社会网络分析在关键词网络分析中的实证研究*

An Empirical Study of Keywords Network Analysis Using Social Network Analysis

魏 瑞 斌

(安徽财经大学信息工程学院 蚌埠 233041)

摘 要 以关键词共现为基础提出关键词网络的概念。在中国学术期刊网络出版总库中,检索到 1990~2008 年期刊论文中包括特色数据库和专题数据库的记录 527 条,将每条记录中的关键词进行整理。利用社会网络分析方法 Ucinet 对 30 个高频关键词组成的关键词网络进行分析。研究结果表明,国内特色数据库的研究已经形成了一些研究热点,但目前的研究内容还比较分散,研究深度不够。

关键词 关键词网络 社会网络分析 特色数据库 共词分析

中图分类号 G250.74 文献标识码 A 文章编号 1002-1965(2009)09-0046-04

关键词是表达文献主题概念的自然语言词汇。在期刊论文、学位论文、会议论文等文献当中,存在着一些关键词经常同时出现在同一篇文献的现象,这种现象可以称之为关键词的共现现象。共词分析是对关键词共现现象进行研究的一种重要方法。它是对一组词两两统计它们在同一篇文献中出现的次数,以此为基础对这些词进行聚类分析,从而反映出这些词之间的亲疏关系,进而分析这些词所代表的学科和主题的结构变化。利用共词方法可以概述研究领域的研究热点,横向和纵向分析领域学科的发展过程、特点以及领域或学科之间的关系,反映某个专业的科学研究水平及其发展历史的动态和静态结构,拓展信息检索领域以求帮助用户检索信息等等。共词分析方法主要经历了基于包容指数和临近指数、基于战略坐标和基于数据库内容结构分析的共词分析方法三个阶段[1]。本文尝试利用社会网络分析方法来对共词现象进行研究,希望这种方法能给共词分析的研究带来一些新的思路。

社会网络分析(SNA)是 20 世纪 70 年代以来在社会学、心理学、人类学、数学、通信科学等领域逐步发展起来的一个的研究分支。它不仅仅是一种工具,更是一种关系论的思维方式[2]。国内情报学领域一些学者已经利用社会网络分析方法,在竞争情报、知识管理、图书馆资源配置、学科热点、引文分析、科研人员合著、网络链接、博客网络等方面展开了一系列研究[3]。本文是以关键词为研究对象,通过社会网络的视角来对关键词的共现进行尝试性的研究。

1 基于词共现的关键词网络

笔者认为,围绕某一主题研究者在其研究成果(如期刊论文)中经常会使用一些相同的关键词,这些关键词同时出现在一些文献中则表现为关键词共现。而关键词的共现如同作者合作、文献共被引等情况一样,某一主题领域内的关键词共现,实际上会形成一个虚拟的关键词网络。

从社会网络的视角看,关键词是网络中的一个个节点,而它们的共现则体现为节点之间有直接的联系。在虚拟的关键词网络中,由于是否共现和共现频次的不同,每个节点在网络中具有不同的地位,承担不同的角色。在一定的时间范围内,有些关键词反映的是该主题的研究热点;有些词表示的内容处于不成熟的状态;有些词之间的联系非常紧密,有些词会在网络中显得比较孤立。通过对关键词网络的分析,可以发现隐藏在真实关系网背后的关系网络,它对于了解一个研究主题的成熟度、知识结构、研究的规模等状况具有非常重要的意义。

通过以上分析及参考文献[4~9]的研究,笔者认为社会网络分析方法可以运用到某一主题关键词网络

收稿日期:2009-04-12 修回日期:2009-05-18

基金项目:安徽财经大学青年科研项目"特色数据库建设的实证研究"(编号:ACKYQ0836ZC)的研究成果之一。

作者简介:魏瑞斌(1073-),男,博士,研究方向为文献计量学。

的研究当中。这种方法应用的优势在于它可以从定量的角度来确定某个关键词在网络中的地位,以及定量地反映出词与词之间的关系。结合项目研究的需要和参考文献[10],本文的实证部分选择特色数据库这一主题。同时,本文选取 Ucinet 作为数据处理工具,它是研究者使用较多且可以免费使用的社会网络分析软件。它的优势是可以将关键词网络的特征定量化,同时它具有绘图功能,可以将结果可视化显示。

2　关键词网络的数据来源与数据预处理

2.1　数据来源　从方便获取数据和数据量的角度出发,本文选择《中国学术期刊网络出版总库》作为统计源。检索条件确定为:关键词="特色数据库" or"专题数据库";数据的时间范围是 1990～2008 年;期刊来源是为中国学术期刊全文数据库收录的所有期刊。最终检索到 527 条记录,关键词 1929 个,篇均关键词约为 3.7 个,去重后的关键词为 746 个。

2.2　数据预处理　数据预处理是指将从数据库查询的结果处理为可以使用社会网络分析软件 Ucinet 来直接处理的特定的数据格式。具体过程如下:

a.把检索结果保存在一个 excel 文档当中,将去重后的关键词按出现频次多少排序,最后选择出现频次大于 5 次的关键词作为研究对象(见表 1)。

表 1　"特色数据库"研究领域的高频关键词

序号	关键词	频次	序号	关键词	频次	序号	关键词	频次
1	特色数据库	373	11	数字资源	14	21	ASP	6
2	专题数据库	128	12	网络环境	13	22	CALIS	6
3	高校图书馆	92	13	元数据	13	23	版权	6
4	数据库建设	92	14	地方文献	11	24	高校	6
5	图书馆	68	15	信息服务	10	25	公共图书馆	6
6	数字图书馆	34	16	全文数据库	9	26	河洛文化	6
7	数据库	28	17	党校图书馆	8	27	数字化	6
8	建设	25	18	特色服务	8	28	文献资源	6
9	资源共享	16	19	特色馆藏	8	29	信息检索	6
10	TPI	14	20	信息资源	8	30	资源建设	6

b.利用 excel 的数据透视表和数据透视图功能统计出选定关键词之间的共现频率。在文献[11]研究结果,同时考虑极值对网络的影响,下面分析过程中剔除了"特色数据库"这个关键词。这样可以更加准确地分析其它关键词之间的相互关系及它们在网络中所处的地位。

c.将"特色数据库"以外的 29 个关键词之间的共现频率作为矩阵的元素值,最终得到一个 29 行和 29 列的原始矩阵。

d.原始矩阵所有数据的平均值是 0.6338,取近似值 1。将原始矩阵中大于等于 1 的数值取 1,小于 1 的取值为 0,这样得到一个只有 1 和 0 的二值矩阵。将二值矩阵结果保存为 Ucinet 的数据文件,扩展名为 ##d。

3　关键词网络的分析

利用 Ucinet 的绘图功能,二值矩阵的数据可以转换为一个关键词网络图(图 1)。下面主要从节点和网络的中心性和小团体分析两个方面对图 1 所示的网络进行分析。

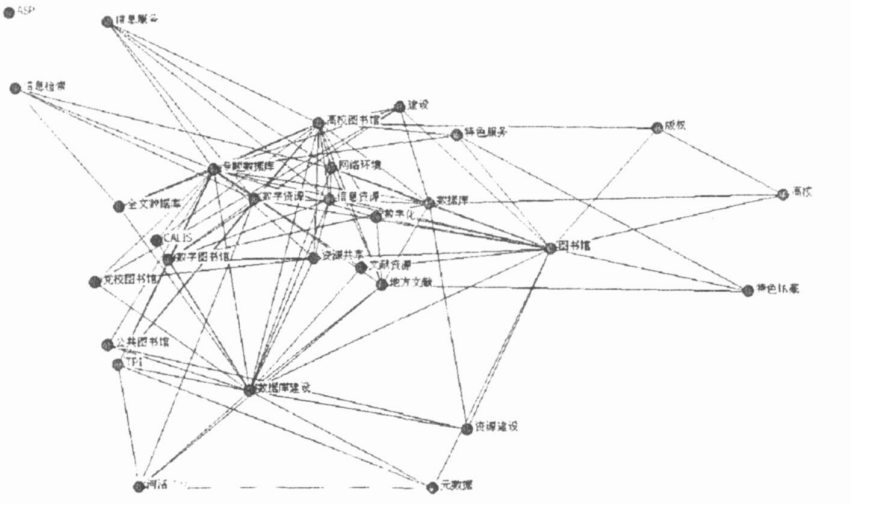

图 1　关键词网络图

3.1　中心性和中心势分析　节点的中心性是指图 1 所示网络中每个关键词在网络中处于什么地位。中心势是反映整个关键词网络中各个节点的差异性程度。由于根据计算方法的不同,节点中心性分为点度中心度、中间中心度和接近中心度三种;网络中心势也分为点度中心势、中间中心势和接受中心势三种,中心势的取值范围介于 0 和 1 之间。

利用 Ucinet 对可以得到网络中每个节点的中心性数值(见表 2)。下面从五个方面对表 2 当中的数据进行分析。

a.点度中心度在本文中反映的是某个关键词与其它词是否共同出现在某篇文献中。如图书馆的绝对中心度是 23,表示它与网络中其它 23 个关键词至少同时出现在某一期刊论文当中。ASP 的绝对中心度是 0,表明它没有和网络中其它词共同出现在某篇论文当中。点度中心度越高,则反映其在网络中的地位越高,越有可能是主题研究中的热点。从点度中心度数据判

断，前 10 个词是当前特色数据库研究的热点所在。

b. 中间中心度在本文中是测量指网络中某个关键词影响它关键词共同出现在一篇期刊论文中的能力大小的指标。从表 1 看，“专题数据库”、“图书馆”、“数据库建设”这三个词的中间中心度与其它词相比，显得非常突出。这表明在特色数据库研究过程中，这些词影响其它词是否共现的能力较强。

c. 接近中心度是衡量的是网络中的某个节点不受其它节点“控制”的能力。在本文中它表示的是网络中某个关键词与其它关键词共现的机率大小。接近中心度越小，表示某个关键词越容易与网络中的关键词出现在一篇期刊论文当中，如图书馆、专题数据库等。

d. 每个节点的三个中心性指标表现不同。图 2 是将表 1 数据规一化处理（表 2 中的 A、B、C 各列的数值除以每一列的最大值）的结果。从表 2 可以看出，有的关键词的三个指标数据都反映其处于网络中的核心地位，如图书馆、专题数据库、数据库建设、高校图书馆等。有些词其绝对中心度高，但是中间中心度和接近中心度较低，如特色馆藏。这可能由于某方面的研究只是集中在少数研究人员研究成果当中造成的。从图 2 整体上看，三个指标在反映中间部分关键词在网络中的地位是相同的，而两头差异比较明显。

表 2 关键词的中心度数据

序号	关键词	A	B	C	序号	关键词	A	B	C
1	图书馆	23	130.387	67	16	河洛文化	5	3.589	82
2	专题数据库	19	139.05	64	17	信息服务	5	1.519	82
3	数据库建设	17	114.395	66	18	党校图书馆	4	0.918	82
4	高校图书馆	16	83.345	67	19	全文数据库	4	1.875	81
5	特色馆藏	12	1.167	88	20	数字化	4	0.515	79
6	数据库	10	30.304	73	21	信息检索	4	2.019	84
7	数字资源	9	23.533	74	22	元数据	4	2.559	81
8	地方文献	8	22.075	75	23	资源建设	4	3.486	81
9	数字图书馆	8	7.728	77	24	CALIS	3	0	82
10	网络环境	8	14.934	75	25	版权	3	2.426	85
11	TPI	7	7.625	78	26	高校	3	0.792	88
12	信息资源	7	5.735	76	27	公共图书馆	3	0.854	83
13	资源共享	7	5.102	76	28	文献资源	3	0	80
14	建设	6	5.102	78	29	ASP	0	0	812
15	特色服务	6	4.97	80		平均值	7.31	21.24	103.31

注：表 1 中的 A 表示关键词的绝对中心度；B 表示绝对中间中心度；C 表示绝对接近中心度。每个关键词的相对中心度、相对中间中心度和绝对接近中心度表中没有列出。

e. 绝对中心势反映的是网络的集中趋势。中间中心势表示网络中中间中心度最高的节点的中间中心度与其它节点的中间中心度的差距。差距越大，则网络的中间中心势越高，表示该网络中的节点可能分为多个小团体而且过于依赖于某个节点传递关系。图 1 网络的点度中心势是 7.52%，中间中心势是 16.14%。由于网络中存在一个孤立的点（ASP），因此无法计算其接近中心势。从结果看，图 1 所示网络的集中趋势较弱，这反映出当前特色数据库的研究内容还比较分散。中间中心势较低，反映出当前特色数据库的研究还没有形成一些核心的研究内容，也就是研究的深度还不够。

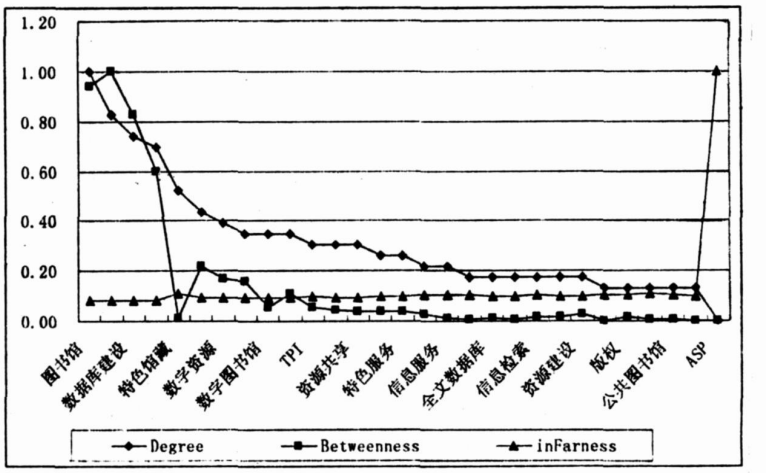

图 2 三个中心性指标规一化数据

3.2 小团体分析 社会网络中的小团体指团体中一小群人关系特别紧密，以至于形成一个次级团体。关键词网络的小团体是指网络中的关键词同时作为某些文献的关键词，相互之间的关联度较强。小团体分析中有四种类型：节点程度分析，如 k－plex、k－core、Lambda Sets；节点距离分析，如 n－clique、n－clan、n－club；绘图分析；小团体密度分析[12]。本文从节点距离和小团体密度两个方面对图 1 的小团体现象进行分析。

3.2.1 n－clique 分析。n－clique 是指小团体内每两个人之间的距离小于等于 n，本文是指每两个关键词之间的距离小于 n，。利用 Ucinet 将图 1 对应数据处理后得到两个 clique，其中 n 为 2，最小规模为 3。利用 Ucinet 小团体分析中的 n－clique 功能，表 1 的 29 个关键词分为了 9 个小团体（限于篇幅，本文只列出 3 个）。

从下面结果看，三个小团体的规模不同，小团体 1 的规模较大，小团体 3 的规模较小。有些关键词同时出现在三个小团体中，如地方文献 高校图书馆等；有些词则只出现在某个团体当中，如 CALIS、TPI。这是由于网络中每两个关键词距离不同而导致的结果。

小团体 1：CALIS TPI 党校图书馆 地方文献 高校图书馆 公共图书馆 建设 全文数据库 数据库 数据库

建设 数字化 数字图书馆 数字资源 特色服务 图书馆 网络环境 文献资源 信息资源 元数据 专题数据库 资源共享 资源建设

小团体 2: 地方文献 高校 高校图书馆 建设 数据库 数据库建设 数字化 数字图书馆 数字资源 特色服务 图书馆 网络环境 文献资源 信息资源 元数据 专题数据库 资源共享 资源建设

小团体 3: 地方文献 高校图书馆 河洛文化 数据库 数据库建设 数字化 数字资源 特色馆藏 图书馆 网络环境 文献资源 信息资源 元数据 专题数据库 资源共享 资源建设

图 3 反映了特色数据库研究领域相关内容之间相互交错的关联结构, 展示了研究主题之间错综复杂的等级关系。图 3 完全依赖于关键词数据之间的天然的联系, 可以较为客观地反映特色数据库领域研究内容之间的内在关联。如地方文献、高校图书馆位于图 3 的最左边, 这些内容是当前特色数据库研究中最为基础的内容; 而 ASP、信息服务、信息检索位于图 3 的最右边, 这些相对而言是本领域较为深入的研究内容。从树状图的聚类效果看, 这些研究内容相互关联的紧密程度不同。

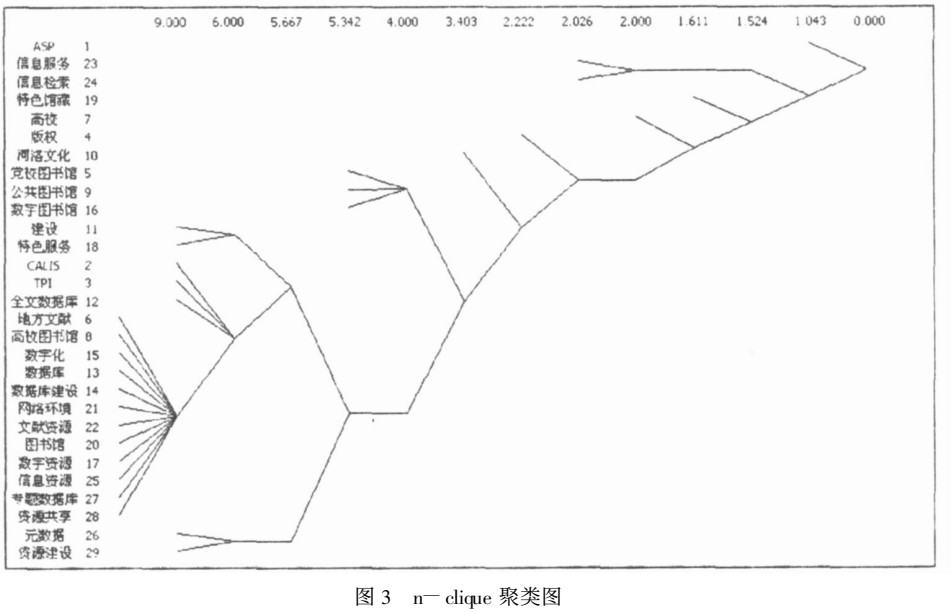

图 3　n－clique 聚类图

3.2.2　凝聚子群的密度。凝聚子群的密度(E－Index)主要用来衡量一个网络中小团体现象是否严重。凝聚子群密度的取值范围为[－1, ＋1]。该值越向 1 靠近, 意味着派系林立的程度越大; 该值越接近－1, 意味着派系林立的程度越小; 该值越接近 0, 表明关系越趋向于随机分布, 看不出派系林立的情形[5]。图 1 网络的凝聚子群的密度值为－0.104, 比较接近于 0, 即图 1 所示网络趋向于随机分布。这反映了图 1 中关键词代表的内容还没有形成一个比较紧凑的派系, 研究状态还比较松散, 研究内容间的关系不是非常紧密。

4　结　论

利用社会网络分析方法及专门的分析工具, 可以把某一主题的关键词网络以可视化的方式呈现出来, 并且利用点中心度等指标把关键词在网络中的地位及其相互关系以量化的形式予以揭示。这种方法可以把某一研究主题的关系词在虚拟关键词网络中的地位及其相互关系定量地描述。它对于了解某一研究主题的研究结构, 构建所在领域的知识地图等方面有独特的优势。本文构建的二值矩阵无法完全反映出关键词共现次数多少对整个关键词网络的影响; 实证过程中数据量较少, 这些都是今后需要进一步探讨的问题。

参 考 文 献

1　冯　璐, 冷伏海. 共词分析方法理论进展[J]. 中国图书馆学报, 2006(2): 88－92

2　约翰斯科特著; 刘军译. 社会网络分析方法(第 2 版)[M]. 重庆: 重庆大学出版社, 2007: 6

3　朱庆华, 李　亮. 社会络分析法及其在情报学中的应用[J]. 情报理论与实践, 2008. 31(2): 179－183

4　刘则渊, 尹丽春. 国际科学学主题共词网络的可视化研究[J]. 情报学报, 2006, 25(5): 634－640

5　裴　雷, 马费成. 社会网络分析在情报学中的应用和发展[J]. 图书馆论坛, 2006(6): 40－45

6　徐媛媛, 朱庆华. 社会网络分析法在引文分析中的实证研究[J]. 情报理论与实践, 2008(2): 184－188

7　刘　蓓, 袁　毅, BOUTIN Eric. 社会网络分析法在论文合作网中的应用研究[J]. 情报学报, 2008(3): 407－417

8　李　亮, 朱庆华. 社会网络分析方法在合著分析中的实证研究[J]. 情报科学, 2008(4): 549－555

9　党洪莉, 孙红霞. 图书情报学博客的社会网络分析[J]. 情报杂志, 2009(1): 180－182, 168

10　郭黎康. 近十年特色数据库研究论文的统计与分析[J]. 农业图书情报学刊, 2007, 19(2): 166－169

11　魏瑞斌. 国内特色数据库研究现状分析[EB]. 图书情报工作网刊, 2008. 12

12　罗家德. 社会网分析讲义[M]. 北京: 社会科学文献出版社, 2005: 134－146

(责编: 白燕琼)