CSSCI来源期刊
RCCSE核心期刊
全国中文核心期刊
陕西省精品期刊
世界学术影响力期刊
中国国际影响力优秀学术期刊
中国人文社会科学核心期刊
“复印报刊资料”重要转载来源期刊
ISSN1002-1965
CN61-1167/G3

情报杂志

JOURNAL OF INTELLIGENCE

QINGBAO ZAZHI QINGBAO ZAZHI

ISSN 1002-1965
11>
9 771002 196220

11 Vol.41

陕西省科学技术情报研究院

中国·西安

2022

JOURNAL OF INTELLIGENCE

Vol.41 No.11 2022

MAIN CONTENTS

Sponsored by

Institute of Scientific and Technical Information of Shaanxi
Society of Scientific and Technical Information of Shaanxi

Edited & Published by

Editional Board of JOURNAL OF INTELLIGENCE
Chief Editor: Zhang Wei

99 Yantalu Xi'an,710054 P.R.China

Tel: (029) 85529749

情报杂志 月刊

Qingbao Zazhi

2022 年 No.11

创刊/1982 年

【第 41 卷　2022 年 11 月 10 日出版】

目　次

编　委　会

主　管：陕西省科学技术厅
主　办：陕西省科学技术情报研究院
主　编：薇　子
副主编：白燕琼
编　辑：王平军　贺小利　刘影梅
　　万园园　王　菊　王育英
本期责编：王育英
出　版：《情报杂志》编辑部
编　辑：《情报杂志》编辑部
地　址：西安市雁塔路南段 99 号
邮　编：710054

电　话：（029）85529749
网　址：http://www.qbzz.net/
E-mail：qbzz@263.net
中国标准刊号：ISSN1002—1965 / CN61—1167/G3
邮发代号：52—117
国外邮发代号：M5090
发　行：中国邮政集团公司陕西省报刊发行局
订　阅：全国各地邮局
印　刷：陕西盛世大宇印务有限公司
定　价：28.00 元

文献耦合网络的引文内容加权研究*
——基于提及次数的方法

卢 超[1] 董 克[2,3]

(1.河海大学商学院 南京 211100;2.武汉大学信息资源研究中心 武汉 430072;
3.武汉大学信息管理学院 武汉 430072)

摘 要:[**研究目的**]通过内容特征对引文网络加权,是实现细粒度计量分析的重要途径。系统揭示内容特征加权策略对引文网络结构形态的影响,有利于深化对引文网络形成机理与应用的认识。[**研究方法**]基于多数据源的计量网络与内容结合方法,融合引文提及次数特征与文献耦合网络,提出了4种内容加权策略,以解决单一机构或期刊公开的引文内容数据不适用于大规模计量网络内容加权的问题,并研究耦合网络的基础形态。[**研究结论**]结果表明该文提出的融合方法具有可行性;引入被提及次数相关特征权重不改变文献耦合网络中节点和边的数目;在内容加权处理的文献耦合网络中,边的权重分布、节点度分布以及节点中心度等指标均有变化;网络中中间中心度较高的节点略有减少,表明内容加权的耦合网络更具连通性。

关键词:文献耦合网络;引文内容特征;引文提及次数;自然语言处理;复杂网络分析

中图分类号:G350 **文献标识码:**A **文章编号:**1002-1965(2022)11-0171-08

引用格式:卢 超,董 克.文献耦合网络的引文内容加权研究[J].情报杂志,2022,41(11):171-178.

DOI:10.3969/j.issn.1002-1965.2022.11.025

Study on Adding Weight to Bibliographic Network Using Citation Content
——A Method Based on Citation Mention

Lu Chao[1] Dong Ke[2,3]

(1. Business School, Hohai University, Nanjing 211100;
2. Center for Studies of Information Resources, Wuhan University, Wuhan 430072;
3. School of Information Management, Wuhan University, Wuhan 430072)

Abstract: [**Research purpose**] Adding weights to citation networks via content features becomes a vital means to fine-grained bibliometric analyses. A systematic discovering of the influence of content feature weighting strategies on the structural form of citation networks is beneficial to deepening our understanding of the forming mechanism of citation networks and their applications. [**Research method**] This study proposes a method—combining multiple citation content features and bibliographic coupling network and designing 4 content-weighted strategies to overcome the barrier caused by the situation where usually only one publisher or journal provides the structured full-text data, which limits the application of the content-based co-citation network on large-scale datasets. Via our proposed 4 strategies to add content weight to the networks, we are able to merge the bibliographic coupling networks with citation mention features. [**Research conclusion**] The results suggest that adding citation mention features to the network, the number of nodes and edges are not changed; and that the weight distribution of nodes, weighted degree distribution and betweenness centrality of nodes are changed. Number of nodes with high degree of betweenness centrality declines, indicating better connections within each content-weighted network.

Key words: bibliographic coupling network; citation content features; number of citations; natural language processing; complex network analysis

收稿日期:2022-06-28 修回日期:2022-08-05

基金项目:国家自然科学基金青年项目“劳动分工视角下科研合作者的科研效能研究”(编号:72004054);武汉大学自主科研项目(人文社会科学)“基于计量分析的我国社科学术态势分析框架”;中央高校基本科研业务经费专项资金项目“科研团队多样性对科研绩效的因果效应研究”(编号:B220201058)。

作者简介:卢 超,男,1991年生,博士,讲师,研究方向:科学学、科学合作;董 克,男,1985年生,博士,副教授,研究方向:文献计量。

0 引 言

利用引用关系构建各类网络进行文献计量研究是图情领域重要的研究内容,引用关系衍生来的引文网络、共被引网络、耦合网络广泛应用于科学主题探测、影响力评价、引文推荐等领域[1]。从复杂网路理论在文献计量学中的应用来看,通过文献间互引关系构建的引文网络是将学术文献抽象成点,并保留点与点间的引用关系。因引用行为有其合法性和目的性,通过引用关系构建的引文网络对解决相关研究问题亦具有其合理性。

然而,从学术论文集到引文网络的抽象过程存在许多局限。举例来看,一个研究话题可表示为若干相关论文的集合,相关内容可用其所有文献的全文本内容表征;通过引文网络(或社区结构)表征研究话题,其抽象过程损失了研究话题本身大量的内容特征。具体来看,一篇学术论文的内容包括两个方面[1]:全文本内容,即其作者解决研究问题过程和结果的阐述;引文内容,即其作者为更好陈述其研究报告而对所引文献的述评。此种述评性的引文内容构成了学术论文间的引用关系。抽象学术文献时,其全文本内容被破坏性地压缩甚至消除,其中的引文内容也被简化为引用数字 0 和 1。这为研究话题的细粒度发现及影响力评价带来极大阻碍[2]。

近年来,文献内容特征广泛应用于网络结构分析。研究表明,内容特征加权作者共被引[3-4]、期刊耦合[5]网络,能优化知识结构和话题识别效果[2-3,6]。计量网络分析和内容分析的有机结合成为重要的研究方向[7]。同时,内容特征对引文网络构建的影响机理尚未充分探索,这导致方法论层面的研究与应用缺乏标准[8]。系统揭示内容特征加权与引文网络结构形态间的关系,是研究话题识别[3]、学术影响力评价[9]等应用研究有效实施的必要基础。

作为一种典型的计量网络,文献耦合网络在影响力评价、引文推荐等研究中应用广泛,特别在研究前沿探测上有一定优势。与其他计量网络相比,文献耦合网络虽基于引用关系建立,但其无需额外全文数据便可开展全文内容和引文网络相结合的研究[5],这一定程度上缓解全文内容来源不足的局限性。然而,已有耦合网络研究对其网络形态的认识依旧存在许多不足[10],特别是内容特征与文献耦合网络的融合研究还较为少见。

针对上述问题,本文提出了一种基于多源数据的文献耦合网络与引文内容数据融合的方法,在结构化全文数据不足的现实情况下,提出文献耦合网络内容加权的研究思路与技术路线,探索融合内容特征的文献耦合网络形态基本特征,以求为相关研究的复现提供借鉴。

1 研究框架

本文研究框架如图 1 所示。首先,使用 Python 爬虫脚本爬取 PLoS 学术论文全文数据,并从 WoS 引文数据库中获取相应的引文数据和学科信息;其次,对所获取 XML 格式的全文数据进行解析,获取其元数据、引文内容特征,并对抽取的内容特征进行量化;第三,选取目标学科构建引文网络,包括文献元数据融合、耦合网络构建、加权策略设计以及内容加权网络构建;最后,比较分析已构建的经典耦合网络和内容加权耦合网络。

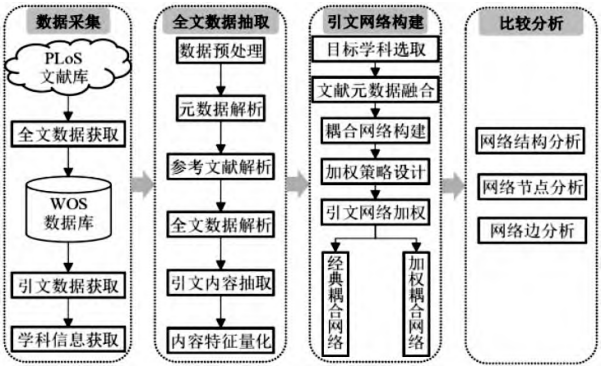

图 1 研究框架

1.1 实验数据获取

1.1.1 PLoS 全文数据及其采集

目前,WoS 和 CSSCI 是代表性的引文数据库,能提供较少噪音的“清洁”数据,但并不提供文献全文。利用学术搜索引擎也可获取引文数据,但同样缺乏结构化全文。几乎所有的全文数据库均提供 PDF 全文,部分提供 html 格式结构化全文,如 Wiley Online 和 Elsevier[11]等。但这些数据库均需付费访问,且限制用户采集和使用数据。期刊方阵里,*Nature*、*Science* 等顶级期刊以及图情领域一些期刊也陆续提供全文数据,但版权会限制数据的采集和使用,且单个期刊对研究主题限制明显;PLoS 旗下所有刊物均提供 XML 格式全文数据,数据处理接口丰富,并且开放获取,为内容与网络结合研究提供更多便利[12-13]。

本文选取 PLoS 中生物医学领域为数据对象,通过联合 PLoS 全文数据和 WoS 数据,构建生物医学领域的内容特征加权文献耦合网络。PLoS 全文数据的采集包括两个步骤:数据的检索与爬取。a.构造检索式进行检索。构造检索式“publication_date:[2003-01-01T00:00:00Z TO 2016-01-15T23:59:59Z]”,从 PLoS 网站共检得 2003 年 1 月 1 日到 2016 年 01 月 15 日期间所有文献,共 176,310 篇(检索链接:https://re-url.cc/GEVAWG),含研究 163,389 篇、综述 471 篇,均

为 XML 格式全文。b.爬取数据。根据检索结果共获得 2,939 个分页网址,每个分页面 60 条记录。利用分页源码获得所有 PLoS 论文绝对链接。利用论文链接,爬取 XML 格式文件。文件记录了文章、作者与被引文献等各类信息。最终,除缺少全文的文献 15 篇,共获取文献 176,295 篇。

1.1.2　WoS 引文数据及其获取

本研究所使用的 WoS 数据来自加拿大蒙特利尔大学 Vincent Larivière 博士提供的 WoS 引文数据,共包括 3 张数据表格:a.WoS_citing。即 WoS 数据库中所有 PLoS 论文信息表,包括 doi、学科等数据。该表共含 218 135 篇论文。b.WoS_citation。即 PLoS 文献及其引文的引用关系表,共有记录 16 646 196 条。该表含 WoS 馆藏号、doi 等信息。c.WoS_ref。即 PLoS 文献引文的元数据表,共有记录 6 808 405 条。

其中,WoS_citing 表包含的 doi 和学科信息用于与 PLoS 全文数据建立连接、识别 PLoS 文献的学科归属;WoS_citation 表用来构建经典文献耦合网络;WoS_ref 表用来和 PLoS 论文的引文信息表进行匹配,预备后期的内容加权网络构建。这 3 张数据表包含 3 种文献身份识别码:doi,WoS 文献馆藏号和数据库本地文献序号,article_id。当某一字段值大量缺失时,可用其他字段进行数据融合,保证数据匹配度和准确率。

1.2　PLoS 全文数据解析与处理

1.2.1　PLoS 全文数据解析

本文使用 NLTK 处理全文数据的分句任务,识别引文内容边界;使用 re 正则匹配全文数据中的关键节点,如引文标记、结构标记等;使用 Elementtree 解析 XML 文档及其结构信息。文献全文数据解析包括文献元数据解析、文献全文解析和参考文献解析三个模块。a.文献元数据解析。文献元数据解析在<article.front>标签区域内析出文章元数据,构建两张表:文章信息表和作者信息表。文章信息包含文献的 doi、标题等信息。作者信息表记录包括前五位作者的姓名、单位等信息。b.文献全文数据解析。文献全文数据解析包括引文内容数据的抽取。通常,学者们选取引文所在句子作为引文内容,但 Teufel 也指出被引文献前后 50 个单词的窗口长度最优[15]。综合以上方案,本研究采用以引文所在句子为中心句,前后最多采集两句的形式,采集最多 5 句话构成一条完整的引文内容。当引文中心句处在段落中,其前后句子数量多于 3 句,本研究则采集 7 句;当引文中心句处在段落边缘,即中心句前后句子数量不足 3 句时,采集的句子总数会在[1,6]之间。c.参考文献解析。参考文献解析后可利用参考文献的作者、标题和时间等信息匹配 WoS 数据库中的引文数据。这样引文数据可与全文数据相融合,为融合内容特征与结构特征打下基础。每篇参考文献存有一个编号在<label>标签里,用于建立与引文内容的映射。参考文献的作者、标题等也会被记录下来。

1.2.2　PLoS 文献学科归属的确定

确定研究话题有利于利用统一口径的学科标准对学术影响力进行归一化[16]。PLoS 根据其机构制定的学科体系为每篇发表的文献提供了学科标签,数量一般为 3~5 个,故很难依据这些多分类的信息来划分每篇文献的学科归属。本研究采用了 Vincent Lariviere 提供的文献学科分类数据,该学科分类数据的分类体系来源于 NSF 的学科分类体系[17]。经过 PLoS 和 WoS 数据集的匹配,得到 180 293 篇可识别身份的文献,其中 140 305 篇文献能利用这种算法得到学科标签。这 140 305 篇文献的学科分布如表 1 所示,其中,约 45.4%的文献从属于临床药学研究、35.5%的研究从属于生物医学研究、9.6%的文献属于纯生物学,仅有约 1%的文献属于人文社科学科。本文选取生物医药(Biomedical Research)和生物学(Biology)为目标学科,因二者间的交叉度高,联合二者可保证文献集的完整性。下文使用"生物医药学"作为两个学科的合称。表 2 显示生物医药学包含的子领域,共计文献 63 279 篇。

表 1　PLoS 研究论文的学科分布表

学科	数量	比率(%)
Clinical Medicine	63 729	45.4
Biomedical Research	49 848	35.5
Biology	13 431	9.6
Psychology	3 120	2.2
Mathematics	2,911	2.1
Earth and Space	1 666	1.2
Health	1 227	0.9
Engineering and Technology	1 125	0.8
Physics	1 119	0.8
Social Sciences	995	0.7
Chemistry	638	0.5
Professional Fields	471	0.3
Humanities	21	0.0
Arts	1	0

表 2　生物医药学科的领域分布

领域	数量	比率(%)
Biochemistry & Molecular Biology	21 438	33.9
Genetics & Heredity	11 159	17.6
Microbiology	5 779	9.1
Ecology	4 031	6.4
Parasitology	3 886	6.1
Botany	3 170	5.0
Virology	2 514	4.0
Cellular Biology Cytology & Histology	2 006	3.2
Marine Biology & Hydrobiology	1 915	3.0
Entomology	1 272	2.0
General Biology	1 165	1.8
Physiology	1 046	1.7
Agricult & Food Science	778	1.2
Nutrition & Dietetic	578	0.9

1.2.3 引文提及次数相关特征的抽取与计算

学术文献的影响力受引文的被提及次数影响较大[2-3,18]，其形式上具有简洁性强和可计算等优势，为应用于网络构建提供便利[19-20]。故在使用相关特征加权文献耦合网络时，本文着重考虑利用引文被提及次数相关特征为文献耦合网络的边加权，即引文被提及次数和引文平均被提及次数。

a.被提及次数。参考其概念[2]，可计算 PLoS 文献中任一引文的被提及次数。由引文内容表，联立公式(1)和公式(2)可得对于任一篇 PLoS 文献 i 的参考文献 j 的提及次数 M_j^i。

$$M_j^i = \sum_1^n cv_k^i, 1 \leqslant k \leqslant n \tag{1}$$

$$c\,v_k^i = \begin{cases} 1, \ if\ c_k^i has j \\ 0, if\ not \end{cases} \tag{2}$$

其中，c_k^i 表示文献 i 中的一个引文内容的片段，cv_k^i 表示引文内容进行量化处理后得到的值。

b.平均被提及次数。根据引文内容表 content_all 可得到每个引文内容中参考文献总数[2]。据此可得到一参考文献 j 在任一篇 PLoS 文献 i 中平均被提及次数 $M_{meanj}{}^i$。

$$M_{meanj}{}^i = \sum_1^n \frac{cv_k^i}{ct_k^i} \tag{3}$$

其中，ct_k^i 表示一引文内容片段中提及的参考文献总数。联立公式(2)和(3)可求 $M_{meanj}{}^i$。

1.3 内容加权网络构建与网络分析指标

如图 1 所示，为构建内容加权的文献耦合网络，首先，对 WoS 数据和 PLoS 数据进行匹配与整合，消除无法匹配的数据；其次，利用融合后的数据构建文献耦合网络，根据引文内容特征，制定基于提及次数的内容加权策略并构建内容加权的文献耦合网络；最后，对构建的多种文献耦合网络进行网络属性(如度分布、聚类系数)分析，比较其异同。

1.3.1 异源数据的匹配

在 PLoS 的全文数据中，施引文献的键值是 doi；参考文献间无完整的的身份识别字段。在 WoS 数据中，文献记录间的独特识别字段是馆藏号 WoS_Id 以及数据源提供的独特识别字段 Article_Id。由于 WoS 并没有完整收录所有文献的 doi 信息，故构建网络的过程中需要对两方数据进行匹配和关联，包括关联两方施引文献、被引文献以及引文内容与被引文献。a.施引文献的匹配。在确定 WoS 数据和 PLoS 数据之间施引文献的身份时，我们利用了 PLoS 文献中的 doi 字段，将所有在 WoS 中无法识别出 PLoS 文献 doi 的文献记录删除，为匹配到的文献之间建立关联。这样做的三个理由：首先，PLoS 中文献之间的 doi 信息完整全面，匹配的准确度高；其次，WoS 数据库中常有 doi 信息错误的情况存在，无法通过 WoS 数据进行映射；最后，利用其它字段信息进行匹配会引发其它问题，如字段信息的消歧问题等。在这三个步骤中，本研究从初始的 WoS 数据中得到生物医学方面的 PLoS 文献共计 63 279 篇，经过匹配 WoS 中的引文数据，得到 PLoS 中的施引文献 63 278 篇，参考文献 1 354 225 篇，引用关系共有 2 851 627 条。b.被引文献的匹配。在 PLoS 数据库中，参考文献的各个字段需要从全文数据中的相关字段中进行采集，因数据格式等问题，抽取质量无法保证。故本文采用字符串匹配的方式对 WoS 和 PLoS 的被引文献进行匹配。匹配中，本研究利用参考文献的标题和第一作者信息构成进行匹配的字符串，过滤字符串中的非数字字母字符；在同一个施引文献中(利用 doi 信息)找出两个数据源中最相似的两篇被引文献建立关联。这样为所有的 PLoS 文献中的参考文献找到其在 WoS 中对应的被引文献。在匹配过程中，本研究发现存在极少数被引文献的 WoS_Id 存在多条不同记录的情况；同时存在 4 031 篇 PLoS 文献的作者错误将同一条参考文献进行了重复引用。由于单篇数据量非常少，本研究选择移除这些错误的数据。经过匹配，得到 PLoS 文献 63 214 篇，被引文献989 016 篇，合计直接引用关系 2 038 854 条。c.引文内容与被引文献的匹配。在对被引文献进行关联之后，本研究利用在 PLoS 抽取的引文内容与参考文献的共同编号对进行匹配过后的引文内容以及被引文献进行关联匹配，共得到 PLoS 文献 62 366 条，被引文献986 828，直接引用关系 2 036 416 条。对此次匹配造成的引用关系缺失，则利用前面步骤获取的引文关系数据进行填充处理。以上，本文实现了 WoS 数据与 PLoS 全文数据的匹配。

1.3.2 文献耦合网络的内容加权策略

PLoS 中的文献及其被引文献经过匹配和消歧过后，最终得到 PLoS 文献 63 026 篇，耦合关系 12 050 612 条。进一步地，本研究将施引文献的引文被提及次数、引文平均被提及次数等两个主要特征对得到的耦合网络的边进行内容特征加权。构建经典文献耦合网络时，两篇耦合文献所构成边的总权重等于这两篇文献耦合的次数。当考虑被引文献在施引文献中被提及次数时，耦合文献的边权需重新调整，如图 2 所示。在经典耦合网络基础上，本研究通过引入不同的内容特征，设计了 4 种内容权重处理策略 s_1，s_2，s_3，s_4 来进一步探究引入内容权重对构建文献耦合网络的影响，并将耦合网络构建策略 s_0(即经典文献耦合网络的构建策略) 的边权结果 作为研究分析的参照。

策略 s_1：该策略仅考虑引文在施引文献中的出现

频次,即引文被提及次数,作为文献耦合网络的内容权重。本研究采用算数平均的方法来调和耦合文献之间权重大小,每一篇共被引文献 i 对其施引文献组成的耦合对(A,B)的权重贡献值 $\omega_i^{A,B}$ 如公式(4)所示。

$$\omega_i^{A,B} = \frac{m_i^A + m_i^B}{2} \tag{4}$$

其中, m_i^A 表示文献 i 在文献 A 中被提及次数, m_i^B 表示文献 i 在文献 B 中被提及次数。耦合对(A,B)总权重 $\omega^{A,B}$ 如公式(5)所示。

$$\omega^{A,B} = \sum \omega_i^{A,B} \tag{5}$$

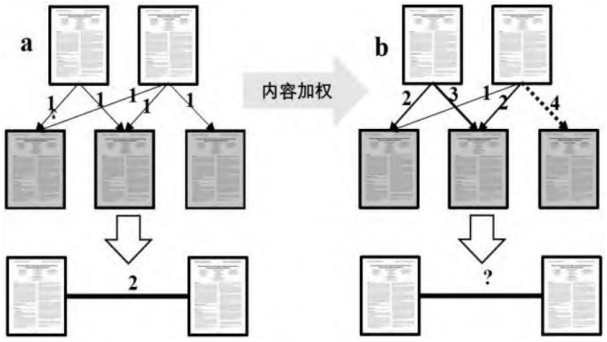

图 2　考虑内容权重情境下文献耦合网络权重的计算问题

策略 s_2 :在考虑引文被提及次数的基础上,进一步引入引文共被提及次数对提及次数的影响,利用共被引次数对每一次提及进行归一化,然后再进行如策略 s_1 中计算方式得到每对耦合文献之间的总权重。故每一篇共被引文献 i 对其引文献组成的耦合文献对(A,B)的权重贡献值 $\omega_i^{A,B}$ 如公式(6)所示。

$$\omega_i^{A,B} = \frac{\sum \frac{1}{c\, m_i^A} + \sum \frac{1}{c\, m_i^B}}{2} \tag{6}$$

其中, cm_i^A 表示文献 i 在文献 A 中被提及时,该引文句所提及参考文献总数; cm_i^B 表示文献 i 在文献 A 中被提及时,该引文句所提及参考文献总数。在计算出每一篇共被引文献和耦合文献对(A,B)的权重之后,本研究继续利用公式(5)计算耦合文献对(A,B)的总权重 $\omega^{A,B}$ 。

策略 s_3 :在考虑引文被提及次数的基础上,同时考虑不同施引文献的出版时间(即引文年龄)的影响。具体做法是,考虑发表时间更晚的施引文献对被引文献的评估比更早的施引文献更具有效性和可参考性。故每一篇共被引文献 i 对其施引文献组成的耦合文献对(A,B)的权重贡献值 $\omega_i^{A,B}$ 如公式(7)所示。

$$\omega_i^{A,B} = \frac{(PY_A - PY_i)\, m_i^A + (PY_B - PY_i)\, m_i^B}{PY_A + PY_B - 2\, PY_i} \tag{7}$$

其中, PY_i 表示文献 i 的发表时间。在计算出每一篇共被引文献和耦合文献对(A,B)的权重之后,利用公式(5)计算耦合文献对(A,B)的总权重 $\omega^{A,B}$ 。

策略 s_4 :在考虑引文被提及次数与参考文献年龄基础上,继续考虑共被提及次数对提及次数的分配效应。故在策略 s_4 中,每一篇共被引文献 i 对其引文献组成的耦合文献对(A,B)的权重贡献值 $\omega_i^{A,B}$ 如公式(8)所示。最后,利用公式(5)计算耦合文献对(A,B)总权重 $\omega^{A,B}$。

$$\omega_i^{A,B} = \frac{1}{PY_A + PY_B - 2\, PY_i}\left((PY_A - PY_i)\sum \frac{1}{c\, m_i^A} + (PY_B - PY_i)\sum \frac{1}{c\, m_i^B}\right) \tag{8}$$

1.3.3　实验分析指标

本文从网络规模、节点度分布和网络中心度三方面评估包含经典耦合网络在内的 5 个耦合网络的结构形态异同。

a.网络规模。本研究用网络的节点数、边数和网络密度共同来分析这 5 个网络的差异。通过考察不同网络的节点数,可研究不同策略构建的文献耦合网络的数据丢失情况。通过边数和网络密度,可准确了解已构建网络内部的连通性以及不同策略对引用关系构建的影响。

b.网络节点度分布。网络中节点度分布可反映该网络的基本结构形态以及节点之间的基本的连通性质。相关研究表明社交媒体中仅有少部分用户拥有大量好友,显示其意见领袖地位[21];幂律分布网络中弱连接对网络稳定至关重要[22]。故本文将首先考察这 5 种网络边的权重分布,然后对网络的度分布进行分析,查看不同网络间结构的稳定性以及不同权重设置策略对网络结构的影响[22]。

c.网络中心度。中心度一直是度量网络中节点连通性和网络结构的重要指标,因此本文将从中间中心度的视角分析本研究生成网络的中心度的异同。中间中心度根据公式(9)可衡量整个网络的流通效率,其中, x, y 是网络 G 中的任意两个不同于节点 i 的两个节点, p_{xy} 指节点 x, y 间最短路径数, $p_{xy}(i)$ 是指所有经过节点 i 的 x,y 间最短路径数。具有高中间中心度的节点往往显示较高的新颖性[23]。

$$C_B(i) = \sum_{x \neq i \neq y \in V} p_{xy}(i) / p_{xy} \tag{9}$$

2　实验结果分析

2.1　网络规模

如图 3 所示,总体上,利用提及次数特征构建的文献耦合网络与经典文献耦合网络具有相同的耦合文献数量 63,026。因为传统文献耦合网络在构建耦合文献对时考虑文献在全文范围内的引用关系,这与利用提及次数特征构建耦合关系对时所选取的文本范围是

一致的。通常,在施引文献没有出现错误引用的前提下,被引文献一定会同时出现在正文中和参考文献中。在本研究中,发现有极少数文献在参考文献部分重复标注了同一篇被引文献或者在正文处漏标了参考文献等错误。由于错误样本极少,本研究直接过滤了这一部分有错误的数据。同时,我们也注意到由于本研究的匹配算法无法取得 100%的召回率,因此利用策略 s_0 得到的耦合关系对利用提及次数特征得到的耦合网络进行修正。对于修正的边的权重,本研究利用了边权的中间数对缺省值进行填充。总之,利用被提及次数能得到和传统方式相同数量的文献耦合关系。

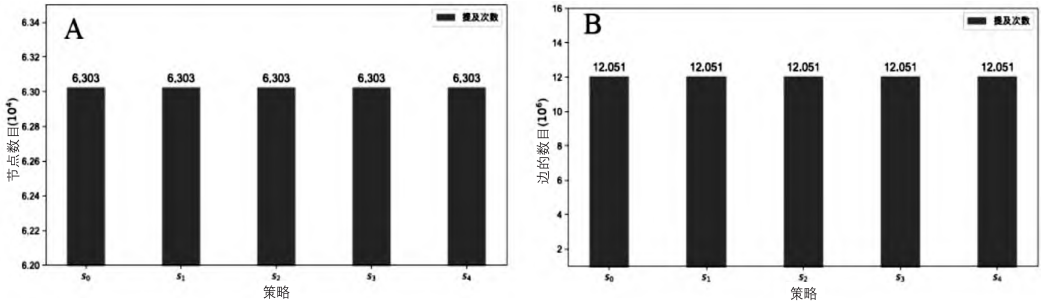

图 3 不同策略下构建的文献耦合网络的大小:(A)节点数目和(B)边数目

边数上,仅使用被提及次数信息构建的耦合网络具有边 12 050 613 条,这与传统方式构建的文献耦合网络的边的数量相同。本研究进一步分析了不同权重计算策略下网络密度的相互关系。总体来看,本研究中的文献耦合网络的密度都比较小,均为 0.005。由于相关研究通常不报告这一指标,与 Jarneving 的研究对比分析,本试验得到网络密度和该研究构建的网络密度相当[24]。因使用提及次数不对网络大小产生影响,故密度不变。

2.2 网络节点度分布

本研究构建的 5 种网络节点的权重分布如图 4(A)所示。总体来看,5 种网络的权重的分布函数在双对数的坐标系下近似呈直线,这表明网络的权重分布近似服从幂律分布,网络节点中的权重具有无标度性。网络大部分的节点权重较小,仅有一小部分节点具有很高的权重[25]。具体来看,在不使用内容特征对文献耦合网络进行加权时,耦合网络中边的权重大部分集中在[1,3],占所有边数量的 98%(如图中策略 s_0 所在曲线所示)。当使用提及次数的特征时,边的权重显著超过了经典策略,如策略 s_4 所在曲线所示。同时,策略 s_1 和 s_3 ,s_2 和 s_4 分别显示了相似的权重分布。这几组权重分别使用了加权平均的方法计算两施引文献间被提及次数的值以及利用引文年龄调节施引文献中被提及次数。这表明年份相近的文献更有可能被引用在一起。引入平均共被提及后,耦合网络的权重分布也出现了较大的变化。对比策略 s_1 和 s_2 ,策略 s_1 在图中的曲线显著高于 s_2 所在的曲线。

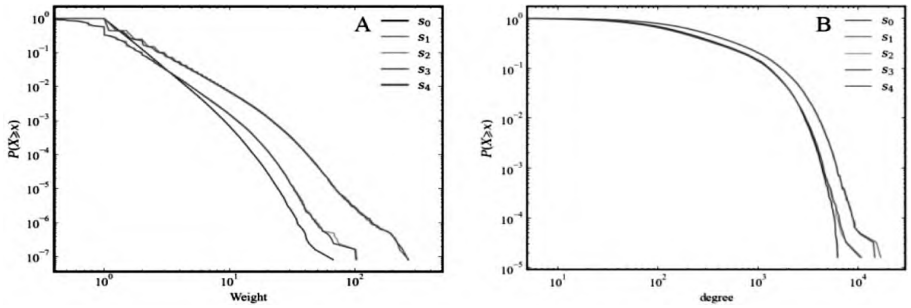

图 4 不同策略下耦合网络节点度分布互补累计分布图:(A)节点的权重;(B)含边权的节点度

综上,我们可看到在利用被提及次数计算得到的权重间有较高相似性,多集中在[1,3],高权重值的边数较少;在引入共被提及次数特征后,被放大的边权重被明显缩小,具备中等高的权重分布曲线;最后,被引文献年龄并不能区别调节耦合文献之间的权重分配。

5 种网络考虑边权的节点度分布如图 4(B)所示。本文中,各网络节点度分布考虑了节点间边的权重。图中网络度分布不再像边权重近似服从幂律分布,而更近似于服从指数分布(在双对数坐标轴上函数图像呈抛物线状)。从函数图像上看,经典文献耦合网络中,80%的节点的度小于 700。与该方法得到近似度分布的权重策略有 s_2 和 s_4 。这两种策略中,网络的节点数与 s_0 策略得到的网络的节点与边的数目相同,因为本研究的度分布计算考虑了边权重。在考虑边权重时,节点的度是所有连接该节点边的权重之和。故策略 s_2 和 s_4 得到网络边的权重是利用共被提及次数进行平均所得。对于一个节点的所有边来看,这些权重之和就近似等于 s_0 求得的权重的和。然而,在考虑被

引文献提及次数的情况下,相较于策略 s_0 ,由策略 s_1 和 s_3 得到的耦合网络中的节点度分布具有更高的概率分布;当度超过 800 时,差异更明显,这部分的节点占据了网络中总结点数的约 90%。类似于上面权重计算的结论,引入时间方面的信息并不能对节点度的分布产生明显的影响。

总的来看,文献耦合网络的度分布(考虑节点权重)近似符合指数分布,不具备长尾特性。在引入被提及次数特征时,我们得到文献耦合网络的度分布曲线处在较高位置,显示了网络中更强的连通性;当考虑被引文线的共被提及次数时,耦合网络的度分布退化为经典文献耦合网络的度分布,网络的连通性有一定减弱;被引文献的年龄在这一部分同样显示了较弱的调节能力,其本质原因是由于其在节点的边权的确定上缺乏调节能力。

2.3　网络中节点中心度

为进一步探测 5 种文献耦合网络的结构特性,本研究统计了这些网络中节点的中间中心度。针对中间中心度的计算,本研究采用采样估计的方式来计算各个点中间中心度,采样的比率为整个网络节点的 1%[26]。为计算某结点的中间中心度,我们选取约 6 200个点对该节点的中间中心度进行估计(网络节点大小见图 5 (A))。由于网络中大部分节点的中间中心度的值普遍较小($< 10^{-5}$),因此我们筛选了中间中心度值不小于 10^{-4} 的节点进行互补累计分布图的绘制。各个网络中筛选到的节点数目见图 5(A)。整体来看,由于整个网络的密度较小,因此网络中节点的中间中心度的值均普遍较低。其中,在利用提及次数构建的耦合网络中,我们筛选得到的高中间中心度的节点的个数最少。策略 s_1 和 s_3 仅分别得到了 15 个和 22 个值高于 10^{-4} 的节点,这样的结果可能是由采样的随机性误差造成的。

各个网络节点中间中心度的分布见图 5(B)。从图中可看出,使用被提及次数和共被提及次数特征的的耦合网络(s_2 和 s_4 曲线所示)显示了较高的中间中心度的分布趋势,表明网络中可能存在更多的社区结构。排除策略 s_1 和 s_3 ,可发现传统权重策略 s_0 所呈现的分布曲线最低,表明传统方法构建的耦合网络节点间中介性强度差异不明显。可能原因是文献间的耦合强度都比较接近,加之网络也比较稀疏,节点的中介性也不容易区分开。

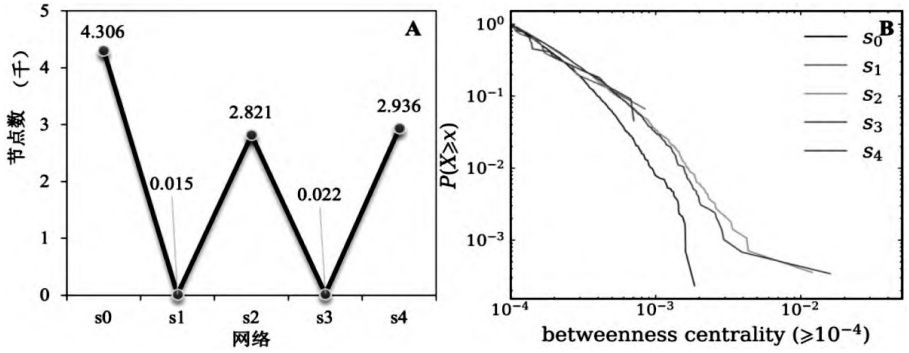

图 5　各网络中中间中心度值高于 10^{-4} 的节点数(A)及其分布(B)

3　研究结论

本研究选取了 PLoS 中的生物医药学领域作为目标学科领域。通过使用自然语言处理方法、复杂网络相关方法,将 PLoS 文献数据和 WoS 数据进行了融合消歧。利用自然语言处理技术抽取了该学科的引文内容,将引文内容转化成可量化的特征,设计了 5 种内容加权策略(包含无内容加权的方案)。通过网络结构形态分析,发现被提及次数相关特征的加权策略不改变网络节点和边的数目;在内容加权处理的网络中,节点的权重分布、度分布以及节点中心度等指标均有显著的变化。文献耦合网络中高中间中心度的节点略有减少,显示内容加权耦合网络具备更好的连通性。

综上,本研究有两点启示:

a.引文内容能丰富耦合网络中的节点关系。内容加权后的文献耦合网络比传统文献耦合网络有更丰富的节点关系。通过内容加权,网络中节点有更高的度分布和权重分布,从而改变耦合网络结构。

b.结合内容特征构建引文网络具有良好的应用前景。日益丰富的内容数据为构建内容加权的引文网络提供更多支持。在不增加复杂性的基础上,内容特征的应用能获得更好的计量研究结果,提高研究成果的各方效益转化[2-3]。

本研究也存在一定的局限性。本研究的主要数据来源于 PLoS 期刊上生物学学科论文。尽管 PLoS 期刊上生物学论文的学术影响力较高,研究结果具有一定代表性。但本研究尚未对其他学科作进一步分析,以进一步提高本研究结论的普适性。未来研究可在以下方面进行深入研究和探索:

a.扩大学科范围和语料集。尽管本文选取的生物

医学领域在PLoS中占有重大比例,但由于PLoS并未包含更多的人文经管等学科,该数据集在更广泛的代表性仍存在一定欠缺。未来可扩大语料,如使用PubMed等数据集开展更广泛性的数据融合,扩充更多研究数据,得出更全面的实验结果,进一步论证相关研究的普适性。

b.探索更多的引文内容特征融合方案。本研究所构建的内容加权网络并未使用位置相关的引文内容特征和全文内容特征[1]。在未来的工作中,可进一步扩大特征选择范围,探索其他特征在引文网络构建中的应用,为新兴研究话题发现以及其他重要的文献计量领域[20]提供方法工具。

参 考 文 献

[1] 卢 超,章成志,王玉琢,等.语义特征分析的深化——学术文献的全文计量分析研究综述[J].中国图书馆学报,2021,47(2):110-131.

[2] Lu C,Ding Y,Zhang C.Understanding the impact change of a highly cited article:A content-based citation analysis[J].Scientometrics,2017,112(2):927-945.

[3] Jeong Y K,Song M,Ding Y.Content-based author co-citation analysis[J].Journal of Informetrics,2014,8(1):197-211.

[4] Kim H J,Jeong Y K,Song M.Content- and proximity-based author co-citation analysis using citation sentences[J].Journal of Informetrics,2016,10(4):954-966.

[5] 李秀霞,马秀峰,程结晶. 融入引文内容的期刊耦合分析[J].图书情报工作,2016,60(11):100-106.

[6] 卢 超,章成志.基于引文内容的单篇学术论文参考文献网络结构研究[J].现代图书情报技术,2014(10):33-41.

[7] Ding Y,Zhang G,Chambers T,et al.Content-based citation analysis:The next generation of citation analysis[J].Journal of the Association for Information Science and Technology,2014,65(9):1820-1833.

[8] Zhao D,Strotmann A.Dimensions and uncertainties of author citation rankings:Lessons learned from frequency-weighted in-text citation counting[J].Journal of the Association for Information Science and Technology,2016,67(3):671-682.

[9] Huang S,Qian J,Huang Y,et al.Disclosing the relationship between citation structure and future impact of a publication[J].Journal of the Association for Information Science and Technology,2022,73(7):1025-1042.

[10] Nicolaisen J.Citation analysis[J].Annual Review of Information Science and Technology,2007,41(1):609-641.

[11] Hu Z,Chen C,Liu Z.Where are citations located in the body of scientific articles? A study of the distributions of citation locations[J].Journal of Informetrics,2013,7(4):887-896.

[12] 徐庶睿,卢 超,章成志.术语引用视角下的学科交叉测度——以PLOS ONE上六个学科为例[J].情报学报,2017,36(8):809-820.

[13] Bertin M,Atanassova I.A study of lexical distribution in citation contexts through the IMRaD standard[J].PloS Negl.Trop.Dis,2014,1(200,920):83-402.

[14] Ding Y,Liu X,Guo C,et al.The distribution of references across texts:Some implications for citation analysis[J].Journal of Informetrics,2013,7(3):583-592.

[15] Teufel S.Argumentative Zoning:Information extraction from scientific text[D].Edinburgh:University of Edinburgh,1999.

[16] Lu C,Bu Y,Dong X,et al.Analyzing linguistic complexity and scientific impact[J].Journal of Informetrics,2019,13(3):817-829.

[17] 佚 名.NSF classification of fields of study[Z/OL](2013-05-22)[2022-07-29].https://www.nsf.gov/statistics/nsf13327/pdf/tabb1.pdf.

[18] Zhang L,Glanzel W,Ye F Y.The dynamic evolution of core documents:An experimental study based on h-related literature(2005-2013)[J].Scientometrics,2016,106(1):369-381.

[19] 卢 超,侯海燕,Ding Y,et al.国外新兴研究话题发现研究综述[J].情报学报,2019,38(1):97-110.

[20] Arino K,Furukawa T,Shirakawa N,et al.Temporal network analysis of emerging technologies:Topic transition in World Wide Web(WWW) conferences[C/OL]//2012 IEEE International Conference on Industrial Engineering and Engineering Management.Hong Kong,China:IEEE,2012:1108-1112[2019-09-20].http://ieeexplore.ieee.org/document/6837914/.

[21] Barabási A-L.Network Science[M].Cambridge,UK:Cambridge university press,2016.

[22] Onnela J-P,Saramäki J,Hyvönen J,et al.Structure and tie strengths in mobile communication networks[J].Proceedings of the National Academy of Sciences,2007,104(18):7332-7336.

[23] Mund C,Neuhaeusler P.Towards an early-stage identification of emerging topics in science-the usability of bibliometric characteristics[J].Journal of Informetrics,2015,9(4):1018-1033.

[24] Jarneving B.Bibliographic coupling and its application to research-front and other core documents[J].Journal of Informetrics,2007,1(4):287-307.

[25] Albert R,Barabási A-L.Statistical mechanics of complex networks[J].Reviews of Modern Physics,2002,74(1):47.

[26] Brandes U,Pich C.Centrality estimation in large networks[J].International Journal of Bifurcation and Chaos,2007,17(7):2303-2318.

(责编/校对:王平军)