A Comprehensive Study on the Relevance and Rating System of Academic Journals

关于学术期刊关联度和评级系统的综合研究

组长：张锦程

组员：陈岳，张瑞熙

2021.05.20

Email：jincheng18@mails.tsinghua.edu.cn

1.1 Background —— Garfield Model

学术期刊的分级与评价、相似性的度量与刻画、区分与聚类是一类重要的研究方向。

Garfield 模型及一些前人的研究在一定程度上解决了衡量学术期刊间关联程度的问题，他们的模型由以下式子定义期刊之间的**关联度指标**：

$$R_{i\&j} = max\left(\frac{H_{i>j}}{ref_i} \times \frac{10^6}{pap_i}, \frac{H_{j>i}}{ref_j} \times \frac{10^6}{pap_j}\right)^1 \qquad (1)$$

Garfield 模型值得借鉴的优点主要在于：

1. 在期刊体量方面，利用 $10^6/\mathrm{Pap}_i$ 因子对期刊的体量进行了**归一化**处理，使得关联程度的定义更加符合实际情况；
2. 在引用量衡量方面，利用 $\mathrm{H}_{i>\mathrm{j}}/\mathrm{Ref}_\mathrm{i}$ 衡量相关性，借鉴了前人的工作，有一定的合理性。

但是本研究认为 Garfield 模型也存在下面几个问题：

1. 在引用量衡量方面，Garfield 模型认为引用和被引是等价的，无法避免小期刊 “抱大腿” 的现象；
2. 没有考虑学术期刊刊龄带来的影响；
3. 只能聚类，而无法解释学术期刊的水平差异，适用范围较窄。

1.2 Background —— JIF Model

此外，我们还考察了现在流行的学术期刊影响因子 (Journal Influence Factor, JIF) 模型，它从总被引和总发文数的比值出发，构建了一套简单的学术期刊评级系统。JIF(IF) 如下定义：

$$JIF_i = \frac{\sum_{j \neq i} H^1_{j>i}}{pap_i} \qquad (2)$$

本研究认为 JIF(IF) 模型存在下面的问题：

1. 无法解决学术期刊间的关联程度问题；
2. 在学术期刊水平评价方面，JIF 受研究领域影响较重 (比方说生物医学期刊 JIF 普遍偏高)；
3. 无法在某一领域内部进行期刊影响力的比较。

基于上述考量，我们尝试从前人的研究基础上出发，构建一个完备的期刊聚类和评级系统：

1. 在 Task1 中，我们通过**改进 Garfield 模型关联度因子生成函数**，结合**模拟退火**算法，完成了基于关联度的期刊**聚类式分级**；
2. 在 Task2 中，我们构建了**网络模型**，并结合网络中心度分析和**投票算法**，完成了**基于影响力的期刊评级**；
3. 在 Task3 中，我们则对 Garfield 模型中的一些参数和假定进行了讨论。

2 Task 1: 基于改进 Garfield 算法的期刊关联度模型

1. 考虑刊龄影响的改进：

在计算 $R_{i>j}$ 时，用 i 期刊当年文章引用 j 期刊近 8 年出版文章的次数代替 Garfield 模型中的 $H_{i>j}$，并修改乘子 10^6 为 10^7 以便于计算；与之对应地，参考文献总量 Ref_i 也取引用近 8 年文章的总量。

选取时间跨度为 8 年的理由：

(1) 可以较完整地反映引用量随年份的变化规律；

(2) 数据的基数不会过小以引起明显的数据波动。

（2019 年 NATURE 共引用 SCIENCE 文章 2601 篇，其中 1038 篇出版于 2011 年及以前）

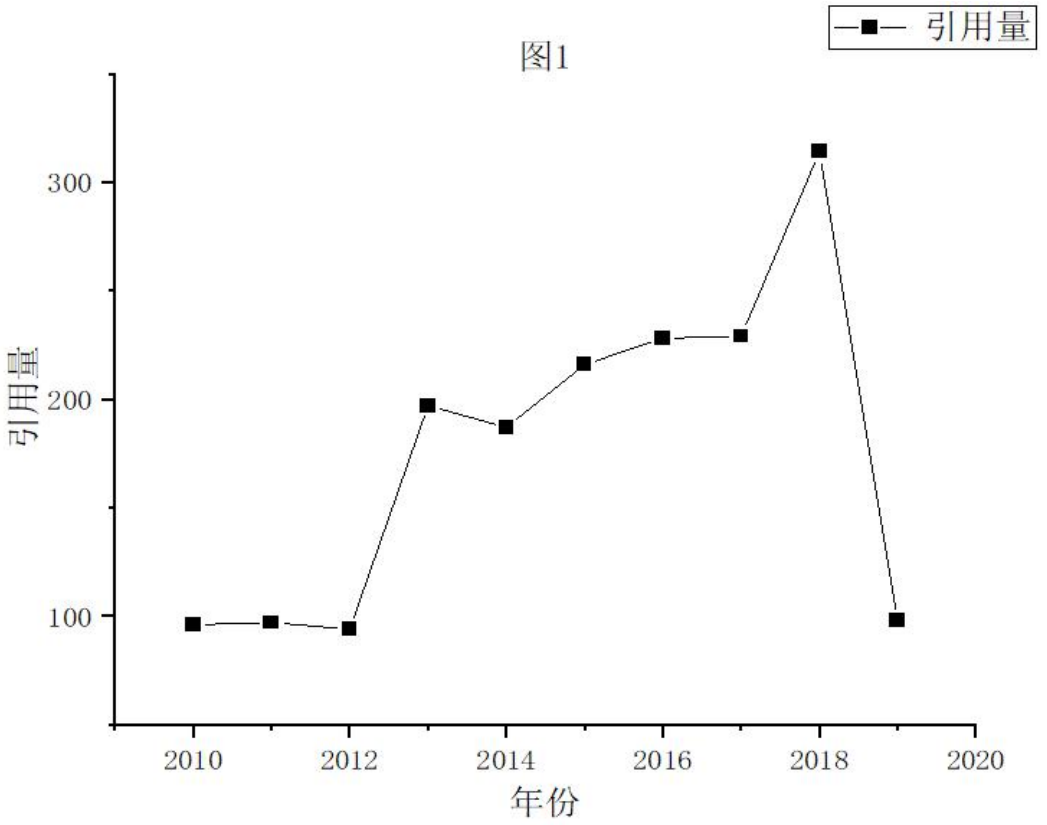

图 1. NATURE 对 SCIENCE 近 10 年引用数据

2 Task 1: 基于改进 Garfield 算法的期刊关联度模型

2. 对关联度因子 $R_{i\&j}$ 生成函数的改进：

记 $R_{i>j} = x$，$R_{j>i} = y$，$R_{i\&j} = z = z(x, y)$，我们认为改进后的 $z(x, y)$ 应满足如下四个公理：

公理1——对称公理. $z(x, y)$ 关于 x、y 对称；
公理2——单增公理. $z(x, y)$ 关于 x、y 单增；
公理3——协同公理. 对于期刊 1、2、3，若 $R_{1>2} > R_{1>3}$，则在 $R_{2>1}$，$R_{3>1}$ 增幅相同时，$R_{1\&2}$ 增幅大于 $R_{1\&3}$ 增幅。特别地，若 $z(x, y)$ 二阶连续可导，则二阶混合偏导数大于0；
公理4——有限公理. x 趋于无穷时，z 趋于有限正数，且该极限随 y 单增。

发现下面的函数满足上述公理：

$$z(x,y) = \frac{xy}{(x+y)}$$

右图为该函数与原 Garfield 模型生成函数的对比，可以看到改进后的算法**只有在 $R_{i>j}$ 和 $R_{j>i}$ 均较大时才会获得较大的关联度因子。**

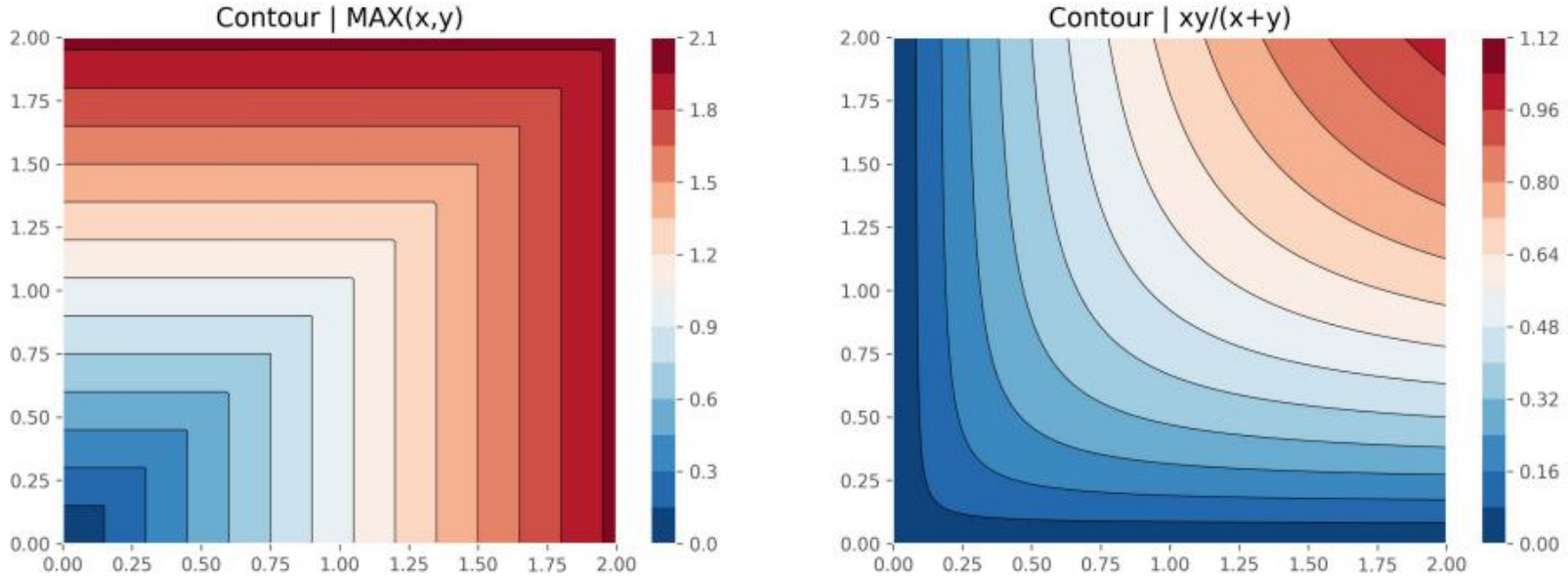

图 2. Garfield 模型和改进 Garfield 模型目标函数对比

2.1 Task 1 —— 关联度计算结果

1. 与 SCI-ADV 高关联度期刊 (排名前十)

Journal	Cited-by-SA	Citing-SA	Relatedness
SCI-ADV	114.35	114.35	57.18
SCIENCE	501.24	49.92	45.4
NATURE	413.29	41.54	37.75
NAT-COMMUN	75.57	34.43	23.65
P-NATL-ACAD-SCI-USA	90.54	30.61	22.88
NATL-SCI-REV	25.71	59.24	17.93
J-R-SOC-INTERFACE	35.63	23.18	14.04
NAT-HUM-BEHAV	26.19	19.64	11.22
PHILOS-T-R-SOC-A	20.08	22.29	10.56

2. 高关联度期刊列表的计算 (列举三个)

Journals	Closely Related Journals(Relatedness)		
GIGASCIENCE	GIGASCIENCE(931.62)	NAT-COMMUN(15.06)	NATURE(14.02)
MAEJO-INT-J-SCI-TECH MIT-TECHNOL-REV	MAEJO-INT-J-SCI-TECH(3132.83)		
NAT-COMMUN	GIGASCIENCE(15.06)	NAT-COMMUN(42.35)	NATL-SCI-REV(18.82)
NATURE	GIGASCIENCE(14.02)	J-R-SOC-INTERFACE(17.96)	NAT-COMMUN(49.46)

认为 64 个期刊间关联度的前 5% 为关系密切，由此算出关联度阈值 R=13.76，并以此作为判定关系密切的标准。

2.2 Task 1 —— 基于 SAA 的期刊聚类式分级

1. 模拟退火算法(SAA)

定义 SAA 目标函数 S 如下

$$S = \sum_{k=1}^{4} \bar{R}_k, \quad \bar{R}_k = \sum_{i \neq j} R_{i\&j} \, / E_k$$

其中 E_k 为第 k 个集合的边数，i, j 表示第 k 个集合中的期刊，**S 为该聚类方式下集合内平均关联度之和。**寻找聚类方法，**使 S 取得最大值。**SAA 算法如下：

1. **设置初温**T，**终温** T1，**退火系数** k，0<k<1，**初始状态** S，**内循环迭代次数** L
2. while $T > T_1$

   for $i = 1, 2, \ldots, L$

   在当前状态 S **附近随机选取状态** S'，**计算目标函数差值** $df = f(S') - f(S)$

   if $df > 0, S = S'$，**接受新状态**

   else，**按** Metropolis **准则以概率** exp(df/T) **接受新状态**

   $T = k * T$，**指数形式缓慢降温**

2.2 Task 1 —— 基于 SAA 的期刊聚类式分级

2. 对聚类的分级

对聚类结果中的每一类求出期刊影响因子 (JIF) 的算术平均，由高到低完成一区到四区的排序

上半区

1 区	2 区	3 区	4 区
DEFENCE-SCI-J	ANN-NY-ACAD-SCI	ACTA-SCI-TECHNOL	ADV-COMPLEX-SYST
FRACTALS	FRONT-BIOENG-BIOTECH	CHIANG-MAI-J-SCI	ADV-THEOR-SIMUL
NAT-COMMUN	INT-J-BIFURCAT-CHAOS	COMPLEXITY	AN-ACAD-BRAS-CIENC
NATURE	ISSUES-SCI-TECHNOL	CR-ACAD-BULG-SCI	ARAB-J-SCI-ENG
PLOS-ONE	J-RADIAT-RES-APPL-SC	CURR-SCI-INDIA	DISCRETE-DYN-NAT-SOC
P-NATL-ACAD-SCI-USA	NATL-SCI-REV	GIGASCIENCE	ENDEAVOUR
SAINS-MALAYS	PHILOS-T-R-SOC-A	IRAN-J-SCI-TECHNOL-A	FRONT-LIFE-SCI
SCIENCE	SCI-ADV	J-ADV-RES	GLOB-CHALL
	SCI-BULL	J-KING-SAUD-UNIV-SCI	INTERDISCIPL-SCI-REV

下半区

1 区	2 区	3 区	4 区
		JOVE-J-VIS-EXP	ISCIENCE
		J-R-SOC-INTERFACE	J-INDIAN-I-SCI
		J-TAIBAH-UNIV-SCI	J-NATL-SCI-FOUND-SRI
		MIT-TECHNOL-REV	KUWAIT-J-SCI
		NAT-HUM-BEHAV	MAEJO-INT-J-SCI-TECH
		P-NATL-A-SCI-INDIA-A	NATL-ACAD-SCI-LETT
		RES-SYNTH-METHODS	NEW-SCI
		ROY-SOC-OPEN-SCI	NPJ-MICROGRAVITY
		SCI-AM	P-JPN-ACAD-B-PHYS
		SCI-DATA	P-ROMANIAN-ACAD-A
		SCIENCEASIA	REND-LINCEI-SCI-FIS
		SCIENTIST	S-AFR-J-SCI
		SCI-PROGRESS-UK	SCI-ENG-ETHICS
		SYMMETRY-BASEL	SCI-REP-UK
		T-ROY-SOC-SOUTH-AUST	

以上模型依然存在一些不足，对于分级的方面依然不能离开 JIF 的框架，而这是 Task2 中所要解决的问题。

3 Task 2: 基于网络模型的 JournalRank 评级系统

期刊之间的引用关系是有向的，简单的例子如 A 经常引用 B，但是 B 几乎不引用A，这种情况通常意味着期刊 B 的等级水平高于期刊 A。所以，若要反映期刊之间的层级关系，我们必须利用引文数据建立**有向网络 (directed network)**。

在这个有向网络中，每个节点代表一个特定的期刊，每个有向边代表引用的过程，边的权重表示 $H_{i>j}/(pap_i \times ref_j)$ 的大小，含有 g 个期刊的引用网络可以用一个 $g \times g$ 的**连接矩阵 $\mathbf{Conn_{g\times g}}$** 来描述：

右图是本研究所采集的 Multidisciplinary Sciences下的 64 个期刊的引用网络，其中 y 轴是引用期刊，x 轴是被引期刊，亮色方块表示网络中该条连接的权重大。

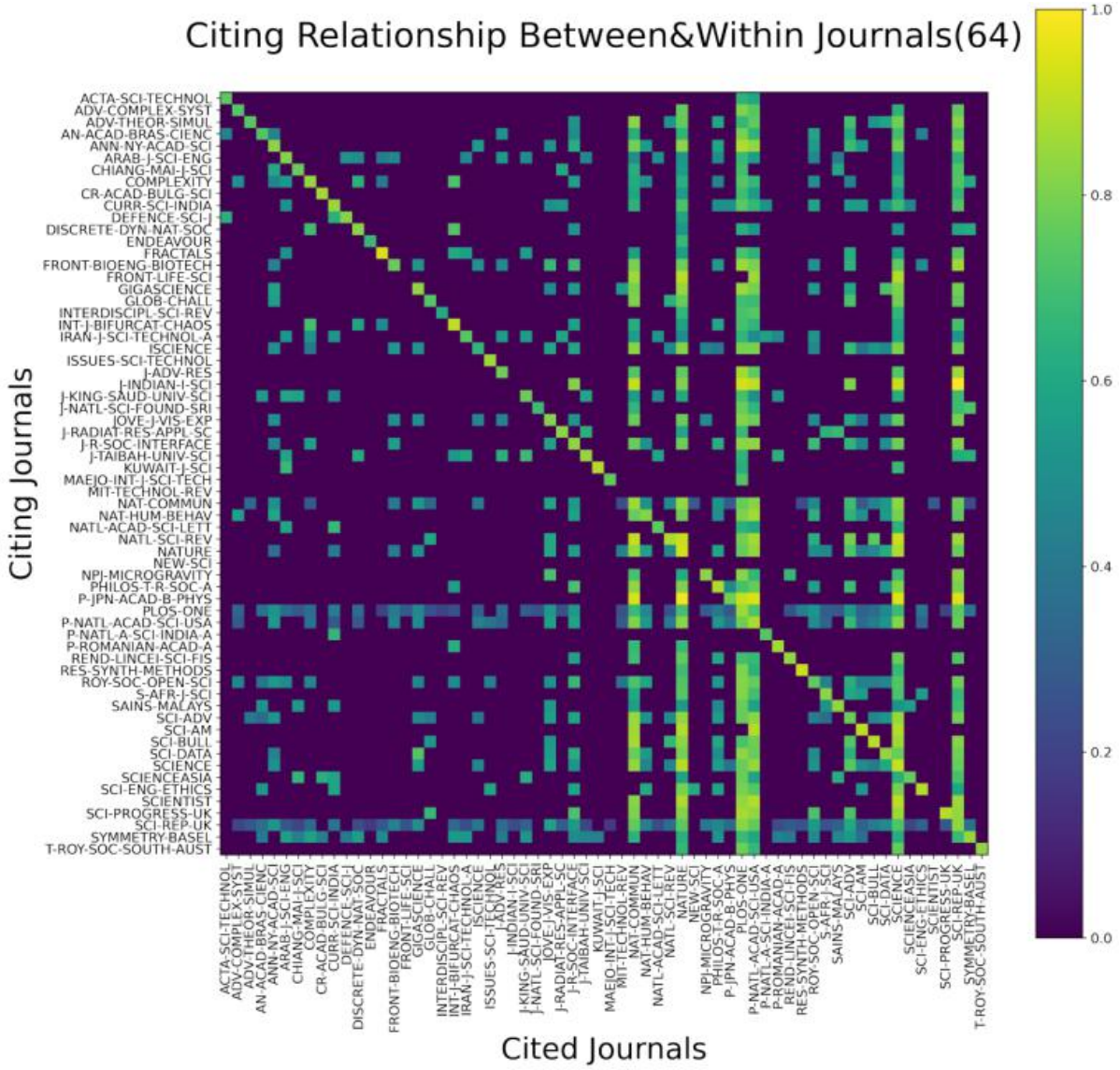

图 3. 不同期刊和期刊内部的引用网络连接矩阵

3.1 Task 2 —— PageRank & JournalRank

Task2 的评级算法主要受 **PageRank 算法**的启发，它由 **Google** 研发，能根据网页间的超链接来计算不同网页的相关性和重要性。对于本研究而言，我们可以把期刊类比为网页，网页之间的超链接关系则等价于不同期刊间的引用关系。因此，借助它可以衡量有向网络中各期刊节点的影响力。

PageRank 算法简单来说分为两步：

1. 给每个网页随机一个初始 PR 值 (下面用 PR 值指代 PageRank 值)；
2. 通过**投票算法**不断迭代，直至达到平稳分布为止。

其中投票算法如下：

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)} \qquad (7)$$

$$PR(p_i) = \alpha \sum_{p_j \in M_{p_i}} \frac{PR(p_j)}{L(p_j)} + \frac{1-\alpha}{N} \qquad (8)$$

其中，B_u 是所有链接到网页 u 的集合，L_v 则是网页 v 的外链数 (即出度 out-degree)。

而在本研究中，我们将外链数 L_v 置换为归一化后的引用数 $\mathbf{H_{i>j}/(Pap_i \times Ref_j)}$，便改良成了**用于期刊分级的 JournalRank 算法。**

3.2 Task 2 —— 中心度因子分析

为了增强投票算法结果的可解释性，我们还辅以传统的**中心度因子分析**，包括：

(1) 度中心度 (Degree centrality):

$$C_D(N_i) = \sum_{j=1}^{g} x_{ij}(i \neq j) \qquad (9)$$

(2) 中间中心度 (Betweenness Centrality):

$$C_B(N_i) = \sum_{i=1}^{g} \sum_{j=1}^{g} N_{i\&j}(between\ N_i) / \sum_{j=1}^{g} N_{i\&j} \qquad (11)$$

(3) 接近中心度 (Closeness Centrality):

$$C_C(N_i) = 1 / \sum_{j=1}^{g} d_{ij}(i \neq j) \qquad (12)$$

(4) 特征向量中心度 (Eigenvector Centrality):

$$C_E(N_i) = ce_i = c \sum_{j=1}^{g} a_{ij} ce_j \qquad (13)$$

3.3 Task 2 —— 分级结果

根据上述的 JournalRank 算法，我们计算出每个期刊在网络中的**影响力**大小，并以此为依据进行了期刊分级，其结果和对应的可视化如下所示。

表 4. Multidisciplinary Sciences 领域期刊分级情况

一区	二区	三区	四区
PLOS-ONE	J-TAIBAH-UNIV-SCI	PHILOS-T-R-SOC-A	P-JPN-ACAD-B-PHYS
NATURE	P-NATL-A-SCI-INDIA-A	SCI-BULL	ADV-COMPLEX-SYST
SCIENCE	JOVE-J-VIS-EXP	GIGASCIENCE	MIT-TECHNOL-REV
SCI-REP-UK	COMPLEXITY	SCI-PROGRESS-UK	NATL-ACAD-SCI-LETT
P-NATL-ACAD-SCI-USA	J-R-SOC-INTERFACE	CHIANG-MAI-J-SCI	NEW-SCI
NAT-COMMUN	T-ROY-SOC-SOUTH-AUST	NATL-SCI-REV	NPJ-MICROGRAVITY
ISSUES-SCI-TECHNOL	SCI-DATA	FRONT-BIOENG-BIOTECH	GLOB-CHALL
DEFENCE-SCI-J	CURR-SCI-INDIA	J-ADV-RES	FRONT-LIFE-SCI
SCI-ADV	J-KING-SAUD-UNIV-SCI	ANN-NY-ACAD-SCI	ADV-THEOR-SIMUL
INT-J-BIFURCAT-CHAOS	IRAN-J-SCI-TECHNOL-A	DISCRETE-DYN-NAT-SOC	SCIENTIST
KUWAIT-J-SCI	SCI-ENG-ETHICS	S-AFR-J-SCI	J-INDIAN-I-SCI
SAINS-MALAYS	SCIENCEASIA	J-NATL-SCI-FOUND-SRI	CR-ACAD-BULG-SCI
P-ROMANIAN-ACAD-A	ROY-SOC-OPEN-SCI	J-RADIAT-RES-APPL-SC	INTERDISCIPL-SCI-REV
FRACTALS	SCI-AM	ISCIENCE	ACTA-SCI-TECHNOL
SYMMETRY-BASEL	NAT-HUM-BEHAV	AN-ACAD-BRAS-CIENC	MAEJO-INT-J-SCI-TECH
ARAB-J-SCI-ENG	REND-LINCEI-SCI-FIS	RES-SYNTH-METHODS	ENDEAVOUR

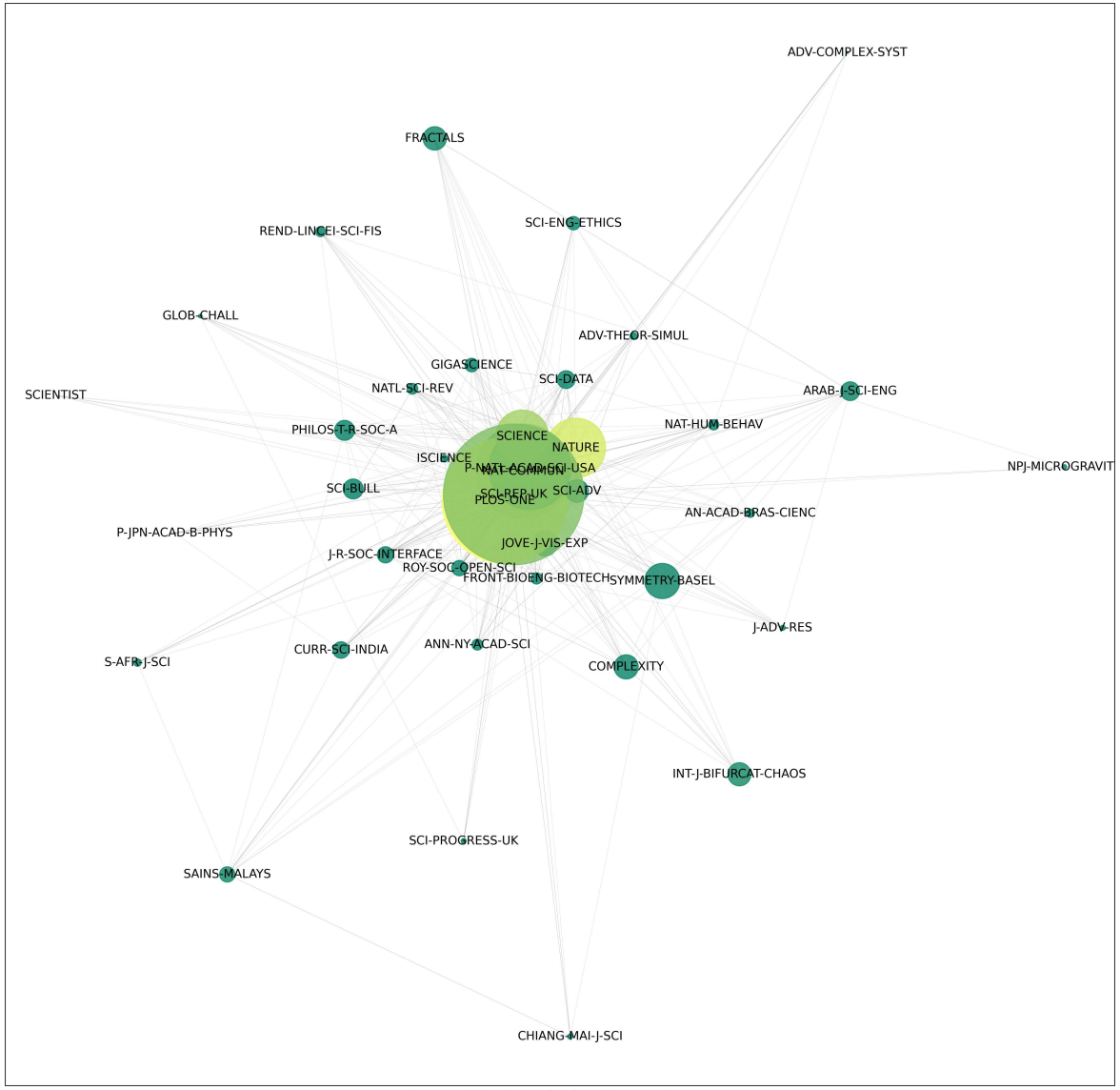

图 4. 引用网络和节点影响力可视化 (JournalRank)

3.4 Task 2 —— 结果讨论

我们还尝试从多个角度出发，来理解 JournalRank 算法的结果。图 5 和图 6 分别展示了基于 JournalRank 的结果与基于 JIF 模型的结果、基于中心度分析的结果之间的关系。可以看到它们**存在一定的相关性**，但是**对于一部分期刊，不同的评价标准会给出相差甚远的结果。**

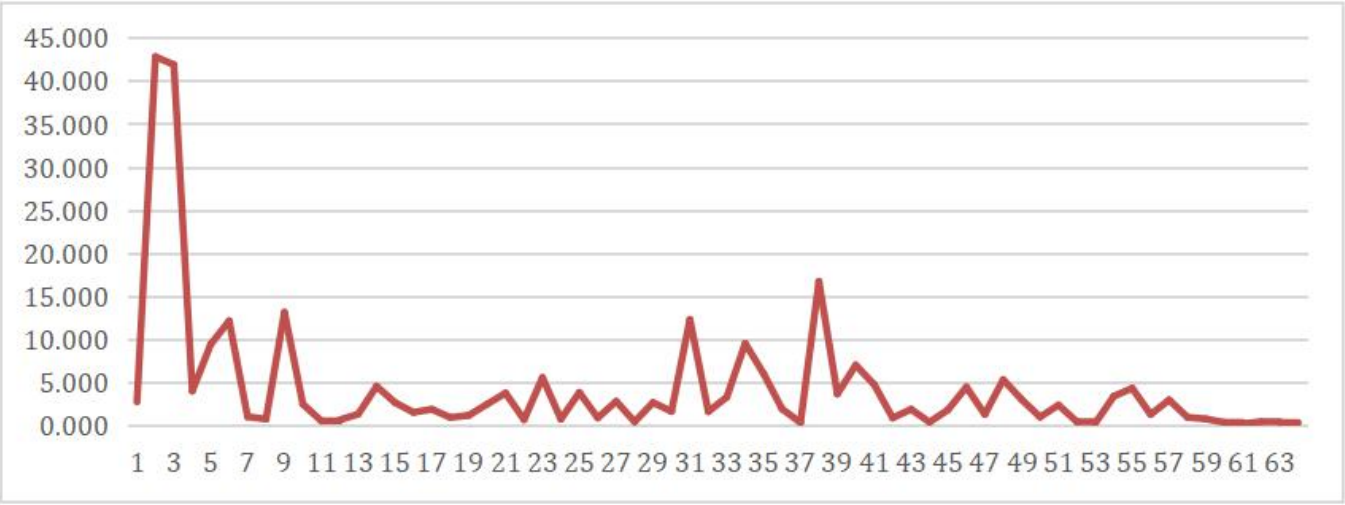

图 5. JournalRank 排名和 IF 之间的关系图

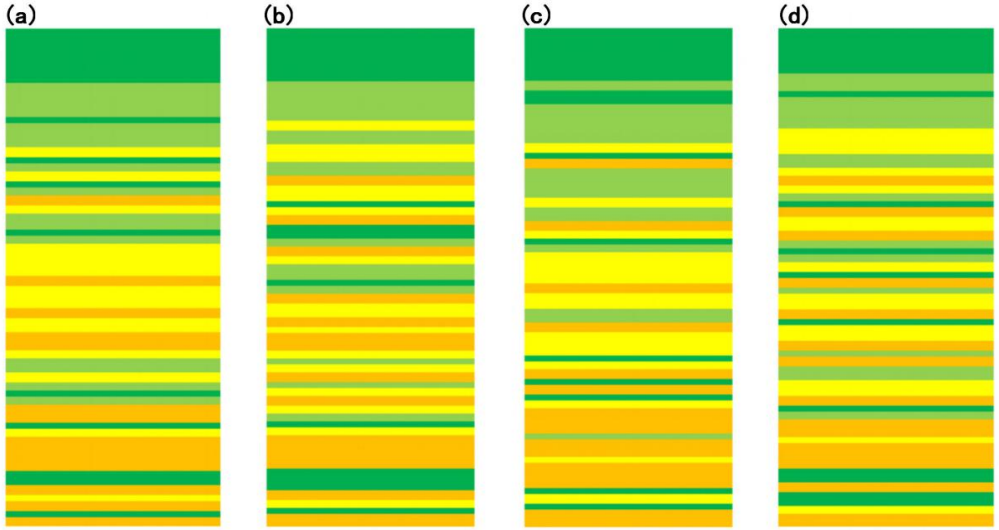

图 6. 采用不同评级方法所导致的期刊排名差异

3.4 Task 2 —— 结果讨论

ISCIENCE 的评级结果就是一个典型的例子，它的 IF 名列 15，中心度因子名列 13-25 之间，而 JournalRank 值很低，排名 46。

表 5. ISCIENCE 期刊各因子数据

名称	排名	Journalrank	$C_D(N_i)$	$C_B(N_i)$	$C_C(N_i)$	$C_E(N_i)$	IF	Pap_i	Ref_i	$Cite_i$
ISCIENCE	46	0.002672	0.41269	0.00143	0.47675	0.10444	4.447	495	28577	1410

结合 ISCIENCE 和其它期刊间的引用关系我们不难得出原因：ISCIENCE 的引用居多，被引很少。正是**引用和被引用关系的不对等性**，导致了 ISCIENCE 期刊评级的降低。这从另一方面也说明了我们算法的优越性，即可以**有效避免小期刊 “抱大腿” 的现象。**

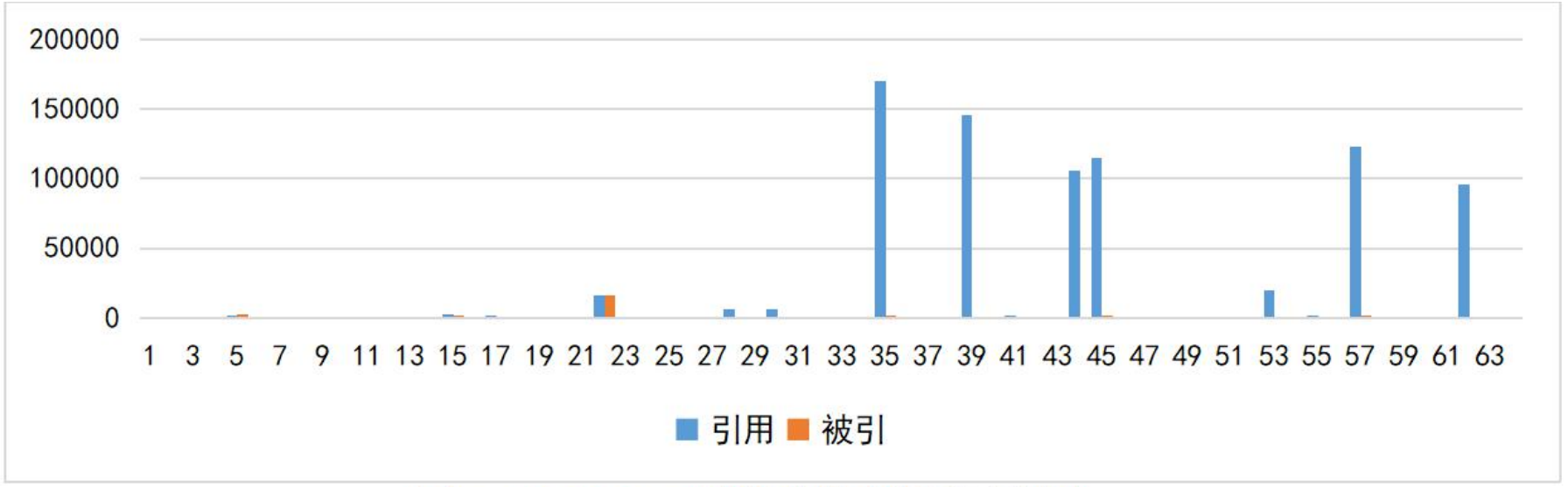

图 7. ISCIENCE 期刊引用关系示意图

3.4 Task 2 —— 结果讨论

另一方面，我们则从相关性的角度出发，考察不同评价因子之间、评价因子和参数 Pap_i，Ref_i，$Cite_i$ 之间的关系。具体计算方法如下：

$$Pear_{ij} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{18}$$

其结果如图 8 所示，值得注意的是 JIF 和大部分的影响力因子几乎不相关，甚至和总引用数弱相关，这充分**说明了 JIF 模型在反映期刊在特定领域影响力方面的不足**。另一方面，可以看到，**和总引用数唯一的强相关变量是 JournalRank**，这也从侧面说明了我们算法的合理性。

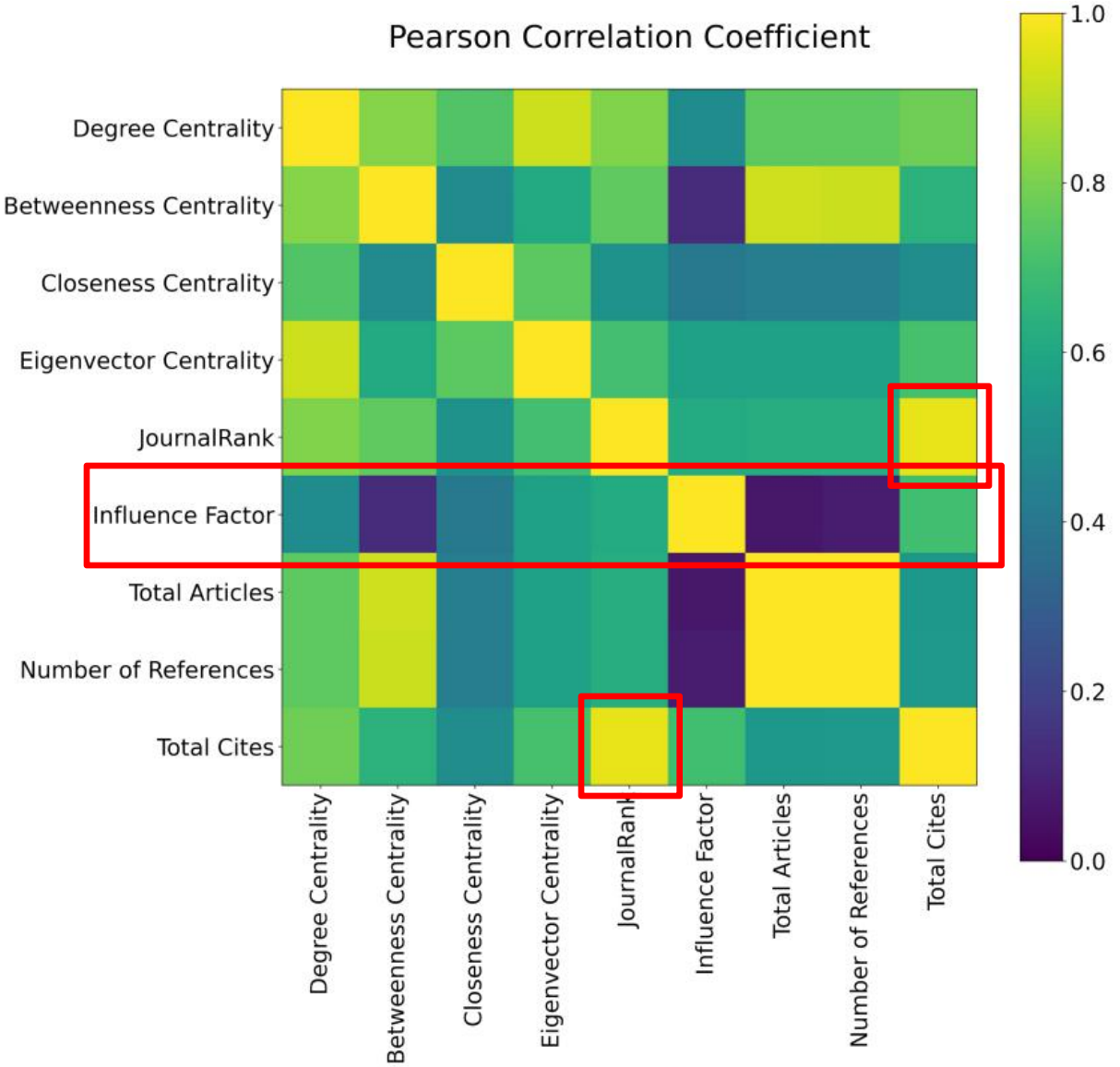

图 8. 不同模型变量之间的 Pearson 相关系数

4.1 Task 3: 关于模型参数的若干讨论

为了探究影响因素间的独立性，我们对 Pap_i、Ref_j、$Cite_i$ 间的关系进行了进一步的探究。

首先，我们发现 **Ref_j 与 Pap_i 的线性相关性极强**，一个解释是，大多数研究人员在写作文章时，引用的文献数量大致一致；

然而，$Cite_i$ 与 Pap_i 的相关性不高，这说明学术期刊上发表的**文章数与学术期刊的总被引数没有必然联系**，进而我们对 Garfield 模型中 "$H_{i>j}$ 与 Pap_i 保持线性关系" 的假设提出一定质疑。

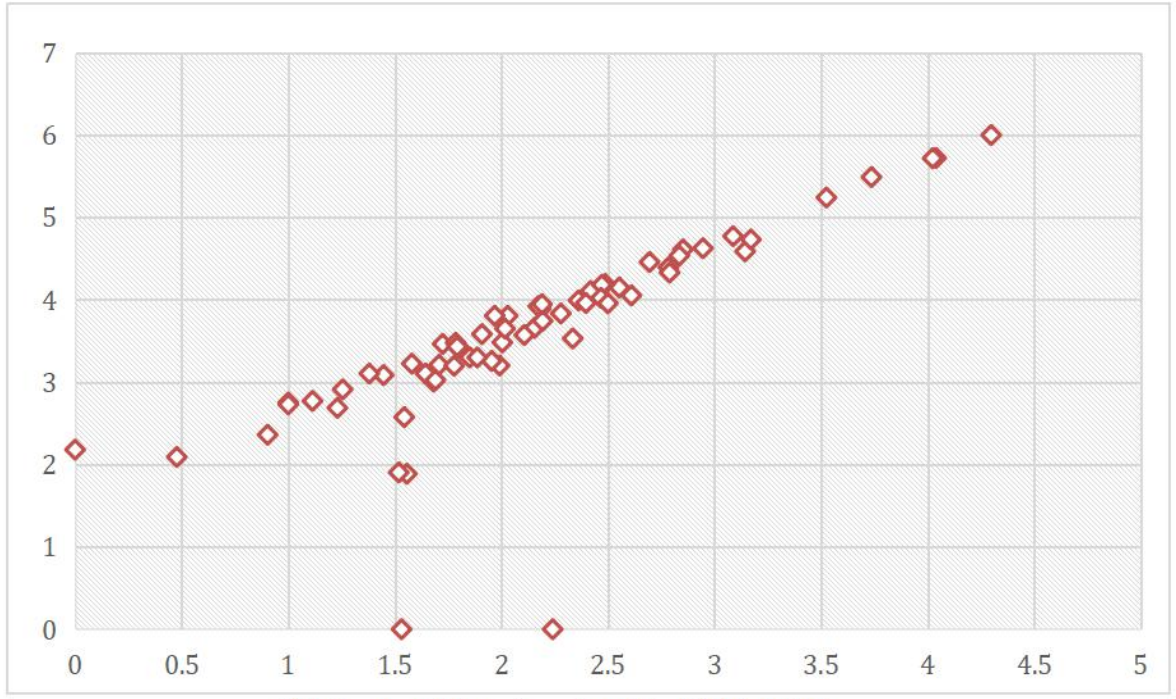

图 9： $\log(Pap_i) - \log(Ref_i)$**散点图**

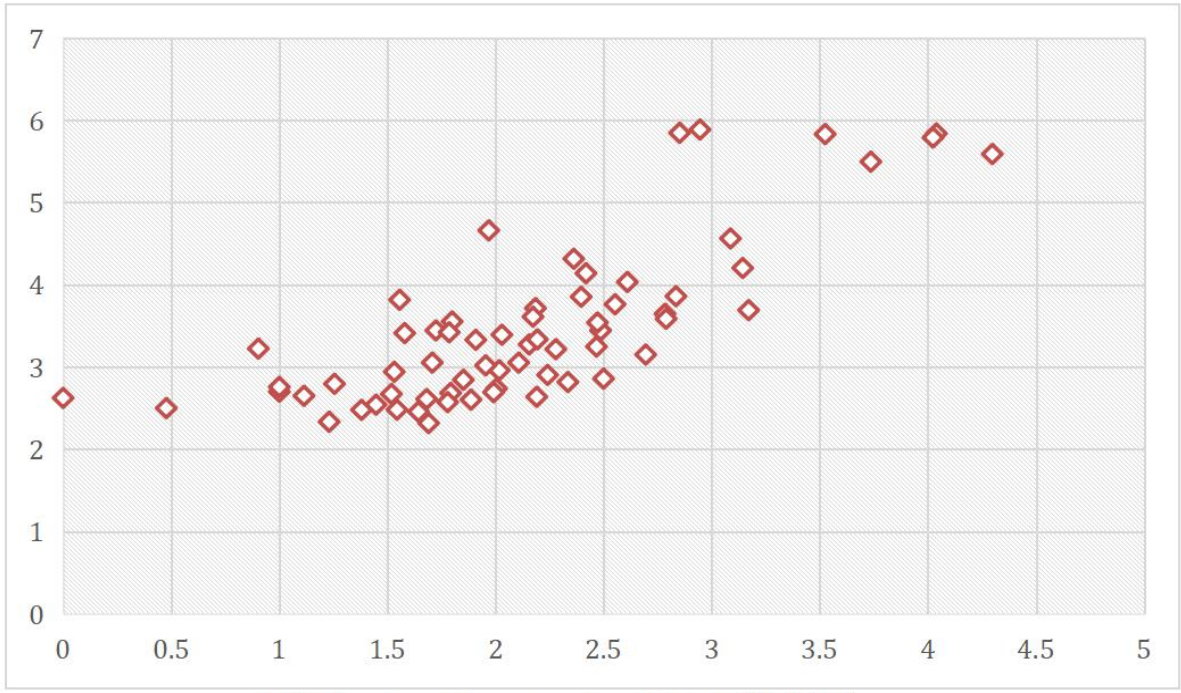

图 10： $\log(Pap_i) - \log(Cite_i)$**散点图**

4.2 Task 3 —— 引入期刊规模参数的合理性分析

沿用了 Garfield 的假设，我们采用了以下方式来定义期刊规模参数：

$$\textbf{期刊规模参数} = \mathbf{Ref_i} * \mathbf{Pap_j}$$

图11(a) 中展示了考虑期刊规模参数后，期刊规模和引用网络边权重之间的相关度，图11(b) 则展示的是不考虑期刊规模参数的模型。

可以看到，在考虑了期刊规模参数后，**边权重和总引用数之间的相关性明显下降，但和总文章数的相关性得到一定的保持。**

综合考虑之下，考虑期刊规模参数的是更好的模型。

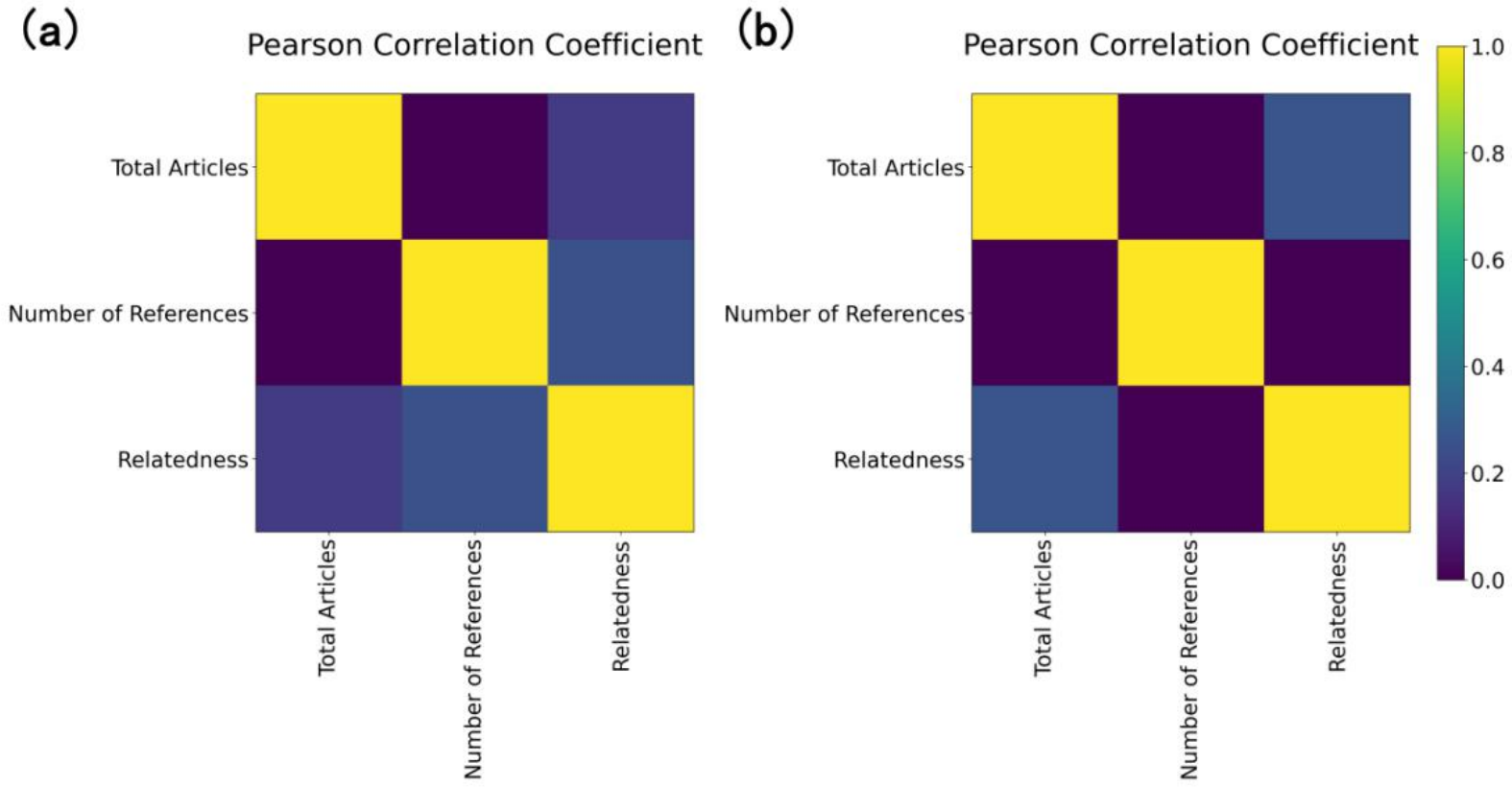

图 11. 期刊规模参数与引用网络边权重的影响

4.3 Task 3 —— 规模参数对模型结果的影响分析

为了研究期刊规模参数对 JournalRank 的具体影响，我们使用下面的式子作为新的网络权重：

$$x_{ij} = Conn_{g\times g}(i,j) = R_{i>j} = Norm(H_{i>j}) \qquad (19)$$

它和原模型的结果对照如图 12 所示，某些期刊的排名会发生剧烈的变化，下面我们结合 NPJ-MICROGRAVITY 期刊的例子来进行原因说明。

从表6可以看到在考虑了规模等因素后，NPJ 的 JournalRank 排名显著上升。这是因为相比于平均发文数 835.390625，NPJ 的发文数量很少，只有 28 篇，**如果不考虑期刊本身的规模**，则 JournalRank 算法会**更倾向于给发文数量多的大刊更高的评分**，而忽视了其期刊中文章的平均质量。

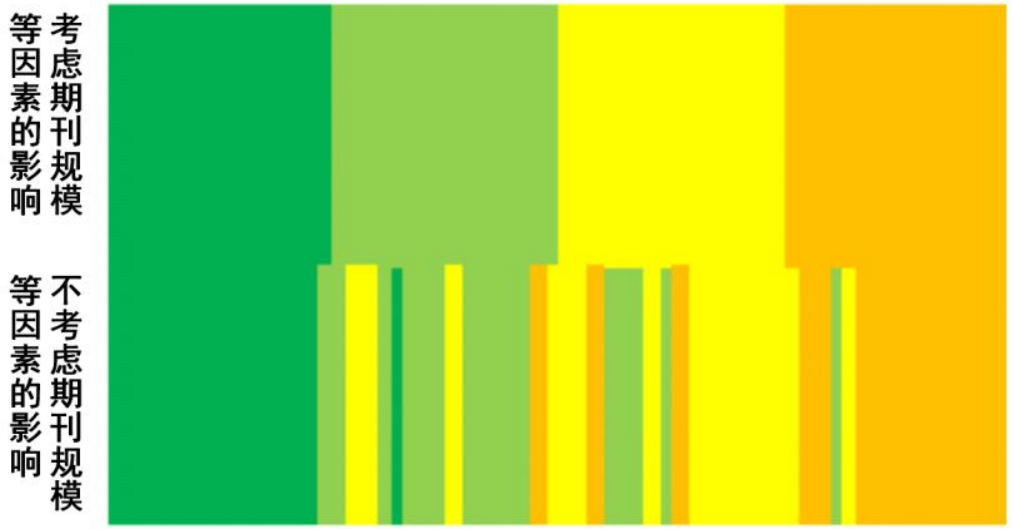

图 12. 期刊规模参数对模型结果的影响

表 6. NPJ-MICROGRAVITY 期刊各因子数据

名称	排名 1	排名 2	JournalRank	$C_D(N_i)$	$C_B(N_i)$	$C_C(N_i)$	$C_E(N_i)$	IF	Pap_i	Ref_i	$Cite_i$
NPJ-MICROGRAVITY	41	54	0. 190476	0. 00024496	0. 46457607	0. 04637865	0. 00329344	3.380	28	1214	347

欢迎大家批评指正！