ISSN 1003-3513

10
2014
总第251期

现代图书情报技术

NEW TECHNOLOGY OF LIBRARY AND INFORMATION SERVICE

中国科学院主管
中国科学院文献情报中心主办

Xiandai Tushu Qingbao Jishu

现代图书情报技术 (月刊)

1985 年创刊
2014 年 第 10 期 总第 251 期
2014 年 10 月 25 日出版

主管：中国科学院
主办：中国科学院文献情报中心
主编：张晓林
执行主编：彭希珺
编辑部主任：李春源
编辑：华 宁
广告审核：苗志刚
编辑 / 出版：《现代图书情报技术》编辑部
地址：北京中关村北四环西路 33 号 (100190)
电话：(010) 82626611-6626
(010) 82624938
传真：(010) 82624938

E-mail：jishu@mail.las.ac.cn
网址：http://www.infotech.ac.cn
印刷单位：北京科信印刷有限公司
发行范围：国内外公开发行
国内发行：北京市报刊发行局
国外发行：中国国际图书贸易总公司
订购处：全国各地邮局
ISSN 1003—3513
CN 11—2856/G2
国内邮发代号：82—421
国外代号：M4345
定价：80 元

广告经营许可证：京海工商广字第 0032 号

目 次

中国图书馆学优秀期刊
中文核心期刊

2014 年　第 10 期　总第 251 期

【数字图书馆】

【知识组织与知识管理】

【情报分析与研究】

【动态】

中国图书馆学优秀期刊
中文核心期刊

MAIN CONTENTS

No.10 Oct. 2014

DIGITAL LIBRARY

KNOWLEDGE ORGANIZATION AND KNOWLEDGE MANAGEMENT

INFORMATION ANALYSIS AND RESEARCH

Sponsored by: Library of Chinese Academy of Sciences
Edited by: Editorial Committee of New Technology of Library and Information Service,
No.33 Beisihuan Xilu, Zhongguancun. Beijing 100190, China
Http: //www.infotech.ac.cn **E-mail:** jishu@mail.las.ac.cn
Distributed by: China International Book Trading Corporation (Guoji Shudian)
M4345 P.O.Box 399, Beijing, China

基于引文内容的单篇学术论文参考文献网络结构研究

卢　超　章成志
(南京理工大学经济管理学院　南京 210094)

摘要:【**目的**】通过对参考文献在学术论文正文中的引用及分布情况的分析，探究参考文献的网络结构形态。【**方法**】基于 575 篇结构化的学术论文数据，利用文本抽取、相似度计算等技术，构建每篇学术论文的参考文献的网络结构，结合实例分析参考文献之间的内在联系及其可能的原因。【**结果**】参考文献间的相似度与其之间的相对距离有一定的负相关性。单篇学术论文中亦存在多样、复杂的网络结构形态。【**局限**】部分全文数据引文标注不够规范，影响实验结果的准确性；参考文献之间相对位置的衡量仍不够精确，需要深入挖掘文本加以解决。【**结论**】从实验结果来看，参考文献的网络结构大致可分为三类，其形成的原因各有不同。单篇论文中参考文献网络仍需深入研究。

关键词: 引文分析　引文内容　网络分析　文本挖掘
分类号: TP393

1　引　言

自从 Garfield 创立科学引文索引(Science Citation Index，SCI)以来，引文分析日渐成熟，最终成为文献计量工作中重要的分析方法。引文分析研究成为文献计量学等相关学科的重要研究领域。影响因子、H 指数等指标是评价期刊、文献质量、科研工作者的学术影响力的重要工具。随着计算机技术的发展，绘制大规模的引证关系图谱成为可能。通过对引证网络研究，能够从计量的角度研究学科之间的相互关系，把握学术动态和学科的前沿发展[1-2]。

然而，有限的文献资源数据只能将引文分析限定在依靠纯粹的引证关系(引与不引、引用与被引、自引与他引等关系)中进行研究。文献中大量的引文细节信息被忽略了，如文献的上下文、具体的引文内容以及引文的极性。因此，有的学者将内容分析添加到引文分析中，从事了相关研究[3-5]。随着自然语言处理技术的不断成熟，全文数据库、结构化全文数据的获取越来越容易，对全文内容甚至细致到引文内容的深度挖掘成为可能，利用自然语言处理(Natural Language Processing，NLP)技术进行引文分析的研究也越来越多[6-9]。借助于 NLP 技术，引文分析能够关注到引证数据在文本中的独特内涵，对于引用动机的甄别具有独特的意义。

在众多的相关研究中更多地关注期刊之间、学术论文之间、作者之间的引文分析[8-10]，目前尚无针对单篇学术论文内部引文网络的研究。通过对单篇论文内部引文内容的挖掘可以探究参考文献在文章中的内在联系，了解作者对其引用文献的理解和把握，为后期进一步了解作者的引文动机提供借鉴。

基于以上考虑，本文采集 600 篇图情领域的学术论文的全文数据，并加以编号。其中有 25 篇正文中没有标记参考文献，不作考虑，但沿用原始文献编号作为标记。利用余下的 575 篇结构化的学术论文全文数据进行单篇论文的引文网络的绘制，探究参考文献在学术论文中的相互关系。

收稿日期: 2014-04-09
收修改稿日期: 2014-05-25

2 相关工作

2.1 引文分析

引文分析法利用文献之间的引证关系研究文献数量特征和内在规律[1]。在此基础上，Garfield[10]又在1972 年提出了影响因子，用来评价期刊的重要性。随后提出的总引证次数、影响因子、即年指标、被引半衰期等相关指标被广泛用来评价期刊、学术论文、机构、作者等的影响力以及他们的聚类情况。Kessler[11]在 Garfield 的基础上提出引文的“耦合分析”用来描述和解释引文之间的多重引证关系。Small[12]、Marshakova[13]又进一步提出“同被引”、“圆环模型”补充引文分析。Liu 等[14]研究发现共引文献的本质在于科研兴趣的互动交流。

但引用行为相当复杂，单纯依靠被引频次无法反映文献中的真实状况[15]。如当论文中出现反面引用时，在计量分析中只能当作正面引用，使分析结果不够精确，影响因子等指标并不能完全体现期刊文献的质量[16-17]；针对不同类型的引用，引证分析都均等看待，存在不合理之处；盲目地依赖引文分析结果进行学术评价，还易造成强迫引用、假引等情况[6]的出现。此外，引用分析工具缺乏共同的标准，分析结果差异较大[2]。因此，基于引用分析的期刊评价的完备性值得质疑[17]。引文分析的缺陷都是因为所有的分析都仅仅考虑文献或期刊的被引频次,而忽视了文献的具体内容。

2.2 引文内容挖掘

全文数据库和自然语言处理技术的成熟为引文分析与内容分析相结合提供了条件，也成为引文分析新的方向[18]。由此，大量的学者开始关注引文内容的挖掘，丰富引文分析理论。刘盛博等[19]结合引文内容，构建引文的检索和推荐系统，发现具有良好的检索和推荐效果；此外，还有 Bradshaw[20] 和 Ritchie 等[21]的研究。引文内容分析还被应用在主题发现和识别中，具有较好的主题识别和分类效果[7-8]。Jeong 等[9]利用 JASIST 提供的结构化数据，进行作者共引分析，并与传统的共引分析进行对比，发现基于内容的作者共引分析能够提供更多的有效信息。Boyack 等[22]利用全文信息进行共引的聚类分析，发现全文下的共引聚类比利用基本引文数据聚合度提高了 30%。依据引文所在的位置对引用进行统计分析，如 Hu 等[23]、Liu 等[24]的相关研究。

综上所述，这些利用引文内容的研究涉及论文的检索与推荐、主题聚类、作者共引分析以及引文位置的研究等，这些研究都是引文内容挖掘与引文分析相结合的成果。但这些研究都仅限于文章、期刊之间的引证关系研究，目前尚未发现基于单篇学术论文中参考文献的内在关系研究。成熟的 NLP 技术，使得仅考虑单篇学术论文的参考文献及其引文的引文分析研究成为可能，这将有利于读者发现参考文献之间的内在关联以及作者的引用意图。

3 总体思路与具体方法

3.1 总体思路

为了实现本文的研究目标——探究学术论文内部参考文献的网络结构，分 6 个步骤展开研究，如图 1 所示:

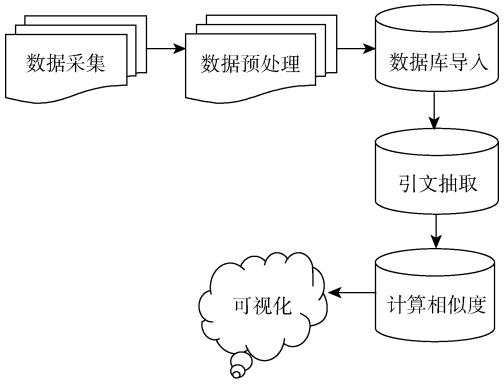

图 1 实验技术路线

(1) 数据采集，获取期刊论文的全文数据，并对论文进行格式化标注;

(2) 数据预处理，对得到的结构化数据进行数据处理和清洗工作，为后期的数据抽取和计算做好准备;

(3) 数据库导入，将清洗后的数据以数据库的形式存储;

(4) 引文抽取，将全文数据中的引文数据抽取出来存入数据库;

(5) 计算相似度，计算每篇文章内引文之间的相似度，作为参考文献之间的相似度以及后期可视化边的权重;

(6) 可视化，根据数据库中每篇文章参考文献之间的相似度绘制参考文献网络结构图，以供分析和讨论。

3.2　关键技术说明

本文利用到的关键技术方法有：引文内容抽取、相似度计算和参考文献关系可视化。

(1) 引文内容抽取

抽取文本中包含引文标记的句子作为引文内容[19](Sentence)文献的编号(Id)、位置信息(Location)以及参考文献所属编号(ReferenceId)。由于不同的期刊对引文的标引有不同的格式要求，因此，本文归纳出 4 种基本的标引模式，以便抽取：

①[A]简单式；

②[A-B](或[A~B])省略式；

③[A, B, C](或[A、B、C])列举式；

④错误类型，比如缺少完整的标记符号，如仅有“[+数字”或者“数字+]”。

针对模式①，直接抽取引文及相关信息；针对模式②，填充抽取，如[1-3]，抽取(1、2、3)三篇参考文献的相关信息和引文，此时引文的内容是一致的；针对模式③，直接抽取列举出的参考文献的相关信息及引文；针对模式④，经实际确认该类错误的标引在文献数据库中已存在(即该论文在发表前就可能已经产生了错误)。针对此种情况，本文选择忽略错误标引的参考文献，不抽取相关信息和引文内容。

在论文中，有较多的交叉引用情况。本文对交叉引用的文献做如下处理：将出现引文文献所有位置的引文都抽取出来并进行合并；其他相关的信息也进行合并，并加以标识。例如某参考文献[6]既出现在文章 A 中的引言部分，也出现在结论部分，则参考文献[6]的引文内容是引言部分和结论部分引文内容合并后的结果。

(2) 引文内容相似度计算

在计算每篇文章中各条引文内容之间的相似度之前，需要将引文表示成空间向量。具体做法如下：利用 ICTCLAS2011①对引文内容进行分词、词性标注；筛选引文内容中的名词、动词、形容词三种词性(具体参见 ICTCLAS2011 词性标注集②)；根据每篇论文的引文内容建立词典，并将每句筛选过的引文内容表示成向量形式。这样得到的向量就可以进行相似度计算。衡量相似度的指标有很多，本文使用 Cosine 值度量相似度[25]，计算方法如公式(1)所示：

$$\mathrm{Similarity}(\vec{a},\vec{b})=\frac{\sum_{i=1}a_i b_i}{\sqrt{\sum_{i=1}a_i^2\sum_{i=1}b_i^2}} \quad (1)$$

其中 $\vec{a},\vec{b}$ 分别表示两组引文内容的向量。在本文中，向量是由引文内容经过向量表示得到。如某篇论文中仅含有两句引文为：“Lewicki 等人[15]将信任定义为依赖于交流伙伴的意愿”、“Fishbein 和 Ajzen[14]指出，信任信念先反映出用户的态度然后能通过态度影响用户行为”。经过分词、词性标注和筛选后变成：{Lewicki; 信任; 定义; 依赖; 交流; 伙伴; 意愿; }、{Fishbein; Ajzen; 指出; 信任; 信念; 反映; 用户; 态度; 影响; 行为; }。这两句引文构成了 16 维词袋为：{Lewicki; 信任; 定义; 依赖; 交流; 伙伴; 意愿; Fishbein; Ajzen; 指出; 信念; 反映; 用户; 态度; 影响; 行为; }，则第一句对应这个词袋生成的 16 维向量为：(1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0)；第二句的向量为：(0,1,0,0,0,0,0,1,1,1,1,1,1,1,1,1)。这两句引文的相似度为：$\frac{1}{\sqrt{7\times10}}=0.1195$。因此，这两句话的相似度为 0.1195。

(3) 数据标准化

由于本文选取的文献之间本身存在一定的差异性，文献类型多样、参考文献数目不同，需要对获得的相似度进行标准化，以获得各文献之间相似度水平的可比性，便于接下来的统计、分析和展示。本文采取极差标准化方法对原始相似度数据进行标准化，计算方法如公式(2)所示：

$$\mathrm{Similarity}_{\mathrm{iStd}}=\frac{\mathrm{Similarity}_i-\min}{\max-\min} \quad (2)$$

其中，$\mathrm{Similarity}_i$ 表示任意两节点之间的相似度，max 表示所有节点间相似度的最大值，min 表示所有节点间相似度的最小值，$\mathrm{Similarity}_{\mathrm{iStd}}$ 表示两点间相似度标准化后的结果。

(4) 结构关系可视化

将每篇文章的参考文献当作节点，得到的相似度设为权重，利用 JS 调用开源的 D3③工具包绘制参考文献的网络结构图。

①http://ictclas.org/index.html.

②http://fhqllt.iteye.com/blog/947917.

③http://d3js.org/.

4 数据与实验结果分析

4.1 学术论文全文数据

为了获取全文数据，并在一定程度上保证论文的质量和引用格式的规范性，笔者从网络上获取 534 篇情报学领域的学术论文[①]以及 66 篇图书馆学领域的论文[②]，共计 600 篇论文的全文数据。但其中有 25 篇论文正文中没有做参考文献引用的标记，故将其忽略。在剩余的 575 篇论文中，CSSCI 期刊论文占 90%以上。这 575 篇文章平均有 5.1 章，7 358.6 个字，平均参考文献数为 18.41 篇，具体分布如图 2 所示，各引用类型的基本统计如表 1 所示，每篇论文引文平均长度分布如图 3 所示。

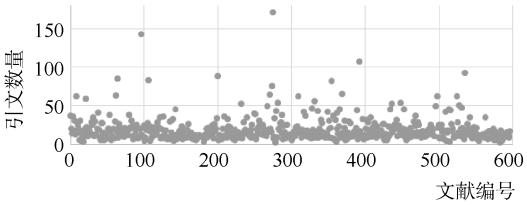

图 2 引文数量统计

表 1 篇均引用类型基本统计

引用类型	简单引用	省略引用	枚举引用	合计
平均频数(每篇)	17.93	0.62	0.13	18.68*

(注：表 1 中合计的平均引用频次比 18.41 大，是由于原文交叉引用参考文献导致。)

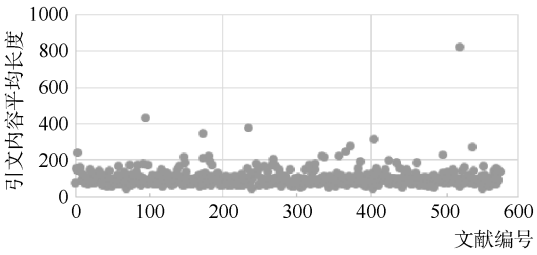

图 3 引文长度统计

论文之间以<Document>…</Document>标签标记，全文内容用<FullText>…</FullText>标签标记，在全文中将篇章段落、标题、内容等以<Introduction∗>、<Chapter∗>等开头标记(∗表示具体在文章中的结构编号)。采集完毕后通过人工校对，添加以<Content∗>和<Heading∗>开头的标签获得信息更加精确的 XML 格式的结构化数据。采集时间：2013 年 10 月 28 日 00:00-2013 年 11 月 28 日 24:00。全文数据标注实例如下所示：

```
<Document><Title>个人信息管理工具使用意愿研究——以智能手机为例</Title>
  <Author>……</Author>
  <Source>……</Source>
  <AuthorInformation>……</AuthorInformation>
  <Abstract>……</Abstract>
<Keywords>智能手机/个人信息管理/技术采纳模型/结构方程模型</Keywords>
  <FullText>
  <Introduction1>……</Introduction1>
  ……
  <Chapter2>
      <Heading2>2  研究模型与假设 </Heading2>
      <Content2>……</Content2>
      <Heading2.1>2.1  趣味性 </Heading2.1>
      <Content2.1>……</Content2.1>
      ……
  </Chapter2>
  ……
  </FullText>
  <References>……</References>
  <PageUrl>……</PageUrl>
  </Document>
```

将得到的 XML 格式的论文解析到数据库中，在全文的<Introduction∗>和<Content∗>两个字段的文本中获取引文信息。经过记录去重，最后得到来自 575 篇文献 11 068 条引文内容。按照公式(1)计算每篇论文内部各篇参考文献之间的相似度，并利用公式(2)进行数据的标准化。

将标准化后的数据进行区间统计，初始区间是[0,0.01],最后一个区间是[0.99,1.00]，共计 100 个等距区间。由 575 篇文献计算得到的 142 732 个相似度的权重标准化(后文将“标准化的相似度”简称为“相似度”或“权重”)，并统计区间绘制分布图，如图 4 所示。图 4 纵坐标是各区间频数，横坐标的序号依次代表各个区间。趋势线是拟合散点图的分布产生的，并给出了趋势线的公式以及相关性系数。可以看出，从第 10 个区间到第 90 个区间基本呈一条直线；从第 90 个区间到第 100 个区间，散点的分布呈上扬趋势。这部分

①http://old2013.cssn.cn/67/6702/.

②http://old2013.cssn.cn/67/6700/.

数据和趋势线拟合情况较差，这也导致了相关性系数比较低。

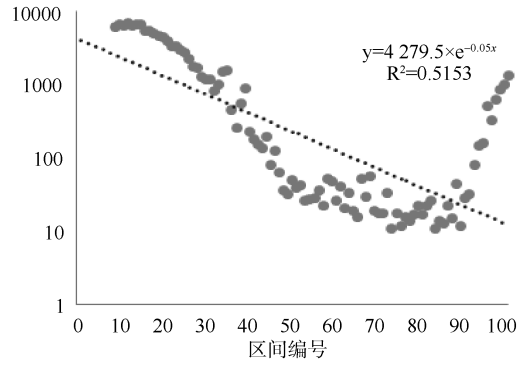

图 4　相似度区间分布

大体来看，参考文献之间相似度的分布总体上呈幂律分布，即 575 篇文献内部参考文献之间的相似度能够从一定程度上反映单篇学术论文内部参考文献网络的结构状况。

本文将得到的各篇文章中引文之间的所有相似度进行汇总，依据相似度及其对应的文献编号生成散点图，如图 5 所示:

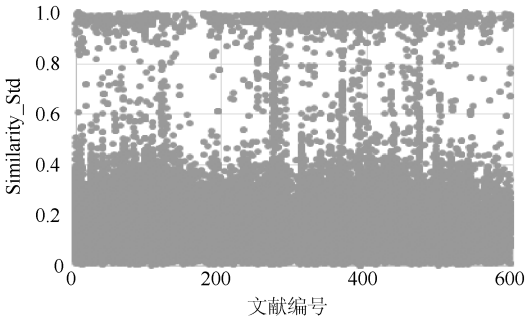

图 5　相似度及其文章编号

从图 5 可以看出，这些参考文献之间的相似度主要集中在两个区域: [0,0.4]以及[0.8,1.0] 这两个区间。且落在[0,0.4]这个区间的点占到 95%，在[0, 0.2]区间中的占总体的 74%。这些点之间的相似度非常小，且占据大部分数据。说明作者引用的这些文献在主题上有一定联系，但这种关联性很弱。这也表明，作者引用参考文献有不同的用处。相对区间[0,0.4]，落在[0.8,1]之间的点较少，这些点出现的可能原因是作者在引用时采用了②、③两种引用模式，导致相似度水平较高。纵观全文的分布来看，落在[0.4,0.8]之间的点较少。这也验证了参考文献之间要么联系得很紧密，要么联系得很松散，而且各自有各自的区域，界限较为清晰。

因此，试图构建单篇学术论文的内部引文网络可行并且有意义。下文将具体探究单篇论文中参考文献之间的差异和联系。

4.2　引文分布及其相似度

根据以上所阐述的问题，文献之间的相似度在[0.4,0.8]之间的分布比较少。因此仍然延续上文图像的绘制方法，绘制引文内容的相似度与引文距离之间的散点分布。本文将这里的引文距离(参考文献之间的距离)表示为参考文献编号之间差的绝对值，最终得到的样本散点分布如图 6 所示:

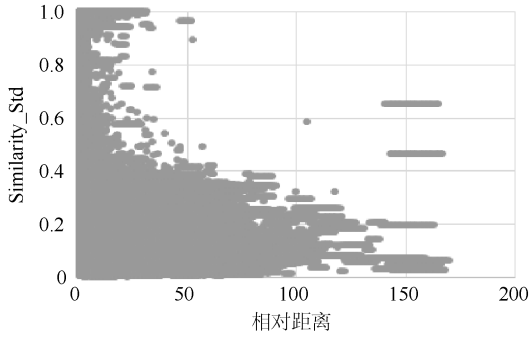

图 6　相似度与参考文献相对距离

图 6 依旧验证了相似度的区间分布。更加值得注意的是，引文之间的距离越远，引文之间的相似度的平均水平越低，反之亦然。同时，随着引文距离的增加，点的密度也在下降，相似度水平也在降低。造成这种现象的原因可能是: 不是所有的学术论文都会有 100 多篇参考文献，所以，一方面引文距离增大，点越稀疏; 另一方面，随着参考文献之间的距离增大，论文探讨的话题和模块发生变化，引用文献的目的和用途也在发生变化，从而使得引文之间的相似度逐渐减低。

图 7 绘制了各个参考文献相对距离下的平均相似度的散点趋势线。总体来看，参考文献之间的相对距离越近，相似度的水平越高。当参考文献的相对距离大于 7 后，文献之间的平均相似度开始趋于稳定，在 0.15 附近波动。当相对文献距离继续增大时，参考文献之间的相似度的波动范围也越来越大。造成这种现象的主要原因有:

(1) 参考文献相对距离大于 7 以后，作者更趋于阐述另外一个话题;

(2) 由于少量交叉引用的存在促使相似度维持在一定水平;

(3) 综述性的文章一般含有较多的参考文献和丰富的交叉引用，会引起不同相似度水平的巨大提升，由于本研究选取文献中综述性文章数量较少，给相似度水平带来较大的波动。

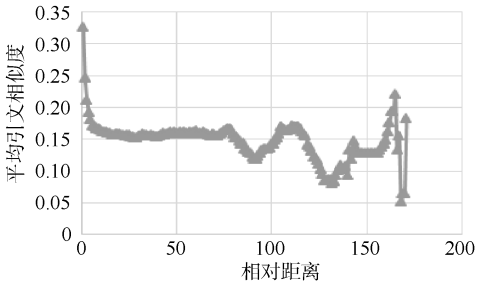

图 7　平均引文相似度及参考文献距离趋势

总的来看，引文之间的相似度与引文之间的相对距离具有一定负相关，相对距离越大，相似度水平越低。参考文献之间的相似度大小存在明显的区域分布特性，而交叉引用的存在似乎打破了这种“区域性”。因此，接下来利用 D3 可视化学术论文的参考文献网络，探讨参考文献的关联。

4.3　参考文献网络结构分析

参考文献之间的相似度非常低，可以认为这两篇参考文献没有“相关性”，需要对其进行过滤。经验证发现相似度小于 0.2 ，引文之间的相关性很小，因此本文设定的阈值为 0.2，即对相似度低于 0.2 的进行过滤。另外，(0.9,1)区间的相似度是由于两篇以上参考文献共同出现在同一句引文内容中产生的。根据图 2，这一部分的相似度虽不符合幂律分布，但参考文献共同出现在同一个句子中说明参考文献之间高度相关，因此需要将其纳入到网络结构的构建中。本文绘制的学术论文的网络结构的边权即相似度，属于[0,1]区间，最终绘制了 575 篇论文参考文献之间的网络结构。其中主要的结构形式有以下 4 种:

(1) 文献中含有大量的参考文献，这些参考文献之间联系紧密;

(2) 参考文献相对结构形式(1)较少，形成明显社区;

(3) 相关联的参考文献进一步减少，社区明显分离，形成“孤岛”;

(4) 参考文献大量减少、参考文献间相似度低，导致结构图中出现大量的孤立点和少量的链状结构，无法形成明显的网络结构。

各个类型的实例如图 8 所示:

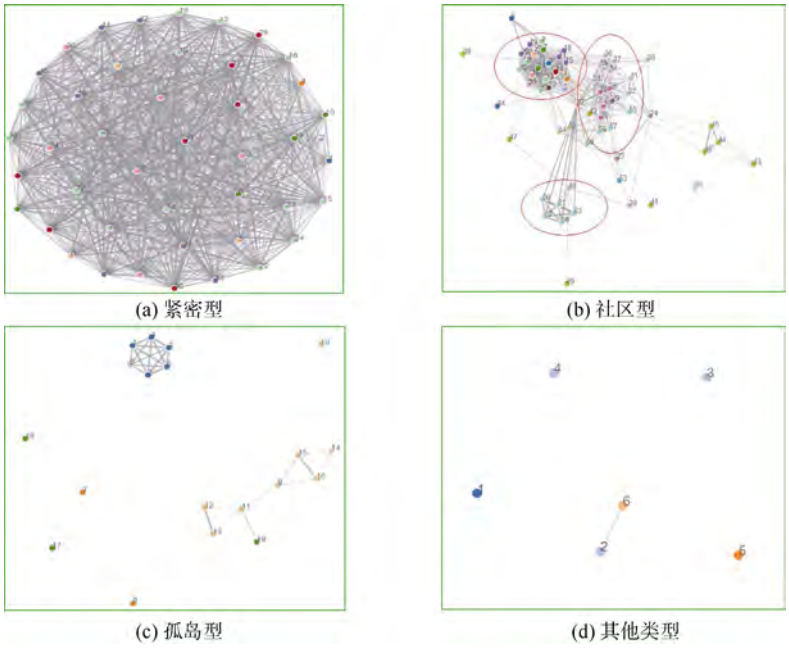

图 8　参考文献网络结构类型示例

这几种结构类型比例分布情况如表2所示，80%以上的学术论文内部的参考文献网络结构有明显的社区划分，且各个部分之间相互联系。但仍有部分论文的参考文献网络联系关系很弱，无法形成社区结构。

表 2　网络结构类型及其文献类型统计

结构类型 / 项目	紧密型	社区型	孤岛型	其他	合计
篇数	5	242	258	70	575
比率	0.01	0.42	0.45	0.12	1.00
文献类型	综述	综述、研究	研究	研究	—
平均参考文献数	85.00	22.30	15.80	9.81	18.41

(1) 紧密型

如图 8(a)所示，在这种论文中有大量的参考文献，且参考文献之间引用的相似度非常高，并没有出现明显的社区分化。在所有绘制的图谱中这样的结构一共有 5 篇论文，形成紧密的“大陆”。返回到数据库中查找原文，这 5 篇论文全部是综述性文章:《国外可用性研究进展述评》、《基于<中图法>的自动分类研究现状与展望》、《国外竞争情报研究进展：概念辨析、问题论域及发展趋势》、《国内知识图谱研究综述与评估: 2004-2010 年》、《1979-2010 年中国图书馆学理论体系研究进展述评》。这些文章都引用大量的参考文献，且引用的类型以省略式为主，这些文献中存在较多的交叉引用。基于本文计算权重的方式，交叉引用、省略式引用以及举例式引用会明显提高相似度水平，即边的权重。从实际的论文写作来看，综述性文章需要对一系列具有相似观点、研究方法和研究结论的优秀成果加以概括和总结，但这种过于概括性的总结也会淹没掉被引文章中大量的信息[22]。在这种复杂结构的图谱中，无法明显看出其内部的网络结构。

(2) 社区型

社区型的网络结构没有紧密型网络结构相似度水平高，但依旧存在明显的社区结构，如图 8(b)所示。绘制的大部分结构图都存在明显的社区，样图来自《知识网络的结构与演化——概念与理论进展》一文，该文是一篇综述，共有 64 篇参考文献。不同于之前综述构成的紧密型结构，该综述形成明显的社区结构。该文章对知识网络结构的理论发展和结构演化做了细致梳理，并作相关述评。其中，理论发展和网络结构类型引用大量的文献，分别为: 29 篇(去重后 19 篇)、34 篇(去重后 30 篇)。绘制的结构图形成明显的三个社区，多出来的一个社区来自原文的“网络的演化”这一主题。在这些社区之间分布少量参考文献，分别为 2 篇、23 篇、24 篇、27 篇。这些参考文献在文中被多处引用，将各个社区连接在一起，使得各个主题有了共通之处。反映在量化计算中就是被多次引用的参考文献的引文内容是多处引文的合并。另外，在其他含有交叉引用的文献中，特别当相对距离较大时，社区结构就会出现明显的连接。当文献出现了社区结构，而参考文献较少或缺乏这种交叉引用时，网络结构就会出现“孤岛型”形态。

(3) 孤岛型

孤岛型网络结构中参考文献在一定程度上形成各自的社区，但在这些社区之间缺乏相似度较高的引文内容。导致这样的结果有几个可能的原因:

①文献本身的参考文献数量有限，无法构建复杂的网络结构;

②这些参考文献主题差异化很大，在引用时就有明显差异化的社区;

③由于构图前的相似度过滤，很多微小的联系被过滤，无法构建这些社区的联系;

④依据之前的一些分析结论，该种结构的论文中缺乏交叉引用。

样图来自于《关于学位论文网络传播的思考》一文，该文章主要阐述了两方面的内容:“校园网内传播本校学位论文的合理性和可行性”和“学位论文网络传播对作者权益的保护作用”。从主题上来看，这两个主题之间的差异较大，此外，该文仅引用 19 篇参考文献，所引用的文献数量相对于“紧密型”和“社区型”结构论文的所引用的文献数量较少，因此没有相对距离较大的交叉引用。这些原因导致了该文章形成图 8(c)中“孤岛型”的示例图。样图中形成两个明显结构紧密的社区，周围伴有一些孤立的参考文献。

(4) 其他类型

该种结构分布的更加简单，这些参考文献之间没有形成明显的社区结构，而仅仅含有一个链状或环状圆状的结构，周围散布很多孤立的参考文献。存在的节点连线主要是相对距离为 1 或 2 的参考文献。形成这种社区的主要原因有:

①参考文献的数量很少，参考文献之间的相似水平过低;

②与紧密型参考文献引用状况不同的是，这类文章的参考文献的描述更加详细，没有过多的省略式引用;

③论文的主题比较分散，作者引用的文章来自不同的领域，以致这种联系十分脆弱,如图 8(d)所示。

样图来自《Weblog 生命周期模式研究》一文，虽

在参考文献部分列出 12 篇文献，但在全文中只标引 6 篇。标引的文献数量过少，而且这些文献分布极为分散，因此无法形成社区结构。

综上，不仅学术论文之间可以形成复杂的网络结构，单篇学术论文引用的参考文献之间也具有一定的网络结构。这种结构非常普遍地存在于学术论文中，结构形态可能还非常复杂。参考文献的数量、学术论文主题划分、参考文献引用的类型及形式都会影响单篇学术论文参考文献网络结构的复杂程度和结构形态。

5 结 语

全文数据的可获取和 NLP 技术的不断成熟，帮助传统的引文分析理论找到了新的发展方向。然而这些相关的引文分析研究更多关注论文之间、期刊之间、作者之间的引证分析，却没有对单篇的学术论文的引文进行深入研究。而单篇学术论文中参考文献之间的相互关系在 NLP 技术的帮助下能够得到更好的解读。因此，本文获取了 575 篇论文的全文数据并进行引文相似度的计算和引文之间网络结构的可视化，在进行学术论文研究时，发现了参考文献之间的联系，相似度水平与相对距离存在一定关系，学术论文内部的参考文献之间也有多样化的网络结构。本文依据网络的形态和复杂程度初步划分成 4 种形态，但由于获取的期刊论文的全文数据存在准确性和规范性问题，会导致后期数据抽取和分析时出现偏差。另外，为了简化操作，将文献编号之差的绝对值作为引文的相对距离，此种方法有待改进。

在进一步研究中还将深入研究参考文献之间的网络结构，试图探究作者论文中话题迁移的引用意图，不同的引用类型对学术论文内部网络的影响以及如何将参考文献网络应用到学术论文质量的评价中。

参考文献：

[1] 邱均平. 信息计量学[M]. 武汉：武汉大学出版社，2007. (Qiu Junping. Information Metrology [M]. Wuhan: Wuhan University Press, 2007.)

[2] 杨思洛. 引文分析存在的问题及其原因探究[J]. 中国图书馆学报, 2011, 37(3): 108-117. (Yang Siluo. The Problems of Citation Analysis and Their Causes [J]. Journal of Library Science in China, 2011, 37(3): 108-117.)

[3] Wakefield R. Networks of Accounting Research: A Citation-Based Structural and Network Analysis [J]. The British Accounting Review, 2008, 40(3): 228-244.

[4] 柯平，贾东琴. 2001-2010 年境外信息管理研究进展——基于相关文献的计量分析和内容分析[J]. 中国图书馆学报，2011, 37(5): 61-74. (Ke Ping, Jia Dongqin. Research Progress on Information Management from 2001 to 2010 at Abroad: Based on the Bibliometric Analysis and Content Analysis [J]. Journal of Library Science in China,2011, 37(5): 61-74.)

[5] Halevi G, Moed H F. The Thematic and Conceptual Flow of Disciplinary Research: A Citation Context Analysis of the Journal of Informetrics, 2007 [J]. Journal of the American Society for Information Science and Technology, 2013, 64(9): 1903-1913.

[6] Yu T, Yu G, Wang M. Classification Method for Detecting Coercive Self-Citation in Journals [J]. Journal of Informetrics, 2014, 8(1): 123-135.

[7] 祝清松，冷伏海. 基于引文内容分析的高被引论文主题识别研究[J]. 中国图书馆学报，2014, 40(1): 39-49. (Zhu Qingsong, Leng Fuhai. Topic Identification of Highly Cited Papers Based on Citation Content Analysis [J]. Journal of Library Science in China, 2014, 40(1): 39-49.)

[8] Liu X, Zhang J, Guo C. Full - Text Citation Analysis: A New Method to Enhance Scholarly Networks [J]. Journal of the American Society for Information Science and Technology, 2013, 64(9): 1852-1863.

[9] Jeong Y K, Song M, Ding Y. Content-Based Author Co-citation Analysis [J]. Journal of Informetrics, 2014, 8(1): 197-211.

[10] Garfield E. Citation Analysis as a Tool in Journal Evaluation [J]. Science, 1972, 178(4060): 471-479.

[11] Kessler M M. Bibliographic Coupling Between Scientific Papers [J]. American Documentation, 1963, 14(1): 10-25.

[12] Small H. Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents [J]. Journal of the American Society for Information Science, 1973, 24(4): 265-269.

[13] Marshakova I V. System of Document Connections Based on References [J]. Scientific and Technical Information Serial of Viniti, 1973, 6(2): 3-8.

[14] Liu Y, Rousseau R. Interestingness and the Essence of Citation [J]. Journal of Documentation, 2013, 69(4): 580-589.

[15] Zhang G, Ding Y, Milojević S. Citation Content Analysis (CCA): A Framework for Syntactic and Semantic Analysis of Citation Content [J]. Journal of the American Society for Information Science and Technology, 2013, 64(7): 1490-1503.

[16] Waltman L R, Costas R. F1000 Recommendations as a New Data Source for Research Evaluation: A Comparison with Citations[EB/OL].(2013-03-18).[2014-04-09]. http://arxiv.org/

ftp/arxiv/papers/1303/1303.3875.pdf.

[17] 叶继元. 首届人文社会科学评价学术研讨会综述[J]. 学术界, 2009(4): 301-304. (Ye Jiyuan. Review of the First Conference of Humanities and Social Science Evaluation Academics in China [J]. Academics in China, 2009(4): 301-304.)

[18] Content-based Citation Analysis: The Next Generation in Citation Analysis[EB/OL]. (2012-11-14). [2014-03-15]. http://www.lis.illinois.edu/Events/2012/09/26/Content-Based-Citation-Analysis-Next-Generation-Citation-Analysis.

[19] 刘盛博, 丁堃, 刘则渊. 基于引用内容的引文检索与推荐系统[J]. 情报学报, 2013, 32(11): 1157-1163. (Liu Shengbo, Ding Kun, Liu Zeyuan.Citation Retrieval and Recommendation Based on Citation Context [J]. Journal of the China Society for Scientific and Technical Information, 2013, 32(11): 1157-1163.)

[20] Bradshaw S. Reference Directed Indexing: Redeeming Relevance for Subject Search in Citation Indexes [C]. In: Proceedings of the 7th European Conference (ECDL'03), Trondheim, Norway. Berlin, Heidelberg: Springer, 2003: 499-510.

[21] Ritchie A, Teufel S, Robertson S. Using Terms from Citations for IR: Some First Results [C]. In: Proceedings of the 30th European Conference on IR Research (ECIR'08), Glasgow, UK. Berlin, Heidelberg: Springer, 2008: 211-221.

[22] Boyack K W, Small H, Klavans R. Improving the Accuracy of Co-citation Clustering Using Full Text [J]. Journal of the American Society for Information Science and Technology, 2013, 64(9): 1759-1767.

[23] 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.

[24] Liu S, Chen C. The Differences Between Latent Topics in Abstracts and Citation Contexts of Citing Papers [J]. Journal of the American Society for Information Science and Technology, 2013, 64(3): 627-639.

[25] Salton G, Wong A, Yang C S. A Vector Space Model for Automatic Indexing [J]. Communications of the ACM, 1975, 18(11): 613-620.

作者贡献声明：

卢超：设计研究方案，设计实验，清洗与分析数据，起草论文；章成志：提出研究思路，讨论研究方案，采集分析数据，论文最终版本修订。

(通讯作者：章成志 E-mail: zhangcz@njust.edu.cn)

Study on the Reference Network of Single Academic Article Based on Citation Content

Lu Chao　Zhang Chengzhi

(School of Economics & Management, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: [**Objective**] To explore the form of the reference networks via the analyzing how the references are cited and disbuted in the content of the academic articles. [**Methods**] Based on the structured data of 575 academic articles, utilize content extraction, similarity computing and other technologies to build the networks of every single article's references and combine examples to analyze the interrelations among them and to find out the reasons. [**Results**] Some negative connections exsist between the similarity of references and their relative distance. Diversification and different models exist in the reference network of a single article as well. [**Limitations**] Some parts of the full-text data are not accurate enough, which affects the results of the experiment.The evaluation of the relative distance among references in this study lacks accuracy. Deep mining of the texts is needed to solve the problem. [**Conclusions**] From the results, the reference network structures can be roughly classfied into three categories, and the causes are different. The reference network of single academic article needs more studies.

Keywords: Citation analysis　Citation content　Network analysis　Text mining