

共引网络中样本量对相似性测度的选择影响

马 振¹, 贾保先²

(1.聊城大学图书馆, 山东 聊城 252000;
2.聊城大学计算机学院, 山东 聊城 252000)
✉ mazhen@lcu.edu.cn; jiabaoxian@lcu.edu.cn



摘 要:选择合适的相似性测度推断共引网络,对于提升网络的关联性和真实性具有重要意义。然而,样本量的大小对相似性测度选择的影响尚未可知。基于样本量大小的敏感性,分别使用两个常用的相似性测度 Phi 相关系数(简称 Phi)和 Ochiai 系数(简称 Och)推断共引网络,通过网络节点属性和拓扑结构对推断的网络质量进行评价。结果显示:与 Phi 相比,Och 推断的共引网络对样本量具有强鲁棒性。随着样本量的变化,Och 推断的共引网络一直都遵循小世界特性,而 Phi 则不符合此特性。研究结论可以推广到其他遵循小世界特性的事务推断网络。同时,研究可以充实网络技术研究领域的基础理论。

关键词:相似性测度;样本量;共引网络;Ochiai 系数;Phi 相关系数

中图分类号:TP393.0 **文献标志码:**A

The Impact of Sample Size on the Selection of Similarity Measure in Co-citation Network

MA Zhen¹, JIA Baoxian²

(1.Library, Liaocheng University, Liaocheng 252000, China;
2.School of Computing, Liaocheng University, Liaocheng 252000, China)
✉ mazhen@lcu.edu.cn; jiabaoxian@lcu.edu.cn

Abstract: Choosing an appropriate similarity measure to infer a co-citation network is of great significance for improving the relevance and authenticity of the network. However, the impact of sample size on the selection of similarity measure is unknown. Based on sample size sensitivity, two commonly used similarity measures, Phi correlation coefficient (referred to as Phi) and Ochiai coefficient (referred to as Och), are used to infer the co-citation network. The inferred network quality is evaluated through network node attributes and topology structure. The results show that compared to Phi, the co-citation network inferred by Och has strong robustness to sample size. As the sample size changes, co-citation network inferred by Och has always follows the small world characteristic, while Phi does not conform to this characteristic. The research conclusion can be extended to other transaction inference networks that follow small world characteristics. At the same time, research can enrich the basic theories in the field of network technology research.

Key words: similarity measure; sample size; co-citation network; Ochiai coefficient; Phi correlation coefficient

0 引言(Introduction)

共引网络的节点交互是通过两个文献节点之间的共引关系推理得出的^[1]。共引关系中的两个节点可视为二元变量,计

算共引关系实质就是计算两篇文献节点之间的共引关联性^[2]。在共引网络中,由于网络结构和节点属性取决于节点间定义的连接索引,每一对节点之间的相似度是通过事务中的节点本身

或共有邻节点的度数进行相似性计算得到的,因此选择合适的相似性测度是创建可靠、健壮的共引网络的必要条件^[3]。相似性测度通常选择 Phi 相关系数(简称 Phi)和 Ochiai 系数(简称 Och)。Phi 反映了两个变量之间的线性关系,对值为零的节点非常敏感,会导致节点之间的低重叠;而 Och 对值为零的节点具有强免疫性^[4-6]。

尽管学者对 Phi 和 Och 已经有了充分研究与讨论^[7-8],然而前人的研究未考虑样本量的大小对相似性测度选择的影响。随着样本量的变化,节点本身及其相互间的关联性都会发生变化,连接的边也可能会受到事务数变化的影响。因此,有必要研究确定哪种相似性测度不受样本量变化的影响。针对上述问题,本文选择两种常用的相似性测度,针对样本量的敏感性,比较哪种相似性测度构建的网络具有强鲁棒性,这对于提升共引网络的质量具有重要意义。

1 相关概念(Relevant concepts)

1.1 计算公式

通过定义连接节点的边构成网络。共引网络是要创建一个无向加权网络,定义包含相关节点的事务,这些事务用于解释两个节点之间是否存在连接。由 N 个事务发展而来的网络 G 用 $G=(D, E)$ 表示,其中 D 是一组 d 节点, E 是一组 e 边, $E \subseteq D \times D$ 。对称的 $d \times d$ 矩阵 $A_G=(a_{ij})_{(ij) \in D \times D}$ 称为 G 的邻接矩阵。如果 $(i, j) \in E$ 是 G 的边,则 $a_{ij} > 0$; 如果 $(i, j) \notin E$, 则 $a_{ij} = 0$; 如果 $i = j$, 则 $a_{ij} = 0$ 。由于要比较 Phi 和 Och 推断的共引网络,因此基于这两个相似性测度定义了边 a_{ij} 。

在 Phi 网络中, a_{ij} 表示 PHI_{ij} , 使用公式(1)计算得到:

$$PHI_{ij} = \frac{C_{ij} \times N - C_i \times C_j}{\sqrt{C_i \times C_j (N - C_i) (N - C_j)}} \quad (1)$$

其中, C_{ij} 是同时包含节点 i 和 j 的事务数, C_i 是包含节点 i 的事务数, C_j 是包含节点 j 的事务数。两个二元变量之间的计算公式包括整个样本量的大小(N), 第一个二元变量 i 中非零值的数量 C_i , 第二个二元变量 j 中非零值的数量 C_j , 以及两个二元变量都非零值的数量 C_{ij} 。因此, 通过公式(1)可以看出: Phi 推断的网络中节点之间的显著相关性随样本量大小 N 的变化会发生变化。

每对节点将产生一个具有统计学意义的 Phi 值连边。用公式(2)计算 t 统计量, 以 $\min(C_{ij}, C_i, C_j)$ 为自由度。在 $\alpha=0.01$, $t > 2.58$ 且 $C_{ij} > \sum C_{ij}/p$ 的条件下建立 Phi 网络, 其中 p 表示最大对数(即 $C_{ij} > 0$ 的对数), 选择 α 为最小化误差。

$$t = \frac{PHI_{ij} \sqrt{C_{ij} - 2}}{\sqrt{1 - PHI_{ij}^2}} \quad (2)$$

在 Och 网络中, a_{ij} 表示为 Och_{ij} , 使用公式(3)计算。与 Phi 不同, Och 只考虑了节点的个体引用率(C_i 和 C_j)及其共引率(C_{ij}), 其中个体引用率和共引率均小于等于样本数量, 即 $C_i, C_j, C_{ij} \leq N$ 。理论上, Och 不直接受样本量大小的影响。

$$Och_{ij} = \frac{C_{ij}}{\sqrt{C_i \times C_j}} \quad (3)$$

由于很难找到 Och 的统计学定义, 因此有学者利用相关系数检验显著性的方法确定 Och 的阈值, 具体步骤如下^[8]。

Step1: 对于每对节点, 计算总体数据集(最大样本量)的 $C_i, C_j, C_{ij}, PHI_{ij}, t$ 值和 Och_{ij} 。

Step2: 求节点对数 $q, t > 2.58$ (即连接在 $\alpha=0.01$ 时显著相关)和 $C_{ij} > \sum C_{ij}/p$, 其中 p 是 $C_{ij} > 0$ 的对数。

Step3: 在给定的 $C_{ij} > \sum C_{ij}/p$ 中, 求出对数等于 q 的 Och 阈值 S_c 。

Step4: 使用 S_c 作为阈值, 在不同的样本量中查找边的数量。

由于 Step3 中计算 S_c 的对数等于 Phi 的最大对数, 因此该方法使用 Phi 和 Och 从最大样本量中推断的网络边数相同。

1.2 节点属性和拓扑结构

评价不同相似性测度推断的网络质量可以从局部属性和全局属性两个方面考虑, 局部属性主要根据的节点度量指标衡量每个节点, 全局属性根据拓扑结构从整体上评价网络质量^[9]。网络节点的度量指标主要包括度中心性、加权重、中介中心性等。表 1 中列出了节点的度量指标定义。

表1 网络节点度量指标的定义

Tab.1 Definition of network node metrics

节点度量指标	定义或描述
边	表示节点之间的联系
网络密度	刻画网络中节点间相互连边的密集程度
度中心性	节点的度中心性越高, 则该节点在网络中就越重要
加权重中心性	考虑边权重的度中心性
中介中心性	节点担任其他两个节点之间最短路径的次数 ^[10]
紧密中心性	节点到网络中其他节点的平均最短距离 ^[10]
聚类系数	反映邻节点之间的相互连接程度 ^[11]
特征向量中心性	通过邻节点的中心性定义该节点的中心性

聚类系数反映了网络的连接紧密程度, 是衡量网络拓扑结构普遍且重要的指标^[12-14]。复杂网络的拓扑结构有很多种, 通常比较常见的有随机网络、无标度网络和小世界网络, 拓扑结构通常有随机、无标度和小世界。如果一组节点之间的连接是以一定的概率随机连接的, 则称为随机网络, 随机网络中节点的度中心性服从二项分布。当网络中少数节点起主导作用时, 称为无标度网络, 无标度网络中节点的度服从幂律分布。当网络中有多个节点簇使得节点之间的距离变小时, 称之为小世界网络, 小世界网络的度可以服从任意分布, 其聚类性能高于随机网络和无标度网络。通常使用特征路径长度和聚类系数衡量小世界网络。

2 数据采样(Data sampling)

选取 Web of Science 数据库中截至 2022 年 12 月 31 日的数据库。为了尽可能地确保研究结果的准确性, 同时考虑到工作量, 选择 *Scientometrics* 期刊, 下载 10 个数量不同的论文题录数据为样本, 涵盖了从小到大 10 个不同数量的数据样本。评估数据的质量, 清理数据(去除社论、校正、会议论文、书评、

信函、提前发表论文等),然后分别使用 Phi 和 Och 推断出 20 个共引网络。

样本包括(1)2008 年发表的 128 篇论文;(2)2013 年发表的 255 篇论文;(3)2020 年发表的 439 篇论文;(4)2021—2022 年发表的 746 篇论文;(5)2020—2022 年发表的 1 185 篇论文;(6)2019—2022 年发表的 1 485 篇论文;(7)2018—2022 年发表的 1 858 篇论文;(8)2017—2022 年发表的 2 230 篇论文;(9)2013—2022 年发表的 3 520 篇论文;(10)2008—2022 年发表的 4 535 篇论文。

创建 20 个共引网络,其中使用 $p < 0.01$ 的 Phi 推断了 10 个共引网络,使用阈值为 0.04 的 Och 推断了 10 个共引网络。例如,使用样本(8),Phi 推断的共引网络有 1 957 条边显著相关,Och 推断的共引网络有 3 894 条边显著相关。

3 实证分析(Empirical analysis)

3.1 不同样本量对网络节点的影响

如图 1 所示,随着样本量的增加,Phi 推断的网络中相关边数显著增加。例如,样本量为 1 485 时,Phi 推断的网络包含 1 265 条边;样本量为 746 时,Phi 推断的网络包含 562 条边;样本量减半时,Phi 推断的网络边数也接近减半。在 Och 推断的网络中,样本量大于 1 185 时,Och 推断的网络边数变化比较缓慢;样本量小于 1 185 时,Och 推断的网络边数变化较明显。

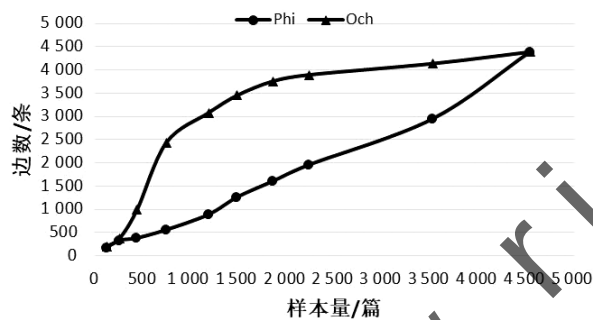


图 1 Phi 和 Och 推断的网络边数

Fig. 1 Edge numbers of Phi and Och inferred networks

网络密度如图 2 所示,Och 推断的网络密度在所有样本中基本保持不变;相反,在 Phi 推断的网络密度发生了较大的变化。这说明样本量的变化对 Och 创建的共引网络的网络密度影响较小。

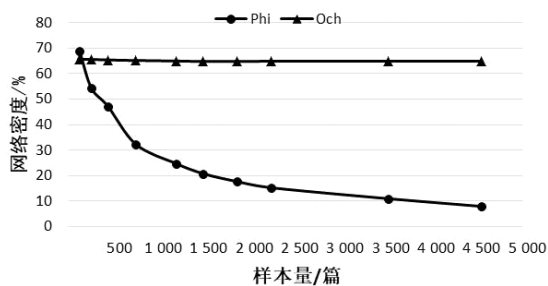


图 2 Phi 和 Och 推断的网络密度

Fig. 2 Phi and Och inferred network density

从图 3—图 8 中可以观察到,样本量大小对共引网络的其他网络度量指标的影响。在 Phi 推断的网络中,平均度中心

性、平均加权重中心性、平均中介中心性、平均紧密中心性、平均聚类系数及平均特征向量中心性都随着样本量的增加而增加。在 Och 推断的网络中,平均度中心性、平均加权重中心性、平均紧密中心性、平均聚类系数及平均特征向量中心性基本保持不变,直到样本量减少到 128 时,指标才有所波动;此外,Och 推断的网络的平均中介中心性是所有网络度量中最不一致的,在样本量减少到 1 185 时,平均中介中心性发生了明显的变化。可见,在样本量较小的情况下,平均中介中心性并不是一个有效的度量指标。

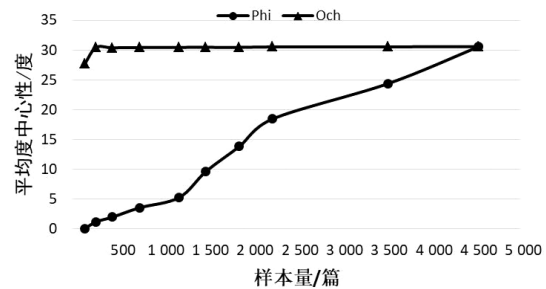


图 3 平均度中心性

Fig. 3 Average degree centrality

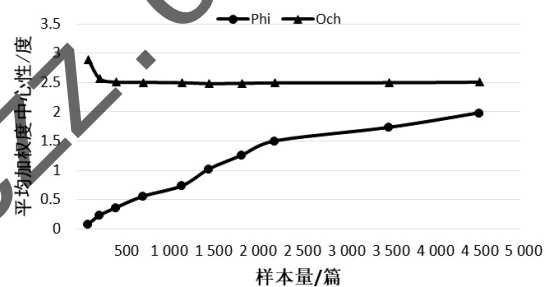


图 4 平均加权重中心性

Fig. 4 Average weighting centrality

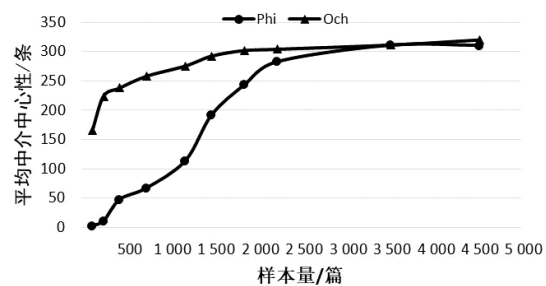


图 5 平均中介中心性

Fig. 5 Average intermediary centrality

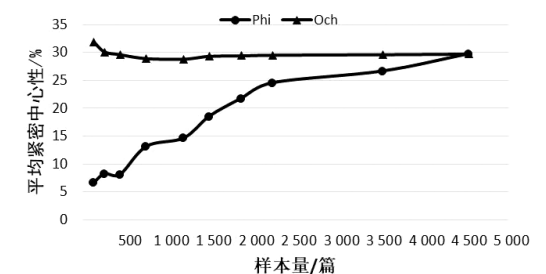


图 6 平均紧密中心性

Fig. 6 Average near centrality

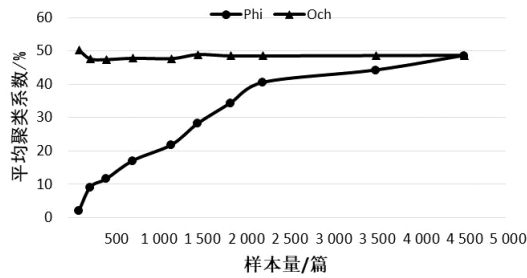


图7 平均聚类系数

Fig. 7 Average clustering coefficient

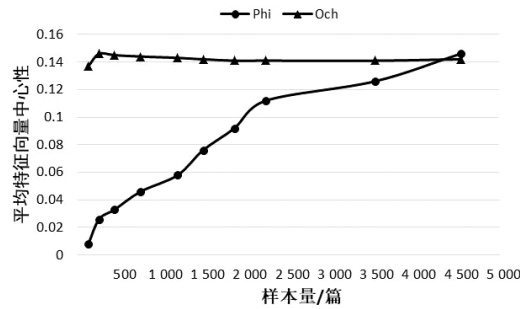


图8 平均特征向量中心性

Fig. 8 Mean eigenvector centrality

3.2 不同样本量对网络拓扑结构的影响

当网络中特征路径长度很短且存在多个节点簇时,网络具有小世界特性。特征路径长度和聚类系数是小世界网络的度量指标。节点 i 和节点 j 之间的距离 d_{ij} 定义为连接这两个节点的最短路径上边的数目。网络的特征路径长度 P 就是任意两个节点之间距离的平均值,计算公式如下:

$$P = \frac{1}{\frac{1}{2}n(n-1)} \sum_{i \geq j} d_{ij} \quad (4)$$

其中, n 是网络中的节点总数。

聚类系数 C 解释了节点形成的小集群,对于共引网络,节点的聚类系数反映了文献节点 i 的直接引用关系。节点 i 的聚类系数可以用公式(5)计算:

$$C_i = \frac{2l_i}{k_i(k_i-1)} \quad (5)$$

其中, l_i 是节点 i 的邻节点间的链路数, k_i 是节点 i 的度。

对于具有小世界特性的网络,需要满足条件 $n \gg k \gg \ln n \gg 1$,确保网络不会分割成多个子网络。此外,必须满足两个条件:首先,网络的特征路径长度(共引网络的特征路径长度用 P_{com} 表示)与相同节点数 n 和平均度 k 的特征路径长度(随机网络特征路径长度用 P_{rand} 表示)大致相同;其次,网络的聚类系数(共引网络聚类系数用 C_{com} 表示)应大于等价随机网络的聚类系数(等价随机网络聚类系数用 C_{rand} 表示)。

为了判断网络的小世界特性,本文将研究重点放在每个网络的最大连通部分上,连通部分包含直接或间接连接的最大连接节点数。例如,样本(8)创建的Phi网络的最大连接组件包

含473个节点,该最大连接组件的平均度为25,即 $n=473$ 和 $k=25$, P_{rand} 和 C_{rand} 可分别通过公式(6)和公式(7)计算得到: $P_{rand} \sim 1.91$, $C_{rand} \sim 0.053$, P_{com} 为2.452(大于 P_{rand}), C_{com} 为0.6414(大于 C_{rand})。此外,满足 $n \gg k \gg \ln n \gg 1$ ($473 \gg 25 \gg 6.16 \gg 1$)。因此,遵循小世界特性。

$$P_{rand} \sim \ln n / \ln k \quad (6)$$

$$C_{rand} \sim k/n \quad (7)$$

如图9所示,随着样本量的增加,Phi推断的网络中最大连接集的节点数也随之增加。然而,Och推断的网络中最大连接集的节点数量基本没有变化,直到样本量减少到255时才急剧减少。

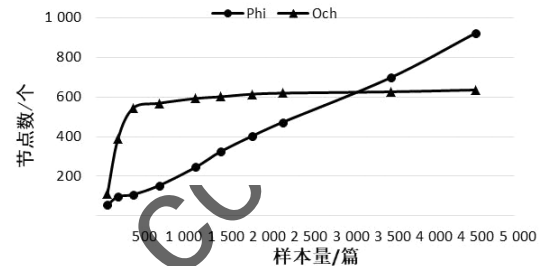


图9 最大连接集的节点数

Fig. 9 Number of nodes in the maximum connection set

关于网络特征路径长度的变化,如图10和图11所示,随着样本量的减小,Och推断的网络中的 P_{rand} 和 P_{com} 几乎保持不变,但是Phi推断的网络显示出一些不一致性。在Phi推断的网络中, P_{rand} 随着样本量的增加逐渐增加, P_{com} 在样本量增加到439之前先增加,之后随着样本量的增加而减少。

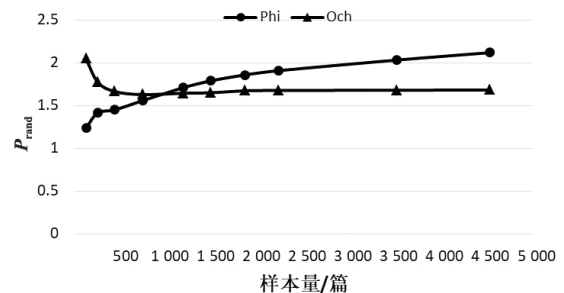


图10 等价随机网络的特征路径长度

Fig. 10 Characteristic path length of equivalent random network

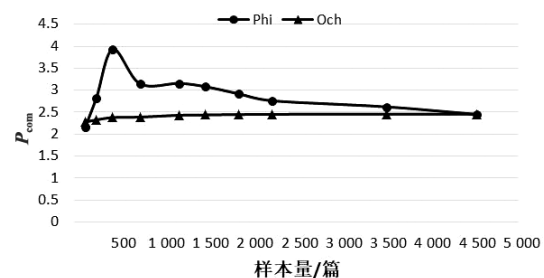


图11 共引网络的特征路径长度

Fig. 11 Characteristic path length of co-citation network

在图 12 和图 13 中可以看到,聚类系数的变化也有类似趋势:在 Och 推断的网络中, C_{rand} 和 C_{com} 基本保持不变,而在 Phi 推断的网络中, C_{rand} 随样本量的增加逐渐减少且变化差异较大, C_{com} 随样本量的变化有轻微的随机变化。

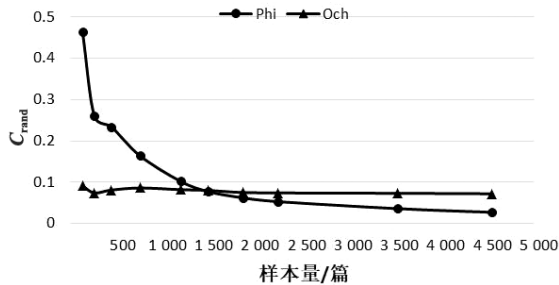


图 12 等价随机网络的聚类系数

Fig. 12 Clustering coefficient of equivalent random network

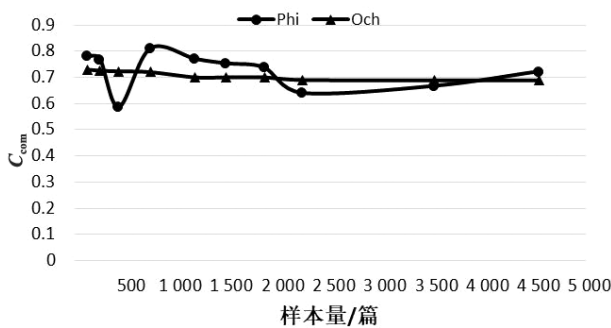


图 13 共引网络的聚类系数

Fig. 13 Clustering coefficient of co-citation network

在 Phi 推断的网络中,样本量为 4 535 时,满足小世界特性。但是,随着样本量的减少, k 和 $\ln n$ 之间的差异变小,这违反了小世界特性的条件要求,形成多个不连通的子网络。此外,在 Och 推断的网络中,小世界特性始终存在。总体而言,Och 保留了网络的整体拓扑结构,样本量较小时也遵循小世界特性,而 Phi 则不符合这一特性。

4 结论(Conclusion)

本文分析了共引网络中样本量对相似性测度选择的影响,分别使用两个常用的相似性测度 Phi 相关系数和 Ochiai 系数推断共引网络,通过网络节点属性和拓扑结构对推断的网络质量进行评价。结果表明:如果目标是寻找高度相关的节点,可以使用 Phi;在样本量较小的情况下,建议使用 Och。本文研究可以丰富网络技术研究领域的基础理论,提升推断关系网络的关联性和真实性。此外,研究结论可以推广到神经网络、语言网络、文本网络等其他遵循小世界特性的通过事务推断得出的网络。

参考文献(References)

[1] 邱均平. 文献计量学[M]. 北京:科学出版社,2019:252-255.

[2] GUILFORD J P. Psychometric methods[M]. New York: McGraw-Hill Book Company,1936:13-22.

[3] 杨利军,张良友. 期刊共被引相似性测度问题的实证研究[J]. 图书情报工作,2010,54(18):139-144.

[4] 曾守楨,骆丹丹. 基于类 Pearson 综合相关系数的概率语言 TOPSIS 多属性决策方法[J]. 系统科学与数学,2021,41(1):126-143.

[5] 高继平,丁堃,刘宇,等. 知识基础与前沿载文间的知识流动分析:以信息领域中的 Gerard Salton 为例[J]. 情报杂志,2009,28(10):98-102.

[6] 康耀红,CHANG K W. 关于 Salton 扩展布尔情报检索模型的一个注记[J]. 情报学报,2002(2):164-166.

[7] CHARTIER J F, MONGEAU P, SAINT-CHARLES J. Predicting semantic preferences in a socio-semantic system with collaborative filtering: a case study[J]. International Journal of Information Management,2020,51:102020.

[8] EGGHE L, LEYDESDORFF L. The relation between Pearson's correlation coefficient r and Salton's cosine measure[J]. Journal of the American Society for Information Science and Technology,2009,60(5):1027-1036.

[9] 孙睿,罗万伯. 网络舆论中节点重要性评估方法综述[J]. 计算机应用研究,2012,29(10):3606-3608,3628.

[10] 胡思文,李兵,何鹏,等. 一种基于 h 指数的软件网络中重要类的度量方法[J]. 小型微型计算机系统,2017,38(2):249-253.

[11] 刘向. 知识网络的形成与演化[M]. 武汉:武汉大学出版社,2014:32-33.

[12] HERNÁNDEZ SERRANO D, SÁNCHEZ GÓMEZ D. Centrality measures in simplicial complexes: applications of topological data analysis to network science[J]. Applied Mathematics and Computation,2020,382:125331.

[13] 马梦珂,倪静. 基于度值和聚类系数的跨单元调度问题优化[J]. 计算机应用研究,2021,38(9):2651-2656.

[14] 杨博,刘大有,金弟,等. 复杂网络聚类方法[J]. 软件学报,2009,20(1):54-66.

作者简介:

马 振(1984-),男,硕士,讲师。研究领域:知识网络,数据挖掘。

贾保先(1982-),男,博士,副教授。研究领域:人工智能,大数据。