networks [J]. World Wide Web, 2014, 17: 743.

[6]　佘宏俊, 胡梦缘. 基于符号网络的边值预测方法研究[J]. 武汉理工大学学报:信息与管理工程版, 2015, 37: 464.

[7]　Papaoikonomou A, Kardara M, Tserpes K, *et al*. Predicting edge signs in social networks using frequent sub-graph discovery [J]. IEEE Internet Comput, 2014, 18: 36.

[8]　张晓琴, 王秀芳. 基于结构平衡理论及 LP 算法的符号网络预测[J]. 云南民族大学学报:自然科学版, 2018, 27: 52.

[9]　Zhu X Y, Ma Y H. Sign prediction on social networks based nodal features [J]. Complexity, 2020, 2020: 1.

[10]　刘苗苗, 郭景峰, 陈晶. 相似性与结构平衡论结合的符号网络边值预测[J]. 工程科学与技术, 2018, 50: 161.

[11]　顾沈胜. 带符号复杂网络的链接预测研究[D]. 扬州: 扬州大学, 2018.

[12]　盛俊, 顾沈胜, 陈崚. 基于隐空间映射的带符号网络上的顶点分类[J], 计算机应用, 2019, 39: 1411.

[13]　Chen X, Guo J F, Pan X, *et al*. Link prediction in signed networks based on connection degree [J]. J Amb Intel Hum Comp, 2019, 10: 1747.

[14]　Hsieh C J, Chiang K Y, Dhillon I S. Low rank modeling of signed networks [C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2012.

[15]　Priyanka A, Garg V K, Narayanam R. Link label prediction in signed social networks [C]//Proceedings of the 23rd International Joint Conference on Artificial Intelligence. California, USA: AAAI Press, 2013.

[16]　苏晓萍, 宋玉蓉. 符号网络的局部标注特征与预测方法 [J]. 智能系统学报, 2018, 13: 437.

[17]　Shen P, Liu S, Wang Y, *et al*. Unsupervised negative link prediction in signed social networks [J]. Math Probl Eng, 2019, 2019: 1.

[18]　Zhang H, Wu G, Ling Q. Distributed stochastic gradient descent for link prediction in signed social networks [J]. EURASIP J Adv Sig Pr, 2019, 3: 1.

[19]　舒坚, 张学佩, 刘琳岚, 等. 基于深度卷积神经网络的多节点间链路预测方法[J]. 电子学报, 2018, 46: 2970.

[20] 张林，程华，房一泉. 基于卷积神经网络的链接表示及预测方法[J]. 浙江大学学报：工学版，2018，52：552.

[21] 王文涛，吴淋涛，黄烨，等. 基于密集连接卷积神经网络的链路预测模型[J]. 计算机应用，2019，39：1632.

[22] Heider F. Attitudes and cognitive organization [J]. J Psychol，1946，21：107.

[23] Hassan A，Abu-Jbara A，Radev D. Extracting signed social networks from text[C]//Proceedings of the TextGraphs-7 Workshop on Graph-based Methods for Natural Language Processing. Jeju，Republic of Korea：Association for Computational Linguistics，2012.

[24] Malekzadeh M，Fazli M A，Khalidabadi P J，*et al*. Social balance and signed network formation games [C]//Proceedings of the 5th Workshop on Social Network Mining and Analysis. New York：ACM Press，2011.

[25] Kunegis J. Applications of structural balance in signed social networks [J/OL]. Computer Science，(2014-2-13)[2020-08-21]. http://arXiv. org/pdf/1402. 6865v1.

[26] Lv L Y，Zhou T. Link prediction in complex networks：A survey [J]. Physica A Stat Mech Appl，2011，390：1150.

[27] Rodriguez J D，Perez A，Lozano J A. Sensitivity analysis of k-fold cross validation in prediction error estimation [J]. IEEE T Pattern Anal Mach Intell，2010，32：569.

[28] Kosir A，Odiĉ A，Tkalĉiĉ M. How to improve the statistical power of the 10-fold cross validation scheme in recommender systems[C]//Processings of 7th ACM conference on Recommender Systems. Hong Kong：ACM，2013.

[29] Huang Y J，Powers R，Montelione G T. Protein NMR recall，precision，and F-measure scores (RPF scores)：structure quality assessment measures based on information retrieval statistics[J]. J Am Chem Soc，2005，127：1665.

[30] Leskovec J，Huttenlocher D，Kleinberg J. Signed networks in social media [C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York：ACM Press，2010.

[31] 滕少华，苏庆佳，刘冬宁，等. 符号网络预测准确度及时间代价的优化[J]. 工业工程，2017，31：62.

[32] 郭景峰，刘苗苗，罗旭. 加权网络中基于多路径节点相似性的链接预测[J]. 浙江大学学报：工学版，2016，7：1347.

[33] Chiang K Y，Whang J J，Dhillon I S. Scalable clustering of signed networks using balance normalized Cut [C]//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York：ACM，2012.

[34] Chiang K，Hsieh C，Natarajan N，*et al*. Prediction and clustering in signed networks：a local to global perspective [J]. J Mach Learn Res，2014，15：1177.

[35] Javari A，Mahdi J. Cluster-based collaborative filtering for sign prediction in social networks with positive and negative links [J]. ACM T Intel Syst Tec，2014，5：1.

[36] Kunegis J，Schmidt S，Lommatzsch A，*et al*. Spectral analysis of signed graphs for clustering，prediction and visualization [EB/OL]. (2010-04-02)[2020-08-21]. https://epubs. siam. org/doi/pdf/10.1137/1. 9781611972801. 49.

[37] Khodadadi A，Jalili M. Sign prediction in social networks based on tendency rate of equivalent micro—structures [J]. Neurocomputing，2017，257：175.

引用本文格式：

中　文：刘苗苗，扈庆翠，郭景峰，等. 符号网络中融合聚集系数与符号影响力的链路预测算法[J]. 四川大学学报：自然科学版，2021，58：052003.

英　文：Liu M M，Hu Q C，Guo J F，*et al*. Link prediction algorithm in signed networks based on clustering coefficient and sign influence [J]. J Sichuan Univ：Nat Sci Ed，2021，58：052003.