

一种内容和引用特征融合的知识结构划分方法研究

任红娟

摘 要 通过对现有知识结构划分方法的梳理,总结了当前的知识结构融合方法,用数据表明了融合文献的内容和引用特征进行知识结构划分的必要性。在此基础上,通过实验证明用最大值函数将文献的标题和摘要内容特征以及文献耦合引用特征融合进行知识结构划分的效果,并将其与现有方法进行对比。结果表明:该融合函数可以提高知识结构划分的精度。图2。表4。参考文献17。

关键词 知识结构 融合方法 融合函数 内容—引用

分类号 G350

A Division Method of Intellectual Structure of Combining Content and Citation of Documents

Ren Hongjuan

ABSTRACT Based on systematical investigation of existing methods of intellectual structure, this paper analyzes the current fusion method of intellectual structure, and shows the necessity of combining content and citation features of literature to divide intellectual structure with real data. On this basis, this paper conducts a research with experiments on the effects of maximum function dividing intellectual structure with the use of combining the content features of titles and abstracts of literature with citation features of bibliographic coupling, and compares the method of experiments with existing methods. The results show that this fusion function has improved the accuracy of dividing intellectual structure. 2 figs. 4 tabs. 17 refs.

KEY WORDS Intellectual structure. Fusion method. Fusion function. Content-citation.

1 引言

科学的知识结构研究是跨学科研究的重要选题,通过对科学单元,比如文献、期刊和学科进行分类研究,分析和展示科学的内在结构特征是当前国内外相关领域的热点研究问题^[1]。

科学知识结构划分的方法有很多,有的对同被引分析、文献耦合分析、共词分析、引文分析和共词分析相结合的四中知识结构划分方法进行了系统的分析^[2]。有的将知识结构划分方法分为三类:基于引文分析的方法(包括文献耦合分析方法、同

被引分析方法、直接引文分析的方法)、内容词分析方法(共词方法)和基于引文—内容词的混合分析方法^[3]。知识结构划分方法各有优劣,而上述研究的作者都提及了把文献的词和引用关系结合在一起研究知识结构的方法,并指出这是知识结构分析方法发展的趋势。

王彦雨和池田指出:科学学的发展历程中主要形成了内容无涉的客观文本范式以及内容有关的社会认知构建范式^[4],韩毅等人基于他们提出的两种范式指出:作为科学学重要组成部分的科学计量学研究主要以客观文本范式为主,较少关注社会认知构建范式,引文网络分析更是把客观文本范式发

挥到极致,随着引文网络分析的深入发展,单纯的客观文本分析范式发展空间越来越小,与其他方法整合共同完善引文网络分析方法体系已成为其向纵深演化拓展的重要趋势^[5]。Morris 在“Mapping Research Specialties”一文中用一张盲人摸象图形象地隐喻从一个特征或关系来刻画领域知识结构的局限性和片面性,作者认为每种方法都能够反映科学知识体系的一个侧面,尽管这些方法得到的结果不是完全分离的,但至少可以肯定是存在偏差的^[6]。这些研究从理论和方法都表明将文献的内容和引用特征融合的必要性。

笔者对文献内容和引用融合的知识结构划分方法研究进展进行了系统的分析^[7],其中以 Janssens 为代表的研究团队在 Janssens 博士论文^[8]基础上发表了一系列相关的研究成果^[9-12],结果表明利用线性融合函数和费歇尔的逆卡方方法(Fisher's inverse chi-square)将文献的全文内容词和文献耦合得到文献相似(不相似)矩阵融合的方法比单一方法得到的知识结构划分更加准确。

然而,Janssens 只研究了 0.5 权重的线性融合函数,没有对权重的选择进行进一步的研究,而逆卡方方法计算复杂度非常高,且原始数据的随机化处理 and P 值的转换每次操作都会不一样,导致最终结果不稳定。本研究拟提出一种更加简单的融合函数将文献的内容和引用特征融合,并将得到的结果与 Janssens 提出的方法得到的结果进行对比,验证本文提出的融合函数是否有效。

2 数据与方法

2.1 数据集构建

2005 年,Glenisson 等人为了验证把文献的全文文本挖掘方法和文献计量学方法结合用于映射科学知识结构的有效性,对 2003 年在中国北京召开的第九次国际科学计量学和信息计量学会议(International Conference on Scientometrics and Informetrics, ISSI)上的 19 篇会议论文进行了初步分析,证明了该方法的有效性^[13]。同年,他们又利用相同的方法对 2003 年刊载在 *Scientometrics* 期刊上文献类型为“article”

的文献集合进行了分析^[14]。该论文详细列出了分析文献的标题以及专家给出的这些文献的具体分类结果。因此,本研究考虑到结果的可对比性和最终知识结构划分“金标准”的可获取性,选取 Glenisson 构建的数据集进行分析,从 web of science 中进行数据检索,并将文献的类型限定为“article”,最终得到 79 篇文献,并将得到的文献标题与文章对应,确定每篇文献所属的类别。

2.2 研究方法

2.2.1 文本预处理

文献的内容特征包含标题、摘要、作者关键词、增补关键词、全文等,有的学者根据全文的特点又进一步把全文中的前沿、结论和讨论部分进行区分,甚至对于文章段落中的位置和句子在段落中的位置以及词在句子中的位置进行更细粒度的内容特征项的划分。侯汉清等人指出,在提取反映文本主题内容的概念词串时,不同标引源,即文本的不同部位,对文本的主题表达能力是有区别的^[15]。一般而言,文献的标题都能够用简单的词汇组合表达出最核心的主旨,且处理起来相对比较迅速。因此,本研究选取文献的标题进行文本预处理。采用 wvtool 开源文本预处理软件对文献的标题进行处理,经过去除停用词和词干化处理之后,构建文本向量空间,即原始的[标题词—文献]矩阵。

2.2.2 潜在语义分析

然而,原始文本向量空间非常稀疏,经过计算后文本的相似性之间区分度很低,很难达到很好的知识结构划分的效果。本文采用潜在语义分析(Latent Semantic Analysis,LSA)对原始文本向量进行处理。潜在语义分析,又称潜在语义索引,出发点是文本中的词与词之间存在某种联系,即存在某种潜在的语义结构。它不需要确定的语义编码,仅依赖于上下文事物的联系,用语义结构来表示词汇和文本,用统计的方法实现简化文本向量,达到去除噪音信息、保留最有用信息的目的。潜在语义分析主要是采用数学方法中的奇异值分解来压缩原始的向量空间,从而把高维度的矩阵转化为保留主要模式的低维度矩阵,凸显隐藏在文本中的主要语

义信息。

2.2.3 融合函数

本研究基于保留最主要信息的思想,在 Janssens 等人对知识结构融合方法研究的基础上,提出了一种相对简单的融合函数:最大相似性函数,如公式(1)所示:

$$f(t, c) = \max(S_i, S_c) \quad (1)$$

(其中 S_i 表示从文献内容特征得到的文献相似性, S_c 表示从文献引用特征得到的文献相似性)

该函数的思想是对每个文献对取最大的相似值,最终得到的矩阵是两两文献对之间的最大值形成的新矩阵。假设 S_i 中的所有元素都大于或都小于 S_c 中的元素时,得到的矩阵就是线性融合函数权重为 0 或 1 的特殊情况,但这种极端的情况几乎不可能出现,而且即使有这样的巧合,结果也是在考虑两个角度的情况下得出的。因此,该函数符合融合的思想,且在逻辑上是可行的。

2.2.4 聚类有效性评价

文献计量学中,对于知识结构的划分一般采用层次聚类、因子分解或社团划分的方法对得到的相似矩阵或原始矩阵进行聚类分析,从而得到领域的知识结构。赵星等人指出:学科或领域的划分一直是计量学研究中的难点,研究者的分类常难得到公认和实际应用^[16]。聚类评价的方法包括三种指标:外部指标、内部指标和相对指标。对于聚类结果评价没有统一的标准,但它们的操作宗旨都是本着同类簇对象相似性最大、不同类簇之间对象相似性最小这个原则。本文选取的文献集有专家给出的“金标准”来确定每篇文献所属的类别。因此,将采用最常用的外部评价指标 F-measure 来进行度量。

F-measure 指标在聚类评价中是综合考虑召回率(Recall, R)和准确率(Precision, P)两个指标来评价聚类质量的一种外部评价指标,它以专家或者已知的分类标准作为评价依据,把聚类得到的结果和这个“金标准”做对比,从而得到聚类有效性的量值。表 1 是聚类结果与已知分类标准的可能性矩阵(contingency matrix)^[17]。

表 1 聚类评价可能性矩阵

		聚类结果	
		同类	不同类
分类标准	同类	TP (true positive)	FN (false negative)
	不同类	FP (false positive)	TN (true negative)

由此,可以得到的聚类有效性公式为:

$$P = TP / (TP + FP) \quad (2)$$

$$R = TP / (TP + FN) \quad (3)$$

$$F = 2 \times P \times R / (P + R) \quad (4)$$

P、R、F 均分布在 [0, 1] 区间,它们的值越趋近 1,表明聚类的质量越好。

3 实验研究

3.1 文献内容和引用特征融合的必要性

为了更加明确地说明把文献内容和引用特征融合的必要性,笔者对数据集从文献标题和文献耦合两种关系分别得到的不同文献相似矩阵,选取相关系数大于 0.4 的对象,利用 UCINET 软件进行分析,分别得到图 1 和图 2。

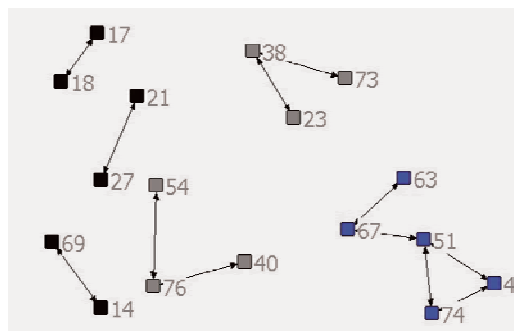


图 1 基于文献标题(内容)的文献
相关系数 > 0.4 的文献关系图

由图 1 可知,从文献标题词得到的文献之间相关系数大于 0.4 的文献仅有 17 篇,且这些文献之间的关系是比较松散的,多数是两两之间的联系。和专家分类的结果进行对比可知,文献 4、51 和 74 属于同一类,都是研究期刊的共被引,被划分为实

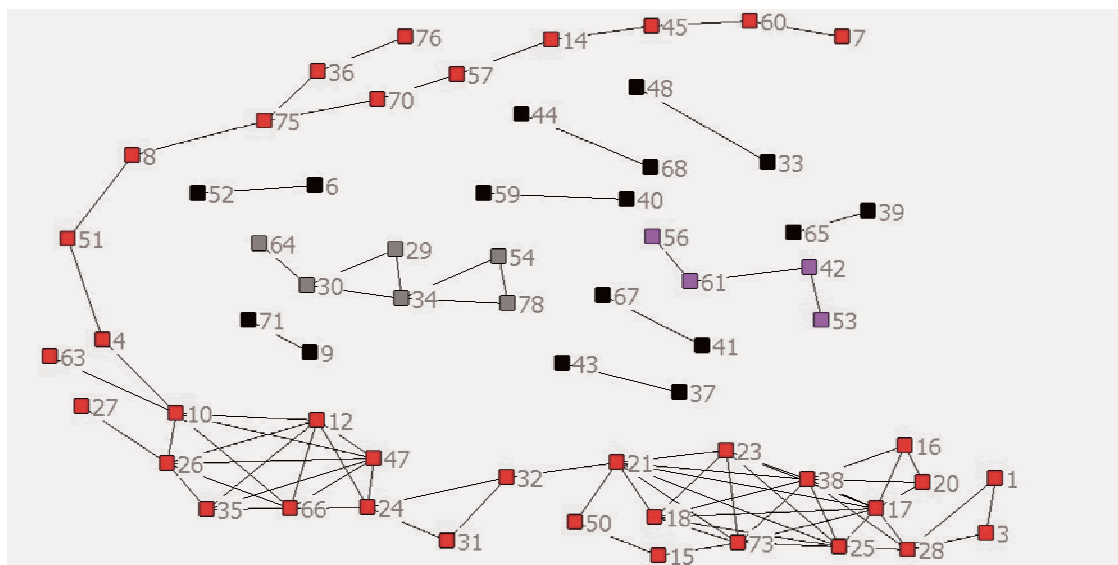


图2 基于文献耦合(引用)的文献相关系数 >0.4 的文献关系图

证研究类,而文献63和67都是研究期刊的引用模式,但是一个属于应用研究,一个属于方法研究,因此被专家分在了不同的类。由此可见,由于词的模糊性,在标题中采用同样的词并不一定属于同一个类,单纯采用内容词不一定能够很准确地划分领域或者研究主题的知识结构。

由图2可知,从文献耦合关系得到的文献之间相关系数大于0.4的文献几乎包含了全部的数据,而且大多数文献之间都有联系。

从这两个图我们可以大体地了解,从不同的分析角度得到的文献之间的关联关系是不同的,有时还会有非常大的差异,而如果能够把从不同角度得到的数据融合在一起,就能够相对全面地反映文献之间的真实关系,从而更加准确地揭示领域的知识结构,避免信息选择单一造成的结果偏差。

3.2 实验结果

利用最大相似性融合函数把文献的标题未潜在语义分析和潜在语义分析的相似矩阵分别与文献耦合相似矩阵融合,得到了如表2和表3所示的结果。

从利用最大相似性融合函数融合的结果来看,融合后知识结构划分的准确率P值比仅采用一种信息源得到的值要高,而F值作为一个综合召回率和准确率的指标,标题未经LSA和文献耦合融

表2 利用max函数将标题和文献耦合相似矩阵融合(未LSA)

	P	R	F
标题	0.3235	0.2903	0.3060
耦合	0.3581	0.3819	0.3696
融合	0.4337	0.3542	0.3899

表3 利用max函数将标题和文献耦合相似矩阵融合(LSA)

	P	R	F
标题 LSA	0.3108	0.2361	0.2684
耦合	0.3581	0.3819	0.3696
融合	0.3769	0.2764	0.3189

合得到的值比单一指标的最大值高。而经过潜在语义分析之后的融合结果F值则低于其中文献耦合得到的F值,这是由于经过潜在语义处理之后,从标题来进行知识结构划分的准确率和召回率都要低于未经潜在语义分析的结果,这与样本数据的特性和潜在语义分析之后有负值出现有一定的关联。

总体而言,作为衡量类和簇统一性的指标P值都比单一信息源绩效高,这表明了该函数的有效性。

3.3 与 Janssens 方法的比较

Janssens 主要是采用了两种融合函数,一种是线性权重的融合函数,作者取权重为 0.5,第二种方法是元分析统计方法中的逆卡方方法,根据数据的分布特征将原始数据转换为概率 P 值,利用费歇尔的逆卡方公式将不同的独立检验结果融合在一起,以求能够从整体上把握事物。这种方法在生物医学领域的应用非常广泛,也已证明了这种方法把不同数值融合在一起的有效性。但是 Janssens 的研究结果表明利用潜在语义分析时的结果不如线性权重融合方法,因此本文仅对没有经过潜在语义分析的数据进行分析。

依照 Janssens 采用的分析流程^[8],利用与其相同的随机化方法,将初始矩阵转化为概率矩阵,利用逆卡方方法把不同的 P 值融合在一起。传统的逆卡方方法没有考虑每个独立检验权重的影响,而 Janssens 采用对数据赋权重的方法对传统逆卡方公式进行了修正,如公式(5)所示,本文选取 λ 为 0.1—0.9 间隔为 0.1 的数值作为权重,将结果与前面融合函数的结果进行对比。

$$P^i = -2\log(P_1^\lambda \cdot P_2^{1-\lambda}) (0 < \lambda < 1) \quad (5)$$

表 4 是利用逆卡方方法得到的文献标题和文献耦合整合的聚类效果表。从表 4 来看, $\lambda = 0.2$ 时,F 最大,而 max 函数的值为 0.3899,比利用逆卡方元分析函数得到的结果要好。并且,逆卡方方法需要经过随机化处理、累积分布函数的 P 值转化、相似性转化等多个步骤,计算的时间复杂度非常高;并且由于对数据可以采用多种随机化方案,即使采用同一个随机化方案得到的结果也不太稳定,而且初始数据的分布规律相对来说也比较难把握。采用不同 λ 值得到的结果差异也较大, λ 值的确定也比较难把握,而 max 函数不但易于操作,而且易于理解,并且无需经过多个步骤的转化,信息冗余和信息损失的概率比较小,因此不会造成太多信息失真和噪声融入,对于实践的作用更大。而且 P 值方法在数据集非常大的情况下效果比较好,数据集较小的情况则不太适合,而本文提出的简单函数不仅能够达到提高知识结构划分精度的效果,而且数据集的数量无论大小都是适用的。

表 4 标题和文献耦合利用逆卡方方法融合的聚类有效性

	标题权重	文献耦合权重	P	R	F
λ	0	1	0.3581	0.3819	0.3696
	0.1	0.9	0.3577	0.4138	0.3877
	0.2	0.8	0.3644	0.4180	0.3894
	0.3	0.7	0.2838	0.4806	0.3569
	0.4	0.6	0.3275	0.2875	0.3062
	0.5	0.5	0.3424	0.2806	0.3084
	0.6	0.4	0.2709	0.3778	0.3155
	0.7	0.3	0.3450	0.2597	0.2964
	0.8	0.2	0.2920	0.4486	0.3538
	0.9	0.1	0.2686	0.4167	0.3266
	1	0	0.3235	0.2903	0.3060

4 结论

本研究提出了一种最大文献相似性函数,把从文献内容特征和文献引用特征得到的文献相似

性融合,实验表明这种方法相比单一的知识结构划分方法更加准确,而且操作也非常简单。

文献的内容特征在表达文献主题方面有不同的作用,本文仅研究了标题,对于其他的内容特征项未加考虑。根据不同文献集合的特征选取不同

的文献特征项,并且与文献的其他引用关系相融合还需要进一步研究。另外,本研究主要选取的样本量相对比较小,而实际的计量分析中数据量相对比较大,且没有“金标准”。采用更大的样本和内部

聚类有效性评价方法来评价知识结构划分的准确性,并在大样本中验证该方法的有效性将是未来研究的重点。

参考文献

- [1] 张琳,梁立明,Janssens F, et al. 混合聚类方法用于科学结构研究——以“心理学、社会学和教育学”领域为例[J]. 科学学研究, 2010, 28(6): 837-845, 808. (Zhang Lin, Liang Liming, Janssens F, et al. Strategy foresight of innovation and development[J]. Studies in Science of Science, 2010, 28(6): 837-845, 808.)
- [2] 卫军朝,蔚海燕. 科学结构及演化分析方法研究综述[J]. 图书与情报, 2011(4): 48-52. (Wei Junchao, Wei Haiyan. Review on analysis methods of science structure and evolution[J]. Library & Information, 2011(4): 48-52.)
- [3] 葛菲,谭宗颖. 基于文献计量学的科学结构及其演化的研究方法述评[J]. 情报杂志, 2012, 31(12): 34-39, 50. (Ge Fei, Tan Zongying. Review of science structure and evolution of bibliometric methods[J]. Journal of Intelligence, 2012, 31(12): 34-39, 50.)
- [4] 王彦雨,池田. 科学文本研究的神化范式及其转变[J]. 科学学研究, 2009, 27(3): 328-333. (Wang Yanyu, Chi Tian. The myth analysis mode of scientific texts and its change[J]. Studies in Science of Science, 2009, 27(3): 328-333.)
- [5] 韩毅,张克菊,金碧辉. 引文网络分析的方法整合研究进展[J]. 中国图书馆学报, 2010: 36(4): 83-89. (Han Yi, Zhang Keju, Jin Bihui. Research progress on methodology integration of citation network analysis[J]. Journal of Library Science in China, 2010, 36(4): 83-89.)
- [6] Morris S A, Van der Veer Martens B. Mapping research specialties[J]. Annual Review of Information Science and Technology, 2008, 42: 213-295.
- [7] 任红娟,张志强. 基于文献内容和链接融合的知识结构划分方法研究进展[J]. 情报理论与实践, 2010, 33(4): 124-128. (Ren Hongjuan, Zhang Zhiqiang. Progress in research on intellectual structure division method based on fusion of document content and link[J]. Information Studies: Theory & Application, 2010, 33(4): 124-128.)
- [8] Janssens F. Clustering of scientific fields by integrating text mining and bibliometrics[D]. Belgium: Katholieke Universiteit Leuven, 2007.
- [9] Zhang L, Janssen F, Liang L M, et al. Hybrid clustering analysis for the domain of Psychology, Sociology & Education [C]//2009 proceeding of ISSI.
- [10] Zhang L, Janssen F, Liang L M, et al. Hybrid clustering analysis for mapping large scientific domains[C]//2009 proceeding of ISSI.
- [11] Liu X H, Yu S, Janssen F, et al. Weighted hybrid clustering by combining text mining and bibliometrics on a large-scale journal database [J]. Journal of the American Society for Information Science and Technology, 2010, 61(6): 1105-1119.
- [12] Liu X H, Glänzel W, Moor B D. Optimal and hierarchical clustering of large-scale hybrid networks for scientific mapping [J]. Scientometrics, 2012, 91(5): 473-493.
- [13] Glenisson P, Glänzel W, Persson O. Combining full-text analysis and bibliometric indicators: A pilot study[J]. Sciento-

metrics, 2005, 63(1): 163 – 180.

- [14] Glenisson P, Glänzel W, Janssen F, et al. Combining full text and bibliometric information in mapping scientific disciplines [J]. Information Processing and Management, 2005, 41(6): 1548 – 1572.
- [15] 侯汉清,章成志,郑红. Web 概念挖掘中标引源加权方案初探[J]. 情报学报, 2005, 24(1): 87 – 92. (Hou Hanqing, Zhang Chengzhi, Zheng Hong. Research on the weighting of indexing sources for Web concept mining[J]. Journal of the China Society for Scientific and Technical Information, 2005, 24(1): 87 – 92.)
- [16] 赵星,谭旻,余小萍,等. 我国文科领域知识扩散之引文网络探析[J]. 中国图书馆学报, 2012, 38(5): 59 – 67. (Zhao Xing, Tan Min, Yu Xiaoping, et al. Exploring the citation networks for knowledge diffusion of humanities and social sciences in China[J]. Journal of Library Science in China, 2012, 38(5): 59 – 67.)
- [17] 杨燕,靳藩, Mohamed K. 聚类有效性评价综述[J]. 计算机应用研究, 2008, 25(6): 1630 – 1632, 1638. (Yang Yan, Jin Pan, Mohamed K. Survey of clustering validity evaluation[J]. Application Research of Computers, 2008, 25(6): 1630 – 1632, 1638.)

任红娟 郑州航空工业管理学院讲师。通讯地址:郑州市二七区大学中路2号。邮编:450015。

(收稿日期:2013 – 02 – 01)