

科研主题演化中三种典型社区发现算法对比研究 ——以植物甾醇信号为例^①

薛 陕^② 董 诚^③ 韩红旗 张均胜 高 雄 王 力

(中国科学技术信息研究所 北京 100038)

摘 要 科研主题识别和演化分析是了解科研领域发展现状和趋势的重要手段。近年来,基于复杂网络和网络演化计量的社区发现算法成为揭示科研主题演化的主要方法,有必要深入探讨该类算法在主题识别和演化分析中的优点与不足。本文选取植物油菜素甾醇领域,构建了关键词共现网络,对比了 Newman MM、Ball Overlapping 和 Blondel 这 3 种典型社区发现算法的科研主题识别和演化追踪结果,并结合专家知识对 3 种算法效果进行分析。实验结果表明,Blondel 算法在科研主题的识别和演化追踪研究中运算速度最快,识别的主题更为准确,能够较好反映领域主题演化。该研究从算法适用性角度为科研主题演化追踪研究提供了参考。

关键词 科研主题演化;社区发现算法;植物油菜素甾醇;关键词共现网络

0 引 言

科研主题演化是衡量科研主题随时间推移表现出的动态性、发展性和差异性的研究。科研主题演化一般包含两方面含义:其一是科研主题内容随着时间推移而发生变化,其二是不同科研主题之间复杂的承继关系。其中,主题识别是演化分析的基础和关键因素。目前主题识别的主流方法可分为基于语言模型的方法^[1-2]和基于网络社区发现的方法^[3-4]。基于网络社区发现的主体识别方法由于速度快、社区划分比较准确,目前已经成为科研主题识别的主要方法^[5-7]。

社区(community)是社会网络中的常见现象,由一群高度聚集、联系紧密的节点聚集组成,在各种知识网络中普遍存在社区结构^[8-10]。社区结构研究可以追溯到 1977 年 Zachary^[11]对空手道俱乐部成员关系网络的研究。Girvan 和 Newman^[12]在对社会

网络的研究中提出了著名的(Girvan-Newman)GN 社区发现算法,随后的研究发现在物理学家合作网络中同样存在社区现象^[13]。Boyack 等人^[14]在利用 7121 种期刊数据集绘制科学景观鸟瞰图的研究中也发现了类似的社区结构。Lambiotte 等人^[15]发现存在于知识网络层面的社区,是一种划分知识领域和学科前沿的新视角。

由于构成网络的数据集的不同,网络中的社区结构往往存在不同特点,不同社区算法对特定网络社区识别效果存在差异^[16]。社区发现算法对主题识别效果直接影响了主题演化结果和路径的判断。因此,有必要对当前主要社区发现算法的效果进行对比研究,了解其效果以及适用性能。本研究以植物甾醇信号相关文献关键词共现网络为实例,选取 3 种典型社区发现算法对其主题发现和演化追踪效果进行了对比,并结合专家知识对植物甾醇信号主题研究现状和演化趋势进行了分析解释。

① 中国工程科技知识中心建设项目(CKCEST-2021-2-6)和中央级公益性科研院所基本科研业务费专项基金青年项目(QN2020-10)资助。

② 男,1984 年生,博士生;研究方向:知识组织与知识工程;E-mail: xues@istic.ac.cn。

③ 通信作者,E-mail: dongc@istic.ac.cn。

(收稿日期:2020-12-24)

1 相关研究

网络社区结构的聚类方法与计算机科学中的图形分割 (graph partition) 和社会学中的分级聚类 (hierarchical clustering) [17-18] 有着密切联系。复杂网络社区识别方法按照聚类算法的不同可以分为以下几类:基于谱平均法的聚类算法、基于分裂的聚类算法、基于凝聚的聚类算法以及基于重叠社区的聚类算法。计算复杂度以及准确性是分析复杂网络社区结构面临的主要问题。如表 1 所示,谱平均法难以适用于社区结构复杂的网络结构,而以 GN 算

法[12] 为代表的分裂算法由于运算复杂不适用于大型网络,因此本文不再选用这两种算法进行研究。目前已有的研究发现基于凝聚的方法[19] 和基于重叠社区[9] 的方法在处理复杂网络社区划分中具有较好的效果[20]。因此本文选取基于凝聚的聚类算法中具有代表性的 Newman MM 算法[13, 19]、Blondel 算法[21] 以及基于重叠社区发现的 Ball Overlapping 算法[22] 作为研究对象。以植物油菜素甾醇研究领域的关键词共现网络为例,对这几种社区发现算法的社区划分速度、准确性以及在演化分析过程中的适用性进行了对比,揭示了它们在主题演化研究中的优点与不足。

表 1 常用社区发现算法的分类与优缺点对比

社区发现算法分类	代表算法	优点与不足
基于谱平均法的聚类算法	Wu-Huberman 算法[23], Capocci 算法[24] 等	速度快,但处理社区界限比较模糊的复杂网络时划分的效果并不好
基于分裂的聚类算法	GN 算法等	比较准确,但是由于引入反复计算的思想,计算复杂度较大
基于凝聚的聚类算法 基于重叠社区发现的聚类算法	Newman MM 算法, Blondel 算法等 派系过滤算法, Ball Overlapping 算法等	相比 GN 算法运算速度快很多,但不支持重叠社区 可以发现重叠社区,但需要事先知晓网络中社区数目 而且算法复杂度较高

2 实验方法设计

2.1 实验数据

以 2010 年为分界点,2010 年之前每年关于植物甾醇信号的研究不到 50 篇,2010 年之后每年文献的数量迅速增长,到 2017 年到达顶峰,有 141 篇文献。因此将文献检索时间限定为 2010 - 2017 年,以检索式“Brassinosteroids”[MeSH Terms] or “Bras-

sinosteroids”[AllFields] or “Brassinosteroid”[All Fields] 从 Pubmed 数据库检索,获得关于植物油菜素甾醇研究文献 962 篇。

2.2 关键词共现网络的构建

对收集到的 962 篇文献进行处理,按年份为尺度对文献进行分割。以关键词为节点,关键词的共现关系为边,构建了每年的关键词共现网络,各关键词网络指标见表 2。

表 2 2010 - 2017 年植物甾醇激素领域关键词共现网络部分指标

	2010	2011	2012	2013	2014	2015	2016	2017
节点数量	403	539	846	812	773	737	675	794
边的数量	3812	5522	8951	8509	7645	7619	6078	7235
平均度	18.91	20.48	21.16	20.95	19.78	20.67	18.00	18.22
弱网络链接组件数量	2	3	1	3	1	3	3	4

2.3 科研主题演化的计算和可视化展现

经过调研,选择文献[25] 提出的相似度计算公

式作为不同时间窗口的主题相似度的测度。该公式基于节点重合度计算两个社区的相似度,能够较好

地反映两个主题之间的相似性。给定社区 M_x 和社区 M_y , 各自对应的词汇集合为 C_x, C_y , 它们的相似度按式(1)定义为

$$Sim(M_x, M_y) = \frac{\sum_{v \in C_x \cap C_y} W(v)}{\min(\sum_{v \in C_x} W(v), \sum_{v \in C_y} W(v))} \quad (1)$$

其中, $W(v)$ 表示节点的频次, $\min(x, y)$ 为 x 和 y 中较小的值。如果前后两个连续时间段中的社区相似度超过设定的阈值, 则认为两个社区存在演化关系。社区 $M_{(T+1)_j}$ 的前驱定义见式(2)。

$$Pre(M_{(T+1)_j}) = (M_{t_i} \mid M_{t_i} \in G_t, Sim(M_{t_i}, M_{(t+1)_j}) > \delta) \quad (2)$$

其中, δ 是可调节的阈值, 根据经验取值为 0.3。为了判断社区的演化形式, 本研究参考了文献[9, 26]所提出的方法, 将网络社区的演化过程定义为 6 种形式, 分别是产生、消亡、分裂、融合、扩张和收缩。主题演化的可视化则采用可视化软件 NEViewer 以河流图形式展现。

2.4 基于专家知识的社区发现算法效果的比较

由于在复杂网络的社区分割中不存在有效的精确解法(该问题是一个 NP 难题)^[27-29], 因此很难定义一个量化指标并从准确性的角度评价不同算法的优劣。为了比较社区发现算法的聚类识别性能, 以确定效果最好的主题识别算法, 本研究邀请领域专家对收集的文献集进行了主题标引, 结合专家标引的结果对 Newman MM 算法和 Blondel 算法的社区

划分以及主题演化分析效果进行了对比。

3 结果及分析

3.1 三种典型社区发现算法社区划分效果的比较

目前基于复杂网络理论的主题聚类算法有很多, 经过调研, 选取其中使用较多、有代表性的 3 种算法: Newman MM 算法、Ball Overlapping 算法和 Blondel 算法, 对文献关键词所组成的共词网络进行了社区划分, 并对这 3 种算法的社区划分性能进行了比较。从运算速度上看, Blondel 算法最快, 仅用 21 s; Ball Overlapping 算法次之, 用时 139 s; Newman MM 算法最慢, 用时 977 s, 约是 Blondel 算法的 46 倍, Ball 算法的 7 倍。社区具体内容如表 3 所示, 在社区识别结果上, 因为 Ball Overlapping 算法是支持重叠社区识别的, 因此产生了很多重复的社区关键词。以识别到的 10 个社区为例, 其中有 8 个社区的关键词都是重复的, 表明该算法对该数据集的支持性不好。Blondel 算法和 Newman MM 算法在社区识别效果上比较接近, Blondel 算法发现了 12 个社区, Newman MM 算法发现了 10 个社区。这 2 种算法发现的社区的代表性关键词也具有较好的解释性, 如 Mutation(突变体)、Plants、Genetically modified(转基因植物)、Signal Transduction(信号传导)等都是植物甾醇激素领域研究的热点主题, 而且基本没有产生重复的关键词。因此之后的研究中可着重针对这 2 种算法作对比。

表 3 3 种社区发现算法的社区内容

Newman MM 算法			Blondel 算法		Ball Overlapping 算法	
社区编号	名称	词量	名称	词量	名称	词量
0	Cell Differentiation/genetics	4	Mutation	53	Brassinosteroids	57
1	Gene Expression Regulation, Plant	86	Cholestanols/ chemical synthesis	7	Brassinosteroids	76
2	Cell Nucleus/metabolism	45	Steroids, Heterocyclic/ metabolism	50	Brassinosteroids	60
3	Animals	29	Cholestanols/analysis	14	Gene expression regulation, Plant	48
4	Steroids, Heterocyclic/ pharmacology	71	Phenotype	26	Brassinosteroids	40
5	Cholestanols/analysis	14	Genes, Plant	41	Brassinosteroids	53

续表 3

6	Gene Expression Regulation, Plant/drug effects	23	Plants, Genetically Modified	43	Brassinosteroids	67
7	Plant Growth Regulators/ pharmacology	46	Arabidopsis/drug effects/genetics	7	Brassinosteroids	36
8	Brassinosteroids	46	Animals	29	Brassinosteroids	49
9	Mutation/genetics	43	Signal Transduction	61	Signal Transduction	39
10	—	—	Brassinosteroids	67	—	—
11	—	—	Cell Differentiation/genetics	4	—	—

3.2 科研主题演化的计算和可视化

由于 2010 - 2017 年间的文献总共有 926 篇,如果这些文献都由专家进行主题标注的话,专家的工作量会很大。因此本文只选取了 2010 - 2014 年的文献集为例。采用 2.3 节所述方法,对这 5 年的植物甾醇激素领域相关文献的主题进行了识别和演化追踪,并通过可视化软件 NEViewer 将主题的演化结果进行可视化展现。图 1 展示了采用 Blondel 算法绘制的主

题演化河流图,图 2 展示了采用 Newman MM 算法绘制的主

题演化河流图。图 2 展示了采用 Newman MM 算法绘制的主

3.3 基于专家知识的社区发现算法准确性的比较

为了评价这 2 种社区发现算法的演化分析结果的优劣,邀请了中科院植物所的专家,利用专家的知识对 Newman MM 算法和 Blondel 算法的社区划分效果进行对比。以 Animals 这个主题为例说明专家的判断结果。Animals 主题在 2010 年的数据集中,Newman MM 算法和 Blondel 算法都发现了此主题。不同的是,Blondel 算法在 2011 年侦测到主题 Animals 的后继主题状态为扩张,而 Newman MM 分区算法没有侦测到主题 Animals 的后继主题。将 Blondel 算法侦测到的主题 Animals 的演化网络从整

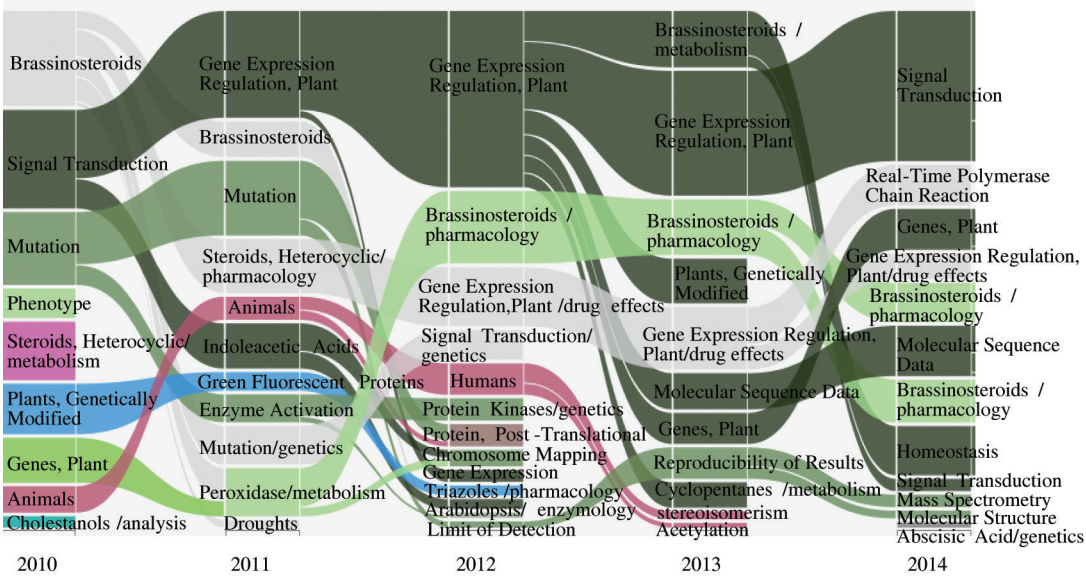


图 1 植物甾醇激素领域的主题演化河流图 (Blondel 算法)

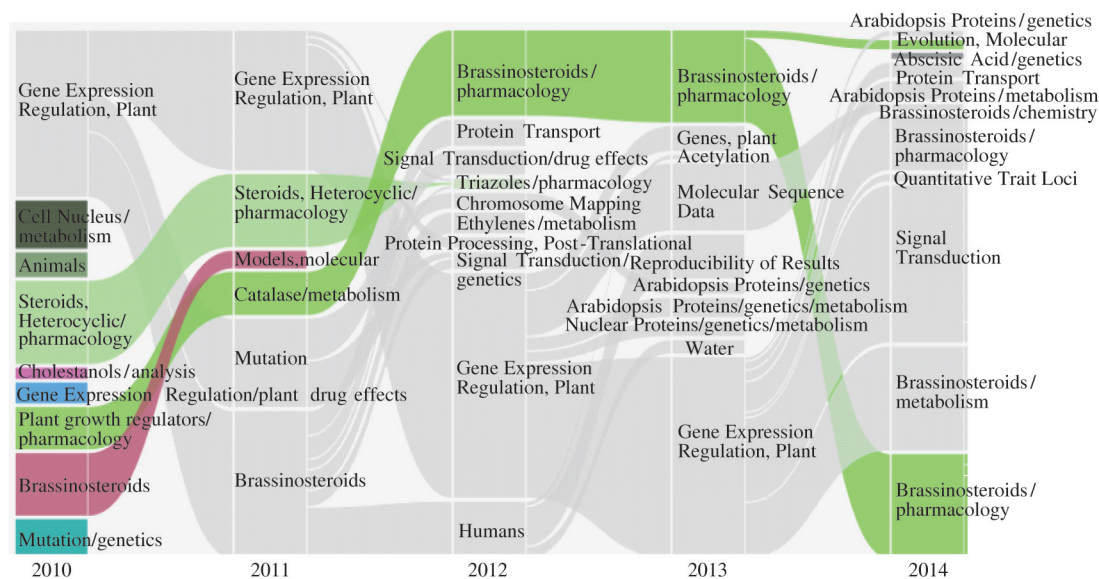


图2 植物甾醇激素领域的主题演化河流图 (Newman MM 算法)

体网络中单独提取出来,以河流图形式进行可视化展现,结果见图3。

如图3所示,2010年组成 Animals 主题的关键
词数量只有 29 个。2011 年该主题发生扩张,关键
词数量增加到 31 个。2012 年该主题进一步扩张,

关键词增加到 71 个,并且中心度最高的关键词由
Animals 演化为 Humans。2013 年该主题发生分裂
产生了 2 个新生主题:Stereoisomerism(立体异构)和
Acetylation(乙酰化)。

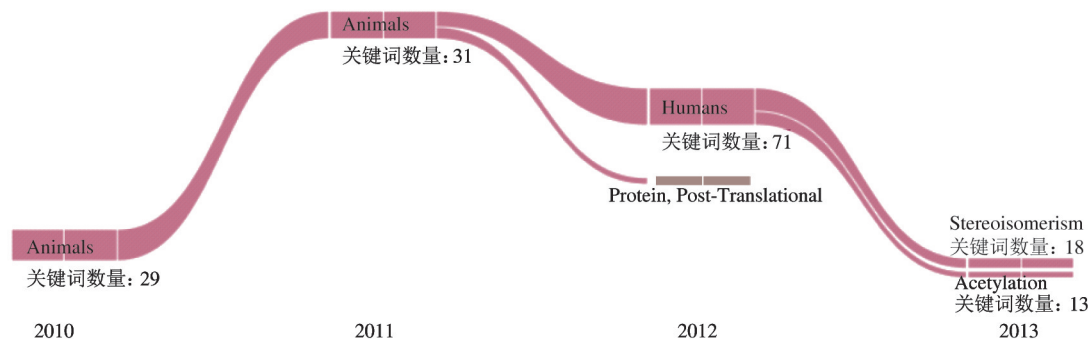


图3 主题 Animals 的演化河流图 (Blondel 算法)

为了验证 Blondel 算法获得的 Animals 主题演
化路径的准确性,需要借助专家知识对 2011 – 2014
年的文献主题进行标注和分析,如果发现有 Animals
相关研究组成的主题则说明 Blondel 算法较为准确。
如果没有发现 Animals 相关研究组成的主题,则说明
Newman MM 算法较为准确。因此中科院植物所
的领域专家受邀对文献集进行了主题标注,标注结
果如表 4 所示。

由表 4 可知,从 2010 年开始到 2013 年,与动物
有关的 Brassinosteroids 研究文献一直存在并呈上升

趋势,2010 年 4 篇,2011 年 6 篇,2012 年 7 篇,2013
年 9 篇。主题内容上可以分为 Brassinosteroids 对动
物细胞的毒理研究、植物甾醇与动物甾醇的功能比
较、Brassinosteroids 的化学修饰。

2010 年主题 Animals 的研究内容主要为 Brassi-
nosteroids 对动物细胞的毒理研究,其中有 3 篇的研
究对象为实验动物(小鼠、大鼠及牛),但也有 1 篇
关于人类胸腺肿瘤细胞的研究。2011 年研究该主
题的文献扩大为 6 篇,但研究还是以实验动物为主。
到了 2012 年研究该主题的文献扩大为 7 篇,其中有

表 4 专家对 Animals 主题文献知的主题标注结果

主题	主题细分	文献	年份
1. Brassinosteroids 的药理效应,对动物细胞的毒理研究	a. 小鼠或大鼠	Developmental toxicity of homo brassinolide in Wistar rats	2010
		Effect of the structure of the brassinosteroid side chain on monooxygenase activity of liver microsomes	2010
		Anabolic effect of plant brassinosteroid	2011
		Akt-dependent anabolic activity of natural and synthetic brassinosteroids in rat skeletal muscle cells	2011
		24-Epibrassinolide, a phytosterol from the brassinosteroid family, protects dopaminergic cells against MPP-induced oxidative stress and apoptosis	2011
		Hypoglycemic effects of brassinosteroid in diet-induced obese mice	2012
		Acceleration of cutaneous wound healing by brassinosteroids	2013
		A review of the phytochemistry and pharmacological activities of raphani semen	2013
	b. 牛肾上腺皮质细胞	Characteristics of steroid metabolism in transgenic Nicotianatabacum plants bearing the CYP11A1 cDNA of cytochrome P450(SCC) from the bovine adrenal cortex	2010
		Brassinosteroids cause cell cycle arrest and apoptosis of human breast cancer cells	2010
		Brassinosteroids inhibit in vitro angiogenesis in human endothelial cells	2012
		Mechanisms of natural brassinosteroid-induced apoptosis of prostate cancer cells	2012
	c. 人类细胞研究	Synthesis and biological activity of probable 24-norbrassinolide biosynthetic precursors	2012
		Increasing the amphiphilicity of an estradiol based steroid structure by Barbier-allylation-ring-closing metathesis-dihydroxylation sequence	2012
		Brassinosteroids and analogs as neuroprotectors: synthesis and structure-activity relationships	2012
		Sterol glycosyltransferases-the enzymes that modify sterols	2011
2. 植物甾醇与动物甾醇的功能比较,动物细胞研究	a. 细胞凋亡	Diverging functions among calreticulin isoforms in higher plants	2011
		Crosstalk in cellular signaling: background noise or the real thing?	2011
		The U-box E3 ubiquitin ligase TUD1 functions with a heterotrimeric G alpha subunit to regulate Brassinosteroid-mediated growth in rice	2013
	b. 信号传导	Brassinosteroid signalling	2013
		Synthesis of fatty acyl derivatives of 24-epibrassinolide	2013
3. Brassinosteroids 的化学修饰	a. 立体异构 (Stereoisomerism)	Design and studies of novel polyoxysterol-based porphyrin conjugates	2012
		Regio- and stereocontrolled synthesis of novel steroidal isoxazolines; a new route to the formation of selectively modified steroid side chains	2013
		Asymmetric synthesis and effect of absolute stereochemistry of YCZ-2013, a brassinosteroid biosynthesis inhibitor	2013
		A new approach to the side chain formation of 24-alkyl-22-hydroxy steroids: application to the preparation of early brassinolide biosynthetic precursors	2013
	b. 乙酰化修饰 (Acetylation)	N,O6-partially acetylated chitosan nanoparticles hydrophobically-modified for controlled release of steroids and vitamin E	2013

6 篇是关于动物细胞的研究,当年关于人类细胞的研究共有 5 篇文献,而关于小鼠的研究下降为 1 篇。2013 年该主题进一步发生分裂产生了 Brassinosteroids 的化学修饰这个研究主题,包括乙酰化修饰 1 篇,立体异构体 3 篇。这与 Blondel 算法获得的 Animals 主题演化路径基本是一致的。

由以上结果可知,从 2010 年开始到 2013 年,Animals 这个主题是存在的,并且其研究内容发生了从以实验动物为主到以人类细胞研究为主的转变。结合专家知识绘制了该主题的演化模式图,证明 Blondel 算法得到的 Animals 主题演化路径是比较准确的。如图 4 所示。

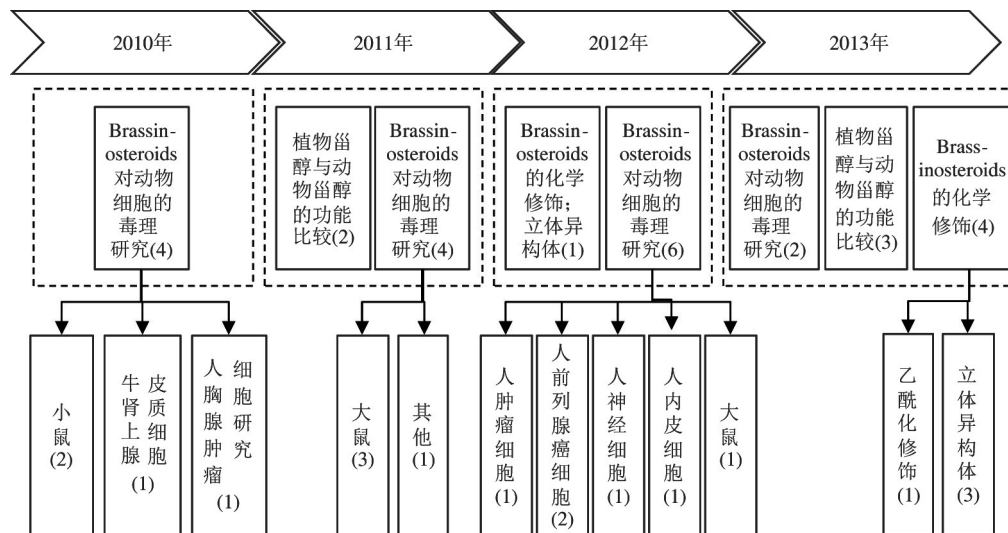


图 4 专家解读绘制的 Animals 主题演化图

4 结 论

本研究对 Blondel 算法、Newman MM 算法和 Ball Overlapping 3 种典型的社区发现算法的社区划分效果和准确性进行了对比研究。在植物甾醇领域的实验研究总结如下。

(1) 3 种社区发现算法的主题识别效果比较表明:在运算速度上 Blondel 算法最快, Ball Overlapping 算法次之, Newman MM 算法最慢。而在社区识别效果上 Blondel 算法和 Newman MM 算法比较接近, 识别到的社区其代表性关键词也具有较好的解释性。

(2) 基于河流图的植物甾醇激素领域的主题演化可视化结果表明:采用同样的社区演化追踪方法, 通过 Blondel 算法划分的社区侦测到的主题演化状态更为丰富, 而基于 Newman MM 算法划分的社区演化状态比较单一。

(3) 以 Animals 主题作为实例, 比较 Newman 和 Blondel 2 个算法发现主题的准确性可以看到:此主题在 2010 年的数据集中 Newman MM 算法和 Blondel 算法都有发现。不同的是, Blondel 算法在 2011 年侦测到 Animals 后继主题状态为扩张, 并且到 2013 年为止都有后续主题的发现; 而 Newman MM 分区算法则没有侦测到 Animals 的后继主题。以上为算法生成的演化路径。

(4) 该领域专家解释的演化路径显示, 从 2010 年开始到 2013 年, Animals 这个主题是一直存在的。并且研究对象经历了以实验动物为主到以人类细胞为主的转变。这证明 Blondel 算法得到的主题 Animals 演化路径是比较准确的。

根据植物甾醇领域文献的实验结果可以得知, 3 种算法中 Blondel 算法获得的主题和演化追踪最为准确, 它不仅可以很好地实现关键词共现网络的社区划分, 而且能更好地发现科研主题的演化。

基于复杂网络理论的社区发现和演化算法为研究科研主题演化追踪提供了一种新的思路。该方法得到的主题演化数据具有较好的解释性, 但同时存在一些不足之处。其一是这种方法比较依赖于关键词数据, 一般只适用于结构性比较强的文献数据; 其二是有些文献关键词的选择存在随意性^[30], 不一定能很好地反映文献的内在特征, 忽略了对文献内容的分析因而具有一定局限性。未来的工作可以尝试在该方法中融入基于文献内部特征提取的方法, 以实现非结构化数据主题演化的研究。

参考文献

- [1] 岳丽欣, 周晓英, 陈旖旎. 期刊论文核心研究主题识别及其演化路径可视化方法研究——以我国医疗健康信息领域期刊论文为例[J]. 图书情报工作, 2020, 64(5): 89-99
- [2] 张斌. 国内共享科研数据热点主题及演化分析: 从主

- 题模型视角[J]. 图书馆学研究,2020(14):11-18
- [3] 王艳东,李萌萌,付小康,等. 基于社交媒体共同网络的灾情发展态势探测方法[J]. 武汉大学学报:信息科学版,2020,45(5):691-698
- [4] 舒文琛,周恩国,李岱峰,等. 基于合著网络社区发现的情报学研究主题演化分析[J]. 情报科学,2020, 38(1):75-81
- [5] Chen C. CiteSpaceII: detecting and visualizing emerging trends[J]. *Journal of the American Society for Information Science and Technology*,2006,57(3):359-377
- [6] Ding Y. Community detection: topological vs. topical [J]. *Journal of Informetrics*,2011,5(4):498-514
- [7] Lu W, Wang J M, Hu J M. Analyzing the topic distribution and evolution of foreign relations from parliamentary debates: a framework and case study [J]. *Information Processing and Management*,2020,57(3):1-14
- [8] Jiang H, Liu Z, Liu C, et al. Community detection in complex networks with an ambiguous structure using central node based link prediction[J]. *Knowledge-Based Systems*,2020,195(12):1-13
- [9] Palla G, Deranyi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. *Nature*,2005,435(7043):814-818
- [10] Lancichinetti A, Fortunato S. Community detection algorithms: a comparative analysis [J]. *Physical Review E Statistical Nonlinear and Soft Matter Physics*,2009, 80(5):056117
- [11] Zachary W W. An information flow model for conflict and fission in small groups [J]. *Journal of Anthropological Research*,1977,33(4):452-473
- [12] Girvan M, Newman M E. Community structure in social and biological networks[J]. *Proceedings of the National Academy of Sciences of the United States of America*,2002, 99(12):7821-7826
- [13] Newman M E, Girvan M. Finding and evaluating community structure in networks[J]. *Physical Review E*,2004, 69(2):026113
- [14] Boyack K W, Klavans R, Brner K. Mapping the backbone of science [J]. *Scientometrics*,2005, 64(3):351-374
- [15] Lambiotte R, Panzarasa P. Communities, knowledge creation, and information diffusion[J]. *Journal of Informetrics*,2009,3(3):180-190
- [16] Danon L, Duch J, Diaz-Guilera A, et al. Comparing community structure identification[J]. *Journal of Statistical Mechanics*,2005,2005(9):09008
- [17] Chopra S, Shim S. A strong formulation for the graph partition problem[J]. *Networks*,2020,75(2):183-202
- [18] Navarro J F, Frenk C S, White S D M. A universal density profile from hierarchical clustering[J]. *Astrophysical Journal*,1996,490(2):493
- [19] Newman M E. Fast algorithm for detecting community structure in networks [J]. *Physical Review E Statistical Nonlinear Soft Matter Physics*,2004,69(6):066133
- [20] Wang X, Cheng Q, Lu W. Analyzing evolution of research topics with NEViewer: a new method based on dynamic co-word networks [J]. *Scientometrics*,2014, 101(2):1253-1271
- [21] Blondel V D, Guillaume J, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. *Journal of Statistical Mechanics: Theory and Experiment*,2008,10:1-12
- [22] Ball B, Karrer B, Newman M E. An efficient and principled method for detecting communities in networks [J]. *Physical Review E Statal Nonlinear and Soft Matter Physics*,2011,84(2):036103
- [23] Wu F, Huberman B A. Finding communities in linear time: a physics approach [J]. *European Physical Journal B: Condensed Matter*,2004,38(2):331-338
- [24] Capocci A, Servedio V D P, Caldarelli G, et al. Detecting communities in large networks[J]. *Physica A: Statistical Mechanics and its Applications*,2005,352(2):669-676
- [25] 王晓光,程齐凯. 基于NEViewer的学科主题演化可视化分析[J]. 情报学报,2013,32(9):900-911
- [26] Palla G, Barabási, Albert-László, et al. Quantifying social group evolution[J]. *Nature*,2007,446(7136):664-667
- [27] Garey M R, Johnson D S. Computers and Intractability: A Guide to the Theory of NP-Completeness [M]. New York:W. H. Freeman & Co.,1979:338
- [28] Huang X, Lakshmanan L V S, Xu J. Community search over big graphs[J]. *Synthesis Lectures on Data Management*,2019,14(6):1-6
- [29] Veremyev A, Semenov A, Pasiliao E L, et al. Graph-based exploration and clustering analysis of semantic

spaces[J]. *Applied Network Science*, 2019,4(1):1-26

[30] 牛奉高, 高旭霞, 李志欣. 基于文献的人工智能领域

研究的网络特征与演化分析[J]. 情报工程, 2019,5

(4): 33-43

Comparison of three typical community discovery algorithms in topic evolution research——a case study of Brassinosteroids research areas

Xue Shan, Dong Cheng, Han Hongqi, Zhang Junsheng, Gao Xiong, Wang Li

(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract

Identification and analysis of the evolution state of scientific research topics are important means to understand the current situation and trend of research topics. In recent years, community discovery algorithm based on complex networks and network evolution metrics has become one of the main methods to reveal scientific research topics. Therefore, it is necessary to explore the advantages and disadvantages of these algorithms in topic discovery and evolution research. Brassinosteroids hormone research area is selected as an example, the keyword co-occurrence network in this field is constructed, and the performances of three typical community discovery algorithms (Newman MM, Ball Overlapping and Blondel) in scientific research topic recognition are compared; the accuracy of the detection of research topic and the evolutionary analysis results is judged by expert knowledge. The results show that the Blondel algorithm has great advantages in the identification and evolution tracking of scientific research topics. It has not only the fastest speed of operation, but also the more accurate topic recognition. It provides a reference for the evolution tracking of scientific research topics from the perspective of algorithm applicability.

Key words: evolution of the research topic, community discovery algorithm, Brassinosteroids, keyword co-occurrence network