

● 王凌燕^{1,2}, 方 曙²

(1. 中国科学院 研究生院, 北京 100049; 2. 中国科学院 成都文献情报中心, 四川 成都 610041)

Q 测度法对探测新兴研究趋势作用的探讨

摘要: 新兴研究趋势的探测已经成为国内外科学计量学研究的一个热点问题。本文引入社会网络中的弱连接概念并分析弱连接的特点, 利用实际数据, 通过 Q 测度方法对两个子集中不同节点在不同性质的网络间的桥接度 Q_w 值进行计算, 得出 Q_w 值较低的节点是覆盖两个子集知识融合的部分, 这里通常有助于新知识的产生, 据此提出 Q_w 值将可作为探测新兴研究趋势的参考判据指标之一。

关键词: 新兴研究趋势; 网络计量学; 弱连接

Abstract: Detecting emerging trends has become a hot spot of scientometrics research at home and abroad. This paper introduces the concept of weak tie in social network and analyzes its characteristics. The paper then uses the practical data and the Q-measure to calculate the bridging degree Q_w value of the different nodes in 2 subsets among networks with different characters, and concludes that the nodes with lower Q_w -values are those that cover the knowledge fusion part of the 2 subsets, which usually helps to produce new knowledge. Thereby, it's proposed that the Q_w value may be used as one of the reference indices for detecting emerging trends.

Keywords: emerging trend; Webometrics; weak tie

新兴研究趋势 (Emerging Trend) 是指对新研究论题进行深入研究并日趋重要的趋势, 这种趋势通常位于科学技术前沿, 通过研究人员研究成果之间的相互引用表现出来。

目前, 探测新兴研究趋势的主要方法之一是进行引文网络的分析。通过构建引文网络, 即把著者、文献、机构等作为节点, 把共引、耦合等关系作为节点之间的连接。分析引文网络的结构和特征, 研究网络中的各个节点以及节点与节点之间的关系形式与关系强度。在由引用关系所确定的不同研究维度的引文网络中, 节点间的连接关系对探测新兴研究趋势到底有什么样的作用和影响? 本文运用 Q 测度方法测度网络节点之间的桥接度, 分析桥接度大小的意义, 来尝试作些探讨。

1 新兴研究趋势的概念

A. Kontostathis 等在 2003 年提出新兴研究趋势的定义是: 随着时间逐渐引起人们兴趣, 并被越来越多的学者所研究或讨论的主题领域^[1]。其实质是指在某个科学技术研究领域备受研究者关注的一组新主题领域集合。

2 新兴研究趋势探测的主要方法

新出现研究趋势分析的方法主要从两个方面入手^[2]。一是把一系列高引用的文章组成聚类串, 检验其在随时间

推移变化中的特征, 从而检测出新出现的研究趋势。二是引文网络的特征分析。引文网络由著者、文献、机构等作为节点, 把共引、耦合等关系作为节点之间的连接^[3]。引文网络主要是从网络的内部关系和结构入手, 研究引文网络中的各个节点以及节点之间的关系形式和关系强度。

从边缘位置的角度对新出现知识进行的研究主要集中在弱连接 (Weak Tie) 上。早在 1965 年 Harary 在其图论文章中提到, 网络中的弱连接可能具有桥接性, 即连接不同团体的桥接功能, 对于信息传播来说, 连接不同团体, 可以传递异质性信息^[4]。但文中没有对弱连接的这种桥接性进行深入的讨论, 到 1973 年 Granovetter 对弱连接的性质和强度进行了深入讨论, 分析的起点来自于对三方组关系 (Forbidden Triad) 的探讨^[5]。因为存在被禁止的三方组关系, 所以强连带很难作为桥, 只有弱连带才可能发挥连接不同团体的桥接功能。对于桥接性的测度可以运用 Q-measure 测度方法对不同节点在不同性质的网络间的桥接度 Q_w 值进行计算, 不同的 Q_w 值大小具有什么不同的意义? 它对新兴趋势探测又有什么帮助和作用? 本文将通过实例分析来作些探讨。

3 探讨 Q-measures 对探测新研究趋势的作用

3.1 Q-measures 方法及其研究现状

R. Rousseau 在 2005 年 10 月 28 日至 11 月 2 日举行的

《美国信息科学与技术会刊》(American Society for Information Science & Technology) 年会上介绍了一个由社会网络学家提出的二分网络的 Q 测度方法。这是社会网络学家 A. L. Flom 在文章“A New Measure of Linkage between Two Sub-networks”中提出的一个新的、可应用于科学计量学的网络方法——Q 测度方法^[6]。它既可以应用于网络中的个别顶点,也可用作网络的整体评价。当连通的、无向网络中存在两个子集团时,这个方法可用来获取两个集团之间的桥梁关系。而后 R. Rousseau 对 Q-measures 进行改进,把它推广到一个有向、带权值的二分网络^[7]。随后 R. Guns 和 R. Rousseau 把 Q-measures 进一步进行推广,不再仅限于一个二分的网络,而是可以对“有限个”有向、带权值网络进行计算^[8]。

国内有学者应用 Q 测度方法对合作网络中的性别进行研究^[9],将 Q 测度的思想转化为计算机语言,当网络规模增大到一定程度以后,采用此算法计算出多个顶点的 Q 值,并分析了一个中型的合作网络,得出 Q 值方法是一个有效的、能反映网络中具有不同性质的两个集团之间交流程度的方法。而运用 Q 测度方法探测新兴趋势这方面的工作尚不多见。

笔者尝试性地选取在连通有向、带权值的网络中进行 Q-measures 计算,所得 Q_w 值用于表示节点在网络间的桥接度,通过对 Q 值指标的数据特征进行比较,探讨它对新兴研究趋势判断的作用。

3.2 在有向权值网络中的 Q-measures 方法

1) 定义最大流量问题。最大流量问题 (Maximum Flow Problem) 是经济学和管理学中的一类典型问题。在一个网络中每段路径都有“容量”和“费用”两个限制的条件下,此类问题的研究试图寻找出:流量从 A 到 B,如何选择路径、分配经过路径的流量,可以达到所用的费用最小的要求。在有向权值网络中考虑到有向网络中每一条边有一个最大传输能力,用以限制在每条弧上传输物质的数量。如果想要找到通过网络的起点到终点的最大流量,这样的问题被称为最大流量问题。我们可以看到一个网络流包含很多弧,并且对于最大流有可能访问一个节点很多次^[10]。

2) 定义有向权值网络中节点 j 到 k 之间的网络容量。定义连接节点 j 到 k 之间的网络最大容量为 j 到 k 之间的网络最大流^[11],记为: $\text{Cap}(j, k)$,值得注意的是 $\text{Cap}(j, k)$ 不等于 $\text{Cap}(k, j)$ 。如果 $j=k$ 则 $\text{Cap}(j, k) = 0$ 。如果节点 j, k 之间无连接则 $\text{Cap}(j, k) = 0$ 。

3) 对于有向权值网络流动介数向心性计算。假设一个网络有 N 个节点,用 $\text{Cap}(j, k)$ 记为节点 j 到 k 之间的网络最大容量,用 $\text{Cap}(j-i-k)$ 来标记当位于节点 j 到 k

之间,当点 i 和其所有的连接被删除后的网络容量。则对节点 i 的介数向心性计算如下:

$$B_w(i) = \frac{\sum_{(j,k) \in S(i)} (\text{Cap}(j,k) - \text{Cap}(j-i-k))}{\sum_{(j,k) \in S(i)} (\text{Cap}(j,k))} \quad (1)$$

当 $S(i) = \{ (j,k) : 1 \leq j \leq N; 1 \leq k \leq N; j \neq i; k \neq i; k \neq j \}$

值得注意的是:

如果 $\text{Cap}(j, k) = 0$, 则对于所有的 i , $\text{Cap}(j-i-k) = 0$ 。

如果 $\sum_{(j,k) \in S(i)} \text{Cap}(j,k) = 0$, 则 $B_w(i) = 0$ 。

4) Q_w : 有向权值网络的 Q-measure 计算。对于有向权值网络的 Q-measure 计算以公式 (1) 为基础。假设一个有向权值网络中有两个子集分别为 G, H , 如果作者 a 属于组集 G , 定义作者 a 为 g_i , 则对于作者 a 的 Q-measure 计算定义为: g_i, h_j 。

$$Q_w(a) = \left(\sum_{i=1}^m \sum_{j=1}^n (\text{Cap}(g_i, h_j) - \text{Cap}(g_i - a - h_j)) + \sum_{i=1}^m \sum_{j=1}^n (\text{Cap}(h_j, g_i) - \text{Cap}(h_j - a - g_i)) \right) / \left(\sum_{i=1}^m \sum_{j=1}^n \text{Cap}(g_i, h_j) + \sum_{i=1}^m \sum_{j=1}^n \text{Cap}(h_j, g_i) \right) \quad (2)$$

相同的定义对于作者 a 属于另一个组集时 ($a = h_k$):

$$Q_w(a) = \left(\sum_{i=1}^m \sum_{j=1}^n (\text{Cap}(g_i, h_j) - \text{Cap}(g_i - a - h_j)) + \sum_{i=1}^m \sum_{j=1}^n (\text{Cap}(h_j, g_i) - \text{Cap}(h_j - a - g_i)) \right) / \left(\sum_{i=1}^m \sum_{j=1}^n \text{Cap}(g_i, h_j) + \sum_{i=1}^m \sum_{j=1}^n \text{Cap}(h_j, g_i) \right) \quad (3)$$

注意,得到的 Q_w 的值总是介于 $[0, 1]$ 之间。

3.3 实证分析

本文选取“Journal Citation Reports” (Sciences edition 2008) 中生物领域的 11 种期刊作为实验分析对象 (见表 1), 这 11 种期刊被划分成两个类别: Biochemistry & Molecular Biology (生物化学与分子生物学) 以及 Genetics &

Heredity (基因与遗传学)。所选的 11 种期刊是这两个领域中具有代表性的刊物, 影响因子有所不同且创刊时间有所不同, 根据时间序列由远及近排列。表 2 展示它们之间的引证关系。其中有部分期刊同时属于这两个类别 (根据 JCR 的学科类别划分), 就其学科覆盖范围和程度暂且把它们放在 Biochemistry & Molecular Biology (生物化学与分子生物学) 类别里面。

表 1 期刊名称及说明

类别	期刊名称及缩写	说明
Biochemistry & Molecular Biology	1. Biochemistry (BUS)	
	2. Biochemical and Biophysical Research Communications (BBRC)	Biophysics
	3. Gene Therapy (GT)	Genetics & Heredity ; Biology Biotechnology & Applied Microbiology ; Medicine, Research & Experimental
	4. Genome Research (GR)	Biotechnology & Applied Microbiology ; Genetics & Heredity
	5. HUMAN Molecular Genetics (HMG)	
	6. JOURNAL of Biological Chemistry (JBC)	
Genetics & Heredity	7. American Journal of Human Genetics (AJHG)	
	8. JOURNAL of Medical Genetics (JMG)	
	9. Human Genetics (HG)	
	10. Genetics (GE)	
	11. Human Mutation (HM)	

为了方便表示, 用 A, B, C, D, E, F, G, H, I, J, K 分别依次对应第 1~11 种期刊。

{A, B, C, D, E, F} 属于组 B (Biochemistry & Molecular Biology)。

{G, H, I, J, K} 属于组 G (Genetics & Heredity)。

其中最大流量 (最大传输容量) 的值通过 Pajek 软件计算得到, 见表 3。

计算关于 A 点的 $Q_w(A)$ 值:

要计算 A 点的 $Q_w(A)$ 值, 除表 3 中的数据外, 还需

表 2 引用关系

	BUS	BBRC	GT	GR	HMG	JBC	AJHG	JMG	HG	GE	HM
BUS	-	15	0	0	0	101	0	0	0	0	0
BBRC	3	-	0	0	9	58	2	0	0	1	0
GT	0	0	-	0	1	1	1	0	0	3	0
GR	0	1	0	-	3	2	7	0	1	2	2
HMG	3	1	0	0	-	15	19	7	3	0	4
JBC	68	27	0	2	14	-	1	1	0	1	3
AJHG	0	1	0	3	9	3	-	3	1	1	6
JMG	0	0	0	1	6	1	1	-	0	0	4
HG	0	1	0	1	11	1	10	6	-	2	6
GE	1	1	0	10	2	4	5	0	0	-	0
HM	0	0	0	0	2	2	5	2	2	0	-

注: 对于 11 种期刊的引用矩阵, 位于第一列的期刊分别引用其他期刊, 也就是每一列为“被引”的数据, 而第一行为“引用”的数据。

表 3 最大传输容量

	A	B	C	D	E	F	G	H	I	J	K
A	-	15	0	12	34	116	34	19	7	7	24
B	73	-	0	12	34	73	34	19	7	7	24
C	5	5	-	5	5	5	5	5	5	5	5
D	18	18	0	-	18	18	18	18	7	7	18
E	37	37	0	12	-	37	38	19	7	7	24
F	75	48	0	12	34	-	34	19	7	7	24
G	28	28	0	12	28	28	-	19	7	7	24
H	13	13	0	12	13	13	13	-	7	7	13
I	37	37	0	12	38	37	38	19	-	7	25
J	23	23	0	17	23	23	23	19	7	-	23
K	13	13	0	12	13	13	13	13	9	7	-

注: 表中数据给出了任意两节点之间的最大传输容量 (Capacities)。

要计算任意两节点除去 A 点后的最大传输容量, 结果见表 4。

把表 3 和表 4 中的数据带入公式 (2) 中可得到:

$$Q_w(A) = \left(\sum_{\substack{i=1 \\ i \neq k}}^m \sum_{j=1}^n (\text{Cap}(g_i, h_j) - \text{Cap}(g_i - a - h_i)) \right)$$

$$\begin{aligned}
 & + \sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq k}}^n (\text{Cap}(h_j, g_i) - \text{Cap}(h_j - a - g_i)) \\
 & / \sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq k}}^n \text{Cap}(g_i, h_j) \\
 & + \sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq k}}^n \text{Cap}(h_j, g_i) \Big) \\
 & = \left(\frac{20 + 2}{408 + 370} \right) = \frac{11}{289} \approx 0.028
 \end{aligned}$$

表4 除去A点后的任意两节点的最大传输容量

Cap(G - A - B) = 25	Cap(B - A - G) = 34
Cap(G - A - C) = 0	Cap(C - A - G) = 5
Cap(G - A - D) = 12	Cap(D - A - G) = 18
Cap(G - A - E) = 28	Cap(E - A - G) = 36
Cap(G - A - F) = 28	Cap(F - A - G) = 34
Cap(H - A - B) = 10	Cap(B - A - H) = 19
Cap(H - A - C) = 0	Cap(C - A - H) = 5
Cap(H - A - D) = 12	Cap(D - A - H) = 18
Cap(H - A - E) = 13	Cap(E - A - H) = 19
Cap(H - A - F) = 13	Cap(F - A - H) = 19
Cap(I - A - B) = 35	Cap(B - A - I) = 7
Cap(I - A - C) = 0	Cap(C - A - I) = 5
Cap(I - A - D) = 12	Cap(D - A - I) = 7
Cap(I - A - E) = 38	Cap(E - A - I) = 7
Cap(I - A - F) = 33	Cap(F - A - I) = 7
Cap(J - A - B) = 20	Cap(B - A - J) = 7
Cap(J - A - C) = 0	Cap(C - A - J) = 5
Cap(J - A - D) = 17	Cap(D - A - J) = 7
Cap(J - A - E) = 22	Cap(E - A - J) = 7
Cap(J - A - F) = 22	Cap(F - A - J) = 7
Cap(K - A - B) = 12	Cap(B - A - K) = 24
Cap(K - A - C) = 0	Cap(C - A - K) = 5
Cap(K - A - D) = 12	Cap(D - A - K) = 18
Cap(K - A - E) = 13	Cap(E - A - K) = 24
Cap(K - A - F) = 13	Cap(F - A - K) = 24

如果节点属于另一个子集即带入公式(3)中进行

计算。

利用上述方法可以得到其他各节点的 Q_w - value 值, 见表5。

表5 各期刊的 Q_w 值

期刊列表	Q_w 值
1. Biochemistry (BUS)	0.028
2. Biochemical and Biophysical Research Communications (BBRC)	0.160
3. Gene Therapy (GT)	0
4. Genome Research (GR)	0.142
5. Human Molecular Genetics (HMG)	0.275
6. Journal of Biological Chemistry (JBC)	0.426
7. American Journal of Human Genetics (AJHG)	0.352
8. Journal of Medical Genetics (JMG)	0.168
9. Human Genetics (HG)	0.171
10. Genetics (GE)	0.207
11. Human Mutation (HM)	0.165

利用上述数据, 运用 Pajek 软件绘制出带权值有向网络图, 见图1。

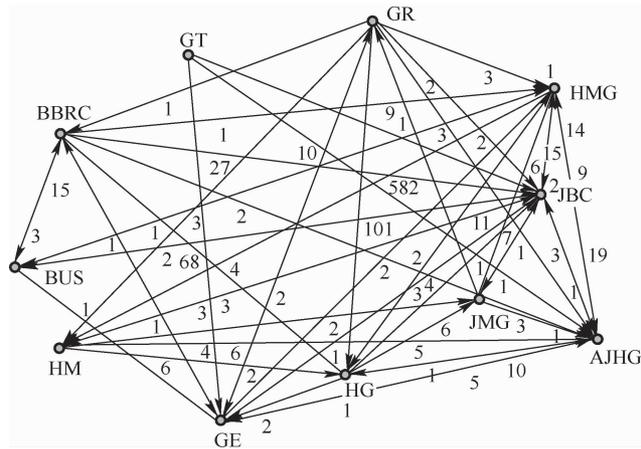


图1 通过 Pajek 绘制的带权值有向网络图

3.4 讨论

通过计算可以看到 Q_w 值最高的是属于 Biochemistry & Molecular Biology 领域的期刊“Journal of Biological Chemistry (JBC)”和属于 Genetics & Heredity 领域的期刊“American Journal of Human Genetics (AJHG)”, 表明 JBC 和 AJHG 节点是在这个网络中桥接度最高的节点。

此外, 同时属于两个集合领域的期刊(被划分到生物化学和分子生物领域进行计算)没有一个期刊具有较高

的桥接度。甚至“Gene Therapy”的桥接度为0，因为它只引用过其他期刊而未被其他期刊所引用，然而这种情况并不很常见，为了保证数据的真实性我们保留了这组数据；而“Genome Research”的网络桥接度也只有0.142，它们在整个网络中具有很低桥接度。

同时可以看到，“Gene Therapy”和“Genome Research”两种期刊都是跨领域的刊物，并且“Genome Research”不但属于生物化学与分子生物、生物遗传学，还属于应用（微）生物技术类别。经过对这两种期刊近两年发表的文章内容进行统计分析，发现这两种期刊对生物科学的前沿交叉领域论文刊载表现突出。

“Genome Research”在基因发现、基因组分析、生物信息学、基因结构和功能、技术革新和应用、最新的基因及物理制图、DNA测序领域方面进行最新报道，把最新的数据资源发表为研究论文、简要报告和摘要等学术性文章，是发布生物领域新兴研究趋势重要的研究平台。

对于期刊“Gene Therapy”，刊载最新的医学前沿以基因和细胞为基础的诊疗技术，并出版该领域最新进展以及展望评论的文章，是发布医学前沿以及基因诊疗技术领域新兴研究趋势成果的重要载体。

可以看到这些跨领域的期刊是融合学科间知识的平台，会促进新知识的产生，但新出现的知识是零星的，只能对局部发生影响，对全局的影响很小。就像期刊“Gene Therapy”基本只有引用的记录而没有被引用的记录，这也是它桥接度低的原因，而它却覆盖了生物化学与遗传生物学两个领域的知识，它又是学科交叉融合的结合点，催化新的知识和研究趋势产生的地方。通过社会网络分析新兴趋势时应当关注桥接度较低的弱节点，它们往往可能是新知识产生的关键点。因此， Q_w 值的大小揭示了桥接度的强弱，桥接度的强弱又反映出节点的强弱连接度。

网络中的小 Q_w 值节点有较大可能产生新的研究主题，我们在进行新兴研究趋势探测时可以把 Q_w 值作为参考判据指标之一。

4 结束语

Q-measure法所测的桥接度是网络节点连接的一种描述，选取期刊作为节点研究是对Q测度运用于新兴研究趋势探测可行性理论的试验性验证研究，但该指标是否真正可用于作为探测新兴研究趋势的重要判据之一，还需要作更多的实证分析来确定，例如用不同学科或技术领域具体研究主题的文献引用数据进行分析验证。进行新兴研究趋势的主动挖掘和判断是一个复杂的处理过程，对于Q-measure方法计算出的 Q_w 值可以作为一个参考的参数指

标，并不是说桥接度低的连接点就一定会产生新知识，更多的是要运用综合分析的方法，运用文本挖掘技术、文献计量学、网络计量学、时间序列分析等多种方法，多角度、多层次，总结更多必要的判断指标，结合学科领域专家的知识才能得到较准确地对新兴研究趋势探测和识别的结果。□

参考文献

- [1] KONTOSTATHIS A, et al. A survey of emerging trend detection in textual data mining [J]. A Comprehensive Survey of Text Mining, 2003 (9).
- [2] 刘玉仙, ROUSSEAU R. 新出现趋势识别和分析方法引介 [J]. 科学学研究, 2009, 27 (7): 995-996.
- [3] WHITE H D, McCAIN K W. Visualizing a discipline: an author co-citation analysis of information science, 1972 - 1995 [J]. Journal of the American Society for Information Science, 1998, 49 (4): 327-355.
- [4] HARARY F. Graph theory and group structure [J]. Mathematical Psychology, 1965 (2).
- [5] GRANOVETTER M. The strength of weak ties [J]. American Journal of Sociology, 1973, 78 (6): 1360-1380.
- [6] FLOM P L, FRIEDMAN S R, STRAUSS S, et al. A new measure of linkage between two sub-networks [J]. Connections, 2004, 26 (1): 62-70.
- [7] ROUSSEAU R, ZHANG L. Betweenness centrality and Q-measures in directed valued networks [J]. Scientometrics, 2008, 75 (3): 575-590.
- [8] GUNS R, ROUSSEAU R. Gauging the bridging function of nodes in a network: Q-measures for networks with a finite number of subgroups [EB/OL]. http://users.telenet.be/Ronald.rousseau/Q_more_groups_ISSL_final.pdf.
- [9] 张文雷, 尹丽春, 等. Q-测度法在合作网络性别研究中的应用 [J]. 科技进步与对策, 2009, 26 (15): 100-103.
- [10] BONAIEH P, HOLDREN A C, JOHNSTON M. Hyper-edges and multidimensional centrality [J]. Social Networks, 2004, 26 (2): 189-203.
- [11] ROUSSEAU R. Q-measures for binary divided networks: an investigation within the field of informetrics [C] //Proceedings of the 68th ASIST Conference, 2008, 42: 675-696.

作者简介：王凌燕，女，1984年生，硕士生。

方曙，男，1957年生，博士生导师。

收稿日期：2010-08-16