中国科学学40年研究主题变迁

——基于特征最大化F指标的文本内容分析

陈 悦[1] Jean-Charles Lamirel[1,2] 刘则渊[1]

（1. 大连理工大学 科学学与科技管理研究所暨WISE实验室，辽宁 大连 116085；
2.University of Strasbourg Synalp-Team-LORIA法国 67000）

摘要：基于F指标的特征最大化的GNG聚类方法，对科学学研究文献文本进行内容分析，绘制了中国科学学近40年的研究主题结构图谱，并附以论文发表时间和作者辅助信息的外生标签梳理出中国科学学研究主题的变迁。这种结合了F指标特征最大化无监督学习方法的分析结果显示科学学研究在近40年逐渐走向成熟，从学科一般属性探讨转向相关学科与知识结构分析，从定性分析转向偏重于定量分析和可视化分析，从科学的一般社会功能研究转向更为具体的经济功能和战略功能研究。

关键词：中国科学学；主题变迁；F指标；无监督学习

中图分类号：G301;G202 **文献标识码：**A **文章编号：**1002-0241（2018）12-0028-18

0 引 言

科学学，即“科学的科学（science of science）”，是科学的自我反思，它以整个科学技术知识及其活动为研究对象，探索科学技术发展的基本规律。早在1910年代便已在波兰萌芽，波兰学者对“科学”研究的态度从形而上学转向实证研究，从单一学科分析转向对科学的整体研究，从而奠定了“科学学”作为一门学科的理论研究基础[1]。贝尔纳（Bernal J D）的《科学的社会功能》（1939）被公认为是“科学学”诞生的标志，这本著作的完成直接导索于有着深刻马克思主义思想渊源的“格森事件”[2]，也就是说，贝尔纳科学学思想直接来源于马克思早在150年以前就深刻论述的“科学的本质及其社会功能”，即科学从来就是社会的，就像社会从来就是科学的一样[3]，这是科学学的基本命题之一。《科学的社会功能》一经出版便引起世界关注，尤其是其中专门讨论到中国的科学[4]，同时也很快引起了中国科学家竺可桢（1890—1974，时任浙江大学校长）、吴学周（1902—1983，时任化学所所长）和任鸿隽（1886—1961，中国科学社创始人之一）等人的关注[5]，并迅速在中国传播开来。

科学学在中国的正式诞生是以1977年钱学森在“现代科学技术”一文中倡议建立一门以现代科学技术为研究对象的称为“科学的科学”的学问为起点[6]，他强调科学学是一门社会科学[7-8]。科学学作为一门综合性交叉学科，其核心内容是以整个科学技术知识及其活动为研究对象，探索科学技术发展的基本规律，其研究范围涵盖了科学的历史研究、哲

收稿日期：2018-07-17

基金项目：大连理工大学《科学学原理》精品课程建设（20160916）

第一作者简介：陈悦（1975—），女，辽宁大连人，教授、博士生导师，博士，研究方向：科学计量学与创新管理。

通信作者：陈悦，chenyuedlut@163.com

注：本文感谢WISE实验室的王智琦、谭建国、宋超、宋凯、郭少聪、周京生、祝嘉欢、徐子彦、杨振力等在主题词翻译、校对和处理方面做出的大量工作。

学研究和社会学研究及经济学研究。正是由于研究对象和研究内容的明确，在中国科学学与科技政策学会及其主办的学术期刊的支持下，有大批学者从事科学学的研究。目前不仅京津沪3所科学学研究所，而且中国科协、科技部科技战略研究院、中科院科技战略咨询研究院、中国工程院战略研究院和一批高校，实际上是科学学基础研究与应用研究的重要力量，科学学正以其特有的姿态顽强而富有生机地发展着。

“科学学”在国际上的发展并非坦途，以贝尔纳奖为线索便可一览其发展轨迹（见图1）。大抵分为3条研究进路，“科学计量学（Scientometrics）”“科学技术与社会（STS）”和“科学知识社会学（SSK）”。普赖斯继承和发展了贝尔纳的科学学理念与范式，深化和拓展了科学学理论与方法[9]，强调科学学的数据基础和定量研究。美国科学社会学家默顿（Merton R K）将科学、技术与社会（STS）相互关系作为一个独立对象进行系统考察，但“排除了对于科学知识内容进行社会学研究的可能性”[10]。随之，科学社会学的研究在发展中被不断分化，尤其是SSK的创建，将科学学研究内容重心从“社会层面”转向“认知层面”，随后又有“人类学”和“伦理学”等分支，实际上已逐渐背离贝尔纳最初的科学学理论范式。但值得关注的是，2018年有2篇关于以“Science of Science”为标题的论文分别发表在Physics Reports和Science上[11-12]，一篇是来自于中国

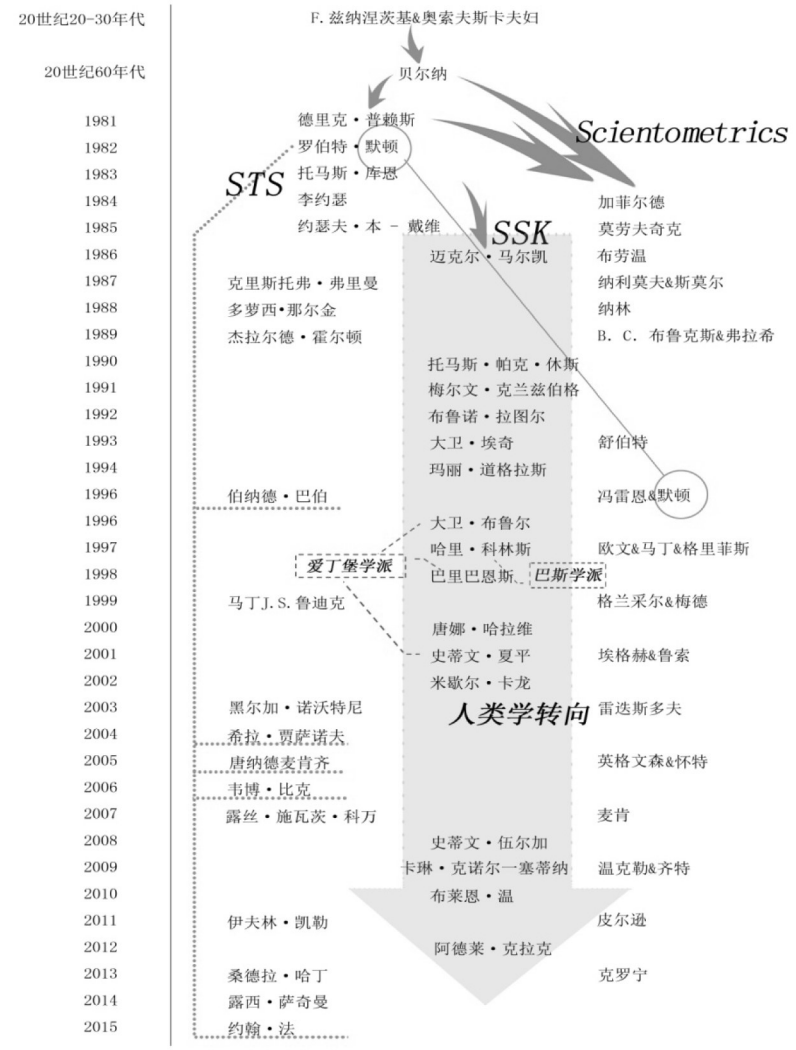


图1　科学学研究进路[13]

北京师范大学系统科学研究团队，另一篇是来自于美国印第安纳大学和莱顿大学的信息可视化研究团队，另外，美国东北大学以物理学为学科背景的复杂性网络研究团队也发表了很多高水平的科学学研究成果，这似乎预示着贝尔纳科学学理论范式下的科学学研究正在回暖。

在中国，相比于其他学科，科学学的研究者们更愿意对学科建设与发展进行反省和回顾[10,14-17]，根据时代背景及时调整方向，逐渐探索出了一条具有中国特色和时代特色的学科发展之路。科学学在中国成立20年之际（1997年），冯之浚先生高度评价“中国科学学从零起步，20年内基本上达到了贝尔纳开创科学学以来国际科学学60年发展的前沿和水平”[18]。刘则渊教授在追忆冯之浚先生的学术贡献时提到[19]：“1977年科学学诞生时我们共同开创了科学学学科建设的第一次浪潮；20年前的世纪之交，我们追踪“冯之浚之问”开创了新世纪科学学发展的新局面，形成了学科建设的第二次浪潮；如今又走过了20年，我们将努力推进科学学学科建设的第三次浪潮。”2010年，刘则渊教授借《中国科学学与科技管理》创刊30年之际，通过对中国科学学主题文献的可视化分析，展现了中国科学学研究30年的前沿与主要领域，呈现出以科学的量化分析为主导的科学计量学和以科学的哲学分析为主导的科学技术学两大知识板块，以及相应的科学学两条互补的发展路径[20]。时至今日，又一个10年，秉承着历史客观性和当代使命感，运用科学学本身与时俱进的研究方法，再次追溯中国科学学40年来的发展历程，试图为中国科学学研究从哪里来，又要到哪里去提供一些启发。

1　数据来源及处理

科学学研究所涉猎的内容和话题是广泛的，其学科边界是模糊的，本文将分析科学学最为核心内容的进展。既完整又精准地提取出“科学学”研究的文献并非易事，在此本文选择了优先考虑“精准”，再次考虑“完整”的原则，将数据检索策略确定为“种子文献+施引文献”。具体而言，以“科学学”和“科学的科学”为检索词在中国知网CNKI数据库中主题检索到2 401篇文献（核心期刊和CSSCI期刊录用，检索时间：2017-10-22），清洗数据后（删除诸如“征文通知”，“会议召开通知”，“杂志简介”和“敬告读者”等类型文献），精炼出1 334篇学术研究论文，以此为种子文献，检索得到此1 334篇学术论文在CNKI中的2 677篇施引文献（去重后），其中有1 539篇为北大核心期刊收录文献。最终，本文以这1 334篇种子论文及其1 539篇施引论文并集而成的2 789篇文献作为研究对象。期刊论文的数量变化趋势（见图2）与刘则渊教授在10年前的判断是一致的，即中国科学学经历了3个阶段，即迅速兴起期（1977—1990）、曲折发展期（1992—2003）和学科复兴期（2003至今）。

提取2 790篇文章的标题、摘要和关键词（1997年以前发表的没有摘要和关键词信息的论文，本文

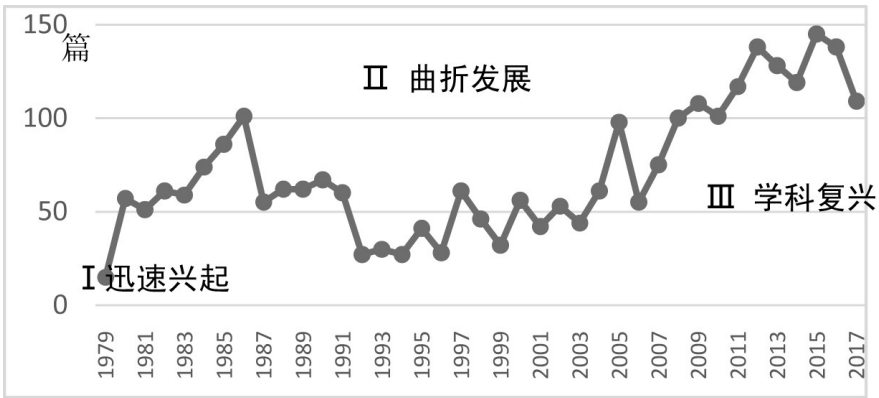


图2　中国科学学期刊文献数量变化趋势

只提取文章的标题信息），利用NLPIR进行分词处理。由于科学学学科的特殊性，软件无法准确分割某些专业术语，如“科学学”、“科学学研究”、“科学逻辑学”、“科学的社会功能”等，因而，本文采取了以下5个步骤进行。（1）建立用户词典。将2 790篇论文的所有关键词（去重后共计9 679个，并将它们的词性标记为名词(n)作为词典导入分词系统；（2）提取名词。将每篇文章分词结果作为一个单元统一编号，并用Python语言提取出其中的名词（标注为/n），自动除去无实意的名词，通常是占一个字符，如“数”、“量”、“人”、“年”等，最终获取13 442个名词；（3）清洗13 442个名词。如将“著者分布”和“作者分布”，“作者合作网”和“作者合作网络”，“知识图谱”和“知识图谱分析”等合并，除去“当局”、“研究”、“分析”、“年份”等无意义的名词；（4）名词翻译。由于本文后面所进行的数据分析程序目前只适用于英文文本，因而将第3步处理后的数据翻译为英文（由于中英文表达上的差异，在此又初步合并了许多中文的异构同义词，如“知识图谱”和“可视化图谱”等都对应为“knowledge mapping”，“贝尔纳”和“J.D.贝尔纳”和“J.D.Bernal”等都对应为“J.D.Bernal”，并将人名、地名和国家进行额外标注，即在相应的词后面增加“name”、“city”、“country”，最终获取11 930个英文名词；（5）将英文名词统一编号（0-11929），并替换2 790篇论文中的名词，将名词在对应文章中出现的频次作为权重，并附以对应文章发表年份，整理后将此部分数据作为中国科学学研究的最初词典。由于信息噪音的存在，本文对所形成的英文初始词典进行了合并等价词、去模糊词和词频控制3步处理（见表1），最终选择1 576个科学学研究的代表词汇。需要说明的是，用这1 576个词汇进行重新检索，并没有丢失任何文献信息，说明了这些词汇用于分析中国科学学研究的有效性。

2　基于F指标的特征选择方法

本文的主要研究方法是基于特征提取与特征选择展开的，即特征最大化方法（feature maximization）。特征最大化是一种无偏度方法，还可以用于分类的质量评估。在无监督学习（如聚类）中，特征最大化方法能够通过提取聚类关联特征来进行聚类标签、主题提取、可视化和聚类优化模型检测。它的主要优点是无参数，适用于高维数据聚类及算法设计，并在分化（discrimination）和泛化（generalization）之间表现出比通常指标（欧几里德、余弦或卡方）更好的和解性（compromise），泛化和分化是学习理论中的一对概念，泛化是从不同中找出共同的特征，分化是从相似中找出区别特征。

2.1　度量特征的F指标

F指标是用于测度某个特征表达某个分类突出特征的能力的指标，本文中所分析的特征是由从文章的题目、摘要和关键词中提取出来的名词来表征，而各个分类是基于GNG算法而形成聚类。若数据集D经聚类得到由一组特征F表征的分区C，那么，某个聚类$c(c\in C)$的关联特征f的度量指标$FF_c f$被定义为“特征召回率（feature recall）$FR_c(f)$”和“特征主导率（feature predominance）$FP_c(f)$”的调和平均值。即：

$$FR_c(f)=\frac{\sum_{d\in c} w_d^f}{\sum_{c\in C}\sum_{d\in c} w_d^f} \tag{1}$$

表1　词典处理过程结果

步骤	合并等价词	删除模糊词	词频控制(>6)
初始数据规模	11 931	11 696	11 571
处理掉的词	—	125	9 995
合并的词	235	—	—
最终数据规模	11 696	11 571	1 576

$$FP_c(f)=\frac{\Sigma_{d\in c}W_d^f}{\Sigma_{f'\in F_c,d\in c}W_d^{f'}} \quad (2)$$

$$FF_c(f)=2\left(\frac{FR_c(f)\times FP_c(f)}{FR_c(f)+FP_c(f)}\right) \quad (3)$$

式中：W_d^f 表示数据 d 的特征 f 的权重，F_c 表示聚类 c 的所有关联特征；$FP_c(f)$ 表示特征 f 表征聚类 c 的能力度量值，$FR_c(f)$ 表示特征 f 表征聚类 c 区别于其他聚类的能力度量值。$FR_c(f)$ 与尺度无关，$FP_c(f)$ 与尺度有关，实验表明这2个指标的组合，即F指标，受特征尺度的影响很微弱，为了保证F指标度量方法与尺度大小完全无关，需要进一步对数据进行标准化处理[21]。F 指标度量方法适用于各种数据加权方法，但需要正值处理（具有负值的特征会被分配在2个不同的特征子集中），数据特征表达的内容便会明晰。

2.2 特征最大化

特征最大化是一个无偏聚类质量指标，它利用了每个聚类中关联数据的属性（即特征）。特征最大化已被证明可以在监督学习中有效选择特征[21]，本文利用 F 指标对无监督学习的文本聚类过程进行特征提取，然后根据最具有代表性的特征标记聚类标签，整个特征选择过程是一个非参数化的处理过程。

特征集 S_c 是由属于分区 C 的聚类 c 的代表性特征构成：

$$S_c=\left\{f\in F_c|FF_c(f)>\overline{FF}(f),\ \ FF_c(f)>\overline{FF}_D\right\} \quad (4)$$

$$\overline{FF}(f)=\Sigma_{c'\in C}\frac{FF_{c'(f)}}{|C_{ff}|},\quad \overline{FF}_D=\Sigma_{f\in F}\frac{\overline{FF}(f)}{|F|} \quad (5)$$

式中：C_{ff} 表示具有特征 f 的 C 的子集。

最终，所有被选择出来的特征构成了特征集 S_C，它是特征集 F 的子集，即：

$$S_C=\cup_{c\in C}S_c \quad (6)$$

也就是说，被判定给某聚类的关联特征的 F 值既要大于所属聚类的 F 平均值，也要大于所属分区所有特征的平均特征值。

2.3 对比度

将“对比度 $G_c(f)$”定义为将某个特征 f 判定给聚类 c 的可能性，即聚类 c 中保留某个特征的性能，$G_c(f)$ 作为一个指标值与聚类 c 中某个特征的 F 指标 $FF_c(f)$ 和整个分区中该特征的 F 指标平均值 $\overline{FF}$ 的比率成正比关系。它可以表示为：

$$G_c(f)=FF_c(f)/\overline{FF}(f) \quad (7)$$

对比度大于1的特征可以用于表征该聚类，对比度越高，越能表征聚类内容。

通过一个例子来说明这种计算方法。数据集（见图3）有2个类（男性（M）和女性（F））和3个描述特征（鼻子大小（*Nose_Size*）、头发长度（*Hair_Length*）和鞋子尺码（*Shoes_Size*））。计算每个特征在每个类中的 F 指标（见图4），进而计算所有特征的平均值 $\overline{F(.,.)}$ 和某个描述特征 x 在2个类的平均值 $\overline{F(x,.)}$，其中：

$$\overline{F(x,.)}=\Sigma_{f\in S_x}\frac{FF_x(f)}{|S_x|} \quad (8)$$

由于 *Nose_Size* 的 F 值（0.3,0.24）都低于平均值 $\overline{F(.,.)}$（0.38），所以忽略该特征。比较 F 值与边际平

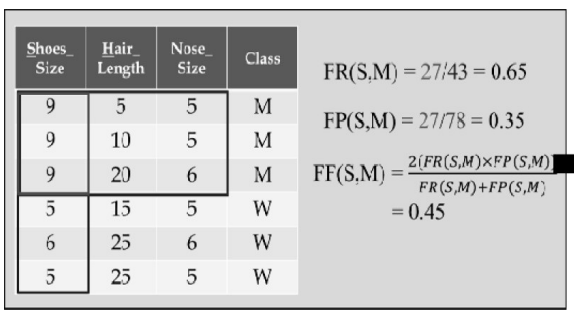


图3　数据集和F指标计算原理

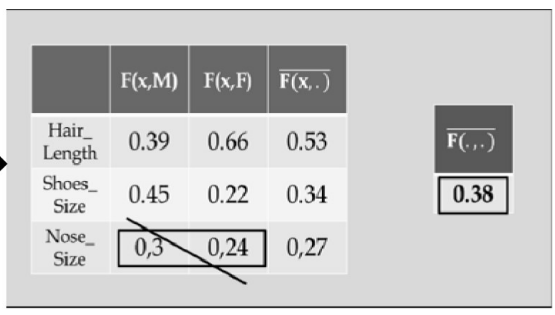


图4　特征选取原理

均值 $\overline{F(x,.)}$，对于男性（*M*），*Shoes_Size* 是代表性特征（0.45>0.34），而对于女性（*W*），特征 *Hair_Length* 具有代表性（0.66>0.53）。

“对比度”是关注被选择特征基于不同类中的边际平均 *F* 值的能动性（activity）和受动性（passivity）。计算过程如图5所示，对比度实际上强化了女性头发长度和男性鞋子尺寸的特征显示，弱化了男性头发长度和女性鞋子尺寸的特征显示。根据特征选择的对比度依据，“头发长度”被选作女性的代表性特征，“鞋子尺寸”被选为男性的代表性特征。

3　数据分析过程

本文基于特征最大化的特征选择算法和高维数据聚类算法，，提取出“中国科学学”的研究主题，其完整的数据处理与分析过程如图6所示。特征最大化方法与聚类方法（如神经网络方法[22]）的适当组合，能够比其他一些方法（如LDA[23]）更有效地提取主题和优化模型[24-25]。本文利用对比度进行聚类可视化，它不仅能解决大数据集交互表征的认知能力超载问题，还可以通过高对比度的共享特征提取显示主题之间的联系。最后，本文利用与聚类相关的数据（论文发表时间和作者）进行聚类外生标签的标注，以为更精准地理解主题变迁提供补充信息。其中，论文发表时间作为主题变迁的重要参考标签，作者信息能更准确地理解研究主题的内涵（鉴于作者数据的不完整，本文没有在文中突出作者的分析，而在对主题的理解中，借助于作者信息）。

3.1　聚类和优化模型检测

K-means 和 GNG 是2种不同的聚类方法[22,26]，前者是基于竞争学习规则（winner-take-all rule）的学习方法，其特点是每个输入点只改变一个质心，后者使用赫布学习规则（Hebbian learning rule）的赢者多取（winner-take-most rule）方法，即不仅要更新赢家

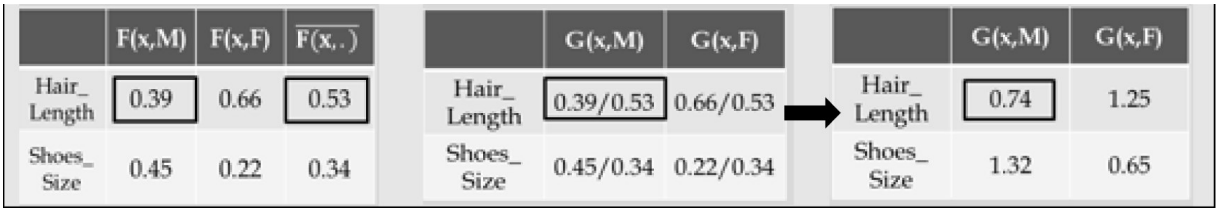


图5　被选择特征的对比度计算原理

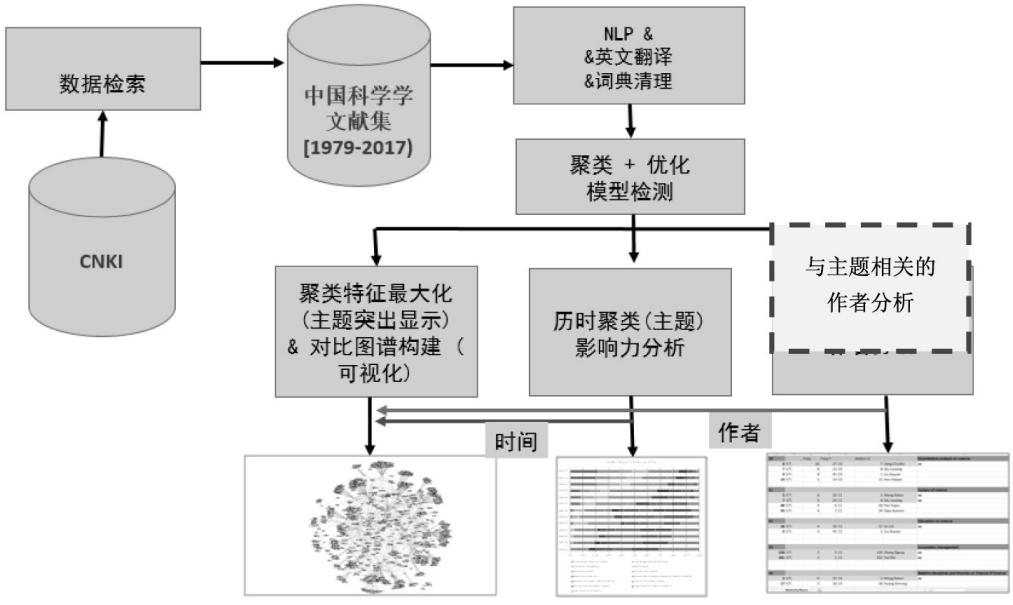


图6　数据分析的完整过程

质心，与赢家质心相邻的所有拓扑近邻点也相应地改变。GNG方法之所以优于K-means方法，是由于它结合了winner-take-most rule的赫布算法而使得学习过程更独立于初始条件，并避免陷入不良的局部最优点，这在许多数据实验中都得到验证[27]。

最优模型的选择取决于特征最大化度量。大多数常用的质量评估器由于对噪声敏感，并不适用于高维数据分析[28]，因而对现实数据分析的结果并不令人满意。因此，本文采用一种更精确的方法，即利用特征最大化和与集群特征的能动性和受动性相关的具体信息来定义聚类质量指标，进而确定最优分区方案。这种分区方法期望公式7描述的对比度最大化，因为特征对比度越大，类内紧凑度越高，类间分离度越大，由此产生*PC*和*EC*这2种不同的指标。*PC*指标，其原理类似于通常模型中的聚类内部惯性，是基于最优分区的能动特征的平均加权对比度最大化的宏观度量指标。*EC*指标，其原理类似于通用模型中类内惯性和类间惯性的组合。是基于能动特征对比度和受动特征的反向对比度的平均加权折衷的最大化。对于包含*K*个聚类的分区，这2个指数分别表示为：

$$PC_k = \arg\max_k \left(\frac{1}{k} \sum_{i=1}^{k} \frac{1}{|s_i|} \sum_{f \in S_i} G_i(f) \right) \quad (9)$$

$$EC_k = \arg\max_k \left[\frac{1}{k} \sum_{i=1}^{k} \left(\frac{|s_i| \sum_{f \in S_i} G_i(f) + |\bar{s}_i| \sum_{h \in S_i} \frac{1}{G_i(h)}}{|s_i| + |\bar{s}_i|} \right) \right] \quad (10)$$

式中：n_i表示聚类i中关联数据的数量；$|s_i|$表示聚类i中能动特征的数量；$|\bar{s}_i|$表示聚类i中受动特征的数量。

综合考虑*PC*和*EC*，本文使用一个组合指标（*CB*）来评估聚类的质量，即选择对应于最高（*PC*+*EC*）指标值（即最优对比度）的*PC*曲线峰值所对应的聚类方案。与通常的质量评估指数不同，如Dunn指数（DU）[29]、Davies Bouldin指数（DB）[30]、轮廓指数（Silhouette index）[31]、Calinski Harabasz指数（CH）或Xie-Beni指数（XB）[32-33]，*CB*指标的主要优势之一在于对任何维度的数据进行分析，都可以生成稳定的结果，而且计算用时少。此外，该指标能够有效地处理其他指标无法处理的二分数据，而且能减少通常聚类结果的嘈杂。实验结果表明，*CB*指标尤其适用于高维文本数据的处理，用来进行验证的文本多是历时文本[25]。

本文分别度量了对应于1-50个聚类的*PC*值和（*PC*+*EC*）值。聚类数量为1的模型被舍弃了，因为单个聚类对于本文没有意义。最终，本文依据最优的*CB*指标选择聚类成13个类的模型（见图7）。这个方法能够优选出某个时期科学学研究主题的数量。依据13个聚类的内容描述特征词给出聚类标签，即科学学近40年来的研究主题（见表2）。

3.2 对比图的绘制

二分图中的节点被分为2个不相交的独立集*U*

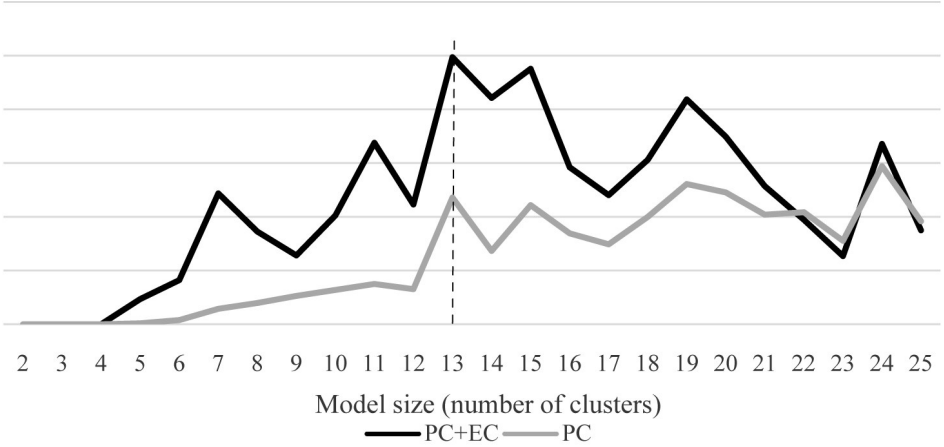


图7 聚类质量评估CB指标原理和优化模型(13个聚类)的选择

和V,一条边将2个集中的节点连接起来。对比图(contrast graph)就是基于特征集S和标签集L之间关系而构建的二分图[34]。理论上,一组标签L应该可以表达关联特征的各种信息,而且特征集F的子集S通过特征选择过程获取。当使用特征最大化方法的时候,边$(u,v), u \in S, v \in L$的权重$c_{(u,v)}$表示标签v的特征u的对比度,它由等式7定义(在公式7中,标签是由聚类的相关数据抽象提炼出来的)。

二元图有以下几点特征。首先,关联特征选择过程删减了连接的数量,从而缓解了图表征所产生的认知超载;第二,当特征词与多个标签连接时,它可以显示标签之间的联系;第三,将该方法与加权导向模型(weighted force-directed model)相结合并可视化[35],能够突出L集中核心的或最有影响力的标签,而与该标签密切相关的特征词也会集聚在标签临近位置。

3.3 补充性外生标签

外生标签(external label)是与数据相关但不影响初始数据分析过程的信息[36],它可以为更精确的内容分析提供重要线索。在聚类基础上,通过补充数据在聚类中的分布来提供相关主题的补充信息。本文主要关注2个外生标签,文章的发表年份和作者。文章的发表年份用于主题变迁研究,突出显示

表2　最优聚类模型的聚类能动特征词与主题标签

聚类	标签	内容(能动特征词)
0#	科学的定量分析	Bibliometrics, citation analysis, journal, indicator, quantity, impact factor, statistics analysis, data, SNA
1#	科研评价	Efficiency, systems engineering, decision making, forcast, evaluation, administration, input and output, efficiency, sustainable development
2#	科学教育与人才培养	Higher education, Ministry of Education, planning, talent cultivation, university
3#	创新管理	Enterprise, knowledge management, collaborative innovation, performance, competitive advantage, technological innovation, integration
4#	科学学相关学科与知识结构	S&T studies, theory of science of science, technology theory, technology philosophy, dialectics of nature, library science, knowledge-based economy, history of science of science, discipline structure
5#	科学学的哲学基础	Philosophy, Marxist doctrine, reality, criticism, ontology, dialectics, human society, materialism, humanism
6#	学科体系	Definition, connotation, discipline system, research method, concept, principle, comparative research, system science, safety, safety principle, safety system
7#	科技政策及科学的社会功能	Scientifilization, S&T development, modern management, productivity, nation, world, emancipation of mind, socialism, social economic development
8#	科学学的学科属性	Natural science, social science, modern science, regular pattern, development principle, edge, interdisciplinary research
9#	科学知识图谱	Research hot topics, software, hotspot, theme, frontier, development trend, knowledge map, data, visualization analysis
10#	科学学的历史	History of science, creator, JD.Bernard, Price, big science, Zhao Hongzhou, scientometrics, Soviet Union, world science, sociology of science
11#	科研产出的出版和管理	Journal, publication, S&T management, S&T system reform, S&T circle, editorial office, institute, S&T policy
12#	科学学组织	Committee, leadership, Chinese Association for Science of Science, conference, symposium, academic exchange, Liu Zeyuan

注:9#科学知识图谱的聚类对比度值最大,该聚类的特征词及其F值为:5.376770 theme, 5.030978 research hot topics, 4.827424 literature, 4.734794 software, 4.697236 frontier, 4.595268 development trend, 4.401170 research topic, 4.342141 hotspot, 4.159228 both at home and abroad, 3.989917 science knowledge mapping, 3.873852 international, 3.801943 expectation, 3.778949 data, 3.721473 knowledge map, 3.648744 visualization analysis, 3.641972 tool, 3.557082 research situation, 3.495185 trend, 3.411639 representative figure, 3.327669 research direction, 3.305491 SCI, 3.195925 field, 3.130080 trajectory, 3.067899 SSCI, 3.047346 multidisciplinary, 3.044982 hotspot field, 2.999260 keyword, 2.971392 webometrics, 2.959471 CSSCI, 2.958249 visualization, 2.930680 clustering, 2.880841 authority, 2.872097 research trends, 2.855751 topic words, 2.828917 database, 2.823275 subject knowledge, 2.779299 distribution, 2.760318 knowledge structure, 2.731361 algorithm, 2.678810 discipline distribution, 2.667470 similarities and differences

各时间区间内的重要主题，这种重要性体现在相对于其他主题的独特性，这有助于更准确地理解研究主题的演变。文章作者用于突出显示在驱动某个研究主题发展中做出重要贡献的作者，尤其是同时驱动和影响若干研究主题发展的作者是很重要的，本文中作者信息是对各时期主要研究主题内容理解的重要辅助信息。

本文的外部标签分析是基于2个不同的指标，即标签频率（label frequency）和标签流行度（label prevalence）。标签l在聚类c中的频率F_c^l定义为：

$$F_c^l = Card\{d \in D\} \; af(d) = c \wedge l \in Extlab_t(d)\} \quad (11)$$

其中$Card$是指基数函数（cardinal function），D是整个数据集，$af(d)$是与某个矩阵关联的数据d的$crisp$函数，$Extlab_t(d)$是与外生标签t关联的数据d的函数。在$crisp$聚类中，每一个数据都是与某个聚类相关，通常用af函数表达：

$$af(d) = \arg\min_k \left(Dist(\vec{k}, \vec{d}) \right) \quad (12)$$

式中：$Dist$表示聚类函数（通常使用欧氏距离），$\vec{k}$表示聚类k的描述向量，$\vec{d}$表示文献d的描述向量。标签流行度（label prevalence）是一个基于聚类的指标。在一个聚类中，如果：

$$'\nexists c' \in C, \; c' \neq c, \; F_{c'}^l > F_c^l \wedge \nexists l' \in L_c, \; l' \neq l, \; F_{c'}^{l'} > F_c^l \quad (13)$$

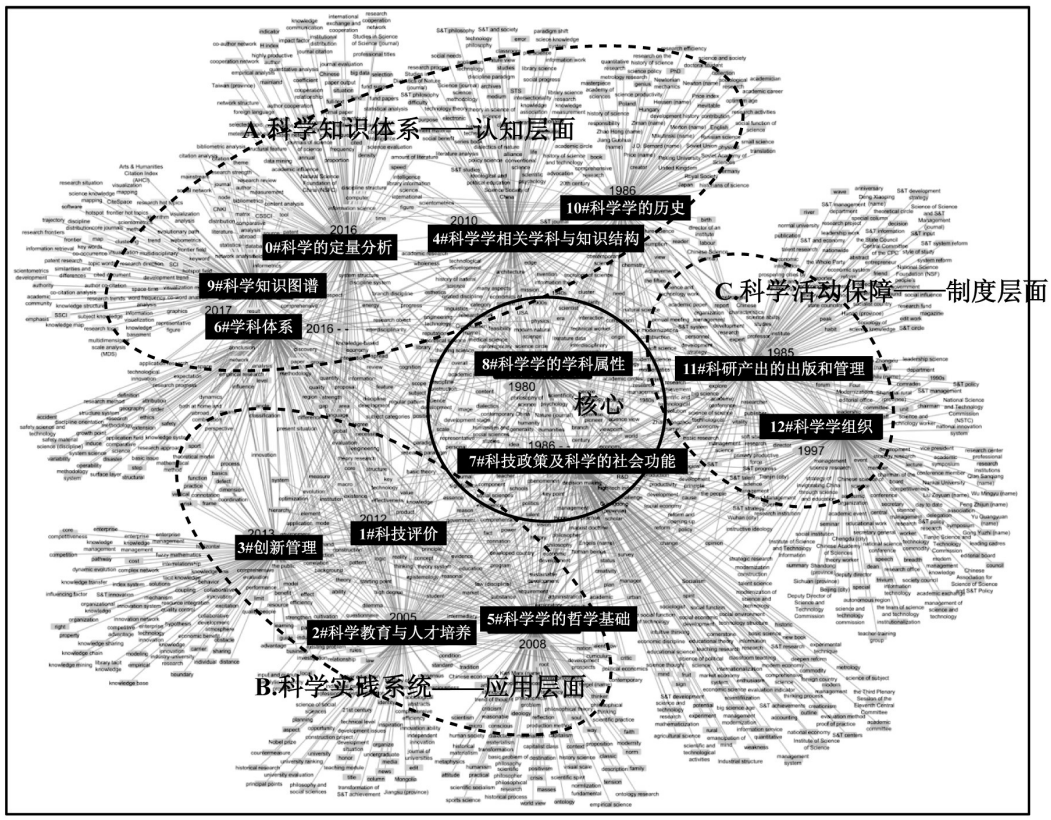
那么标签l就是流行热门的。其中，Lc是聚类c的标签集。流行度是用来表示标签的热门程度。因此，一个标签只能在某个聚类中热门流行，而某些聚类可能没有任何热门标签。

4　数据分析及可视化结果

4.1　科学学研究主题结构

将经特征最大化提取到的特征词以及标签以对比图的构建方法绘制的科学学研究主题结构可视化图谱（见图8，保留了1 074个F指数高于1.5的特征词）。

科学学研究的13个主题在图7中的位置分布表



注：局部放大图见文后附录，文中凸显了9#科学知识图谱，其详细信息见表2

图8　科学学研究主题结构图谱

明其核心内容包含“8#科学学的属性”和“7#科技政策及科学的社会功能”两大主题。科学学是一门反思的学问，因而要直面自身的发展，研究科学学本身的形态及模式；科学学又是一门指导于实践的应用科学，因而要以应用为导向，研究科学的社会功能，服务于科技政策。围绕这2个核心主题，与科学实践活动相一致，科学学研究由三大支撑领域，即“科学知识体系”、“科学实践系统”和“科学活动制度保障”，构成了基于科学的“认知”、“应用”和“制度”3个层面较为完整的逻辑结构体系。“A.科学知识体系”领域从科学的认知层面突出了“10#科学学的历史”、“4#科学学相关学科与知识结构”、“学科体系”、“0#科学的定量研究”和“9#科学知识图谱”五大关联主题。科学学的相关学科和知识结构都蕴含于早期科学学研究之中，大多是从科学技术发展的历史脉络中总结提炼出来的；定量分析是一种重要的科学研究方法，随着科学数据的日益完善和信息可视化技术的发展，科学知识图谱也成为分析科学知识体系的重要方法；作为在整个科学知识体系分类总较为独立并更具有教育规划实践指导意义的是学科体系研究，在这里值得一提的是中南大学的吴超教授咋科学学范畴内较为系统地研究了“安全科学”及相关学科的学科体系研究。“B.科学实践系统”领域从科学的应用层面突出了“1#科技评价”、“3#创新管理”、“2#科学教育与人才培养”和“5#科学学的哲学基础”四大关联主题。科学技术对社会实践的作用是通过创新管理来实现的，随着创新驱动战略的实施，涉及到科学的投入产出评估、科学决策、系统工程等科技评价也日益成为科学实践系统中不可缺少的重要议题；科学实践系统中的主体是科技人才，因而科学教育与人才培养是不容忽视的，而指导整个科学实践系统的根本思想是源于科学学的哲学基础，即马克思主义哲学和恩格斯的自然辩证法。“C.科学活动保障”领域从制度层面突出了“11#科研产出管理与出版”和“12#科学学组织”2个主题，主要是关于期刊和学会的相关建制研究。

4.2 科学学研究主题演变路径

中国科学学研究近40年的13个研究主题，其历史变迁路径如图9所示。在1980年代，科学学在中国刚刚兴起，学术界讨论最多的就是关于科学学学科属性的问题；随着1978年全国科学大会的召开，中国科技体制开始进入改革期，科学学领域的三大期刊，《科研管理》（1978年）、《科学学与科学技术管理》（1980年）、《科学学研究》（1983年）应运逐一创建；随着期刊的发展和科研产出的增加，在1985年代，关于论文的出版和管理成为热门话题；紧接着学术界对深入研究科学学的历史（包括贝尔纳和普赖斯的相关研究）产生浓厚兴趣，尤其是贝尔纳的“科学的社会功能”成为科技政策研究的重要理论依据；接下来，中国科学学经历了一段低迷期，直到1997年，中国科学学与科技政策研究会第三届理事会成立（冯之浚为理事长），这一年刘则渊教授在大连理工大学管理科学与工程博士点下招收科学学方向的博士研究生，从论文发表情况看，科学学研究的组织建制成为这一时期的热门话题；2005年，科学教育与人才培养成为热门话题；2008年，学术界集中在科学学的哲学根源探究，即马克思主义哲学基础；2010年，科学学的相关学科和知识结构成为研究热点，即科学技术学、科学学理论、科学技术史、图书馆科学、自然辩证法、科技哲学、政治与思想教育等都是突显出来的词汇；2012年，学界强调科学活动是一个系统，需要靠系统工程的办法进行评价与管理以提升效率；紧接着，创新管理便成为突出的热门话题；2016年，科学的定量分析和学科体系成为热门；2017年，科学知识图谱的相关研究已成为科学学研究的一个主流话题。

主题聚类中发表论文数量随时间的历年变化趋势（见图10）突出显示了特定研究主题的活跃周期，如趋势线增长表征出新兴主题（0#、9#、1#、3#、6#），趋势线呈峰值状态表征相应时期的热点主题。

科学学的定量研究、科学知识图谱、科技评价、创新管理、学科体系是近些年的新兴主题，也是当今的热门话题。

在一个学科领域，某个研究主题是否成为热点具有一定的偶然性，但在一个较大历史时间尺度里的消长趋势中却蕴含着必然性。依据13个主题每3年发表论文数量的比例，可以看出13个主题在近40年中的流行变化趋势（见图11）。其中“科学的定量分析”、“科学知识图谱”和“创新管理”主题都是在科学学初期没有出现的话题，但近些年来在科学学研究中的地位越来越重要，“定量分析”的主导地位的确立，说明科学学作为一门学科已走向成熟，“科学知识图谱”主题的突出，表明科学学是一门开放的学科，它较好地把计算科学和信息可视化技术结合进来，“创新管理”研究的兴盛表明，科学学是一门与实践紧密结合的学科，它强调科学技术的经济价值，并显示出在当今中国的战略性地位。相比较而言，关于“科学学学科属性”、“科学学的组织建构”、“科研产出的出版和管理”的研究话题热度逐渐减弱，这也标志着科学学研究在中国正逐渐走向成熟和规范。

5　结论与启示

科学学作为一门以实践为导向的基础理论研究，在中国是与国家的改革开放相伴而生的，本文利用数据分析的手段客观地揭示中国科学学研究主题的历史变迁，借以反映其在国家发展过程中发挥的重要作用，同时，中国经济的快速发展与创新实践的日益活跃，也为科学学研究提出了更多的研究话题。

科学学研究在近40年逐渐走向成熟。中国科

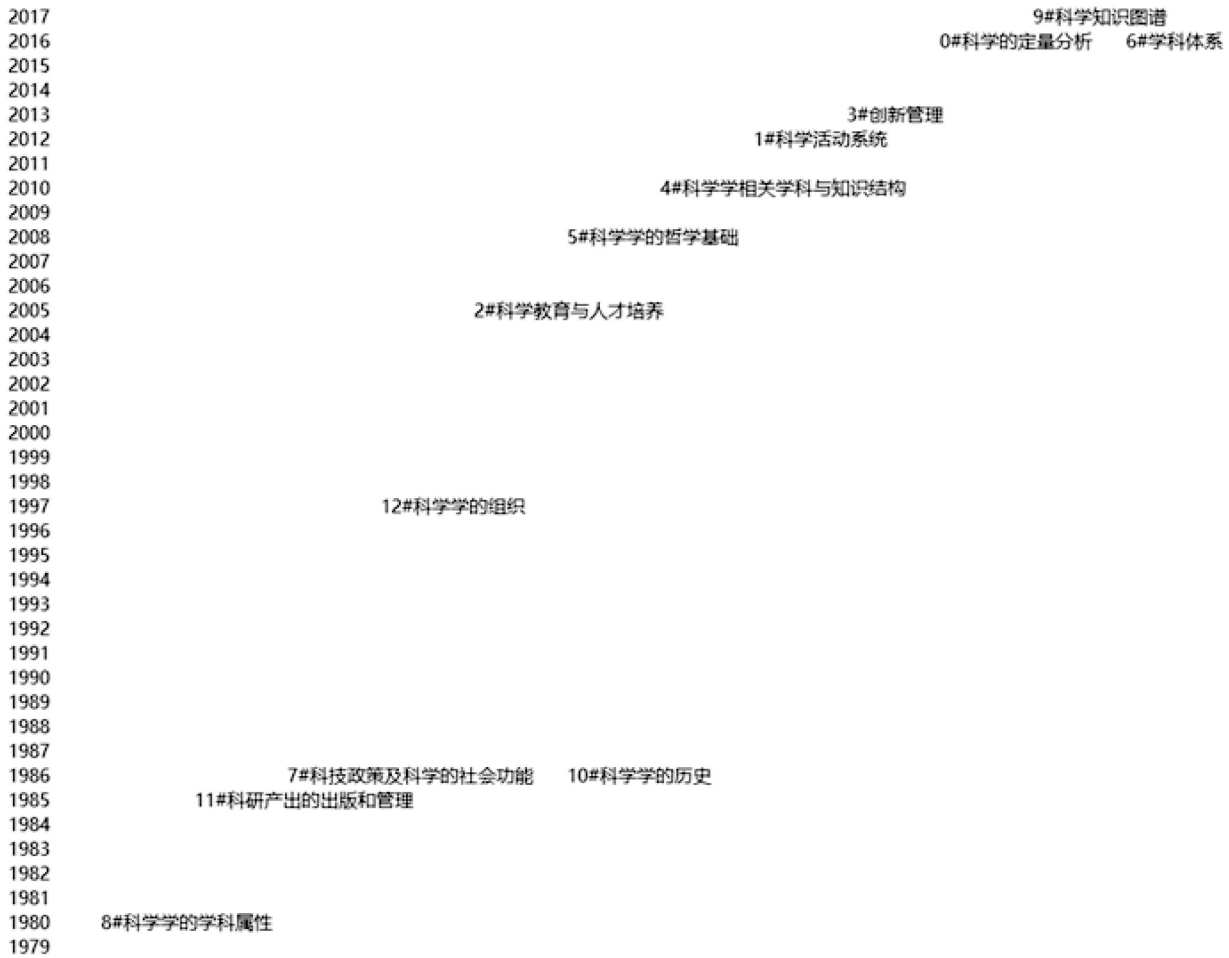


图9　中国科学学研究主题变迁路径图

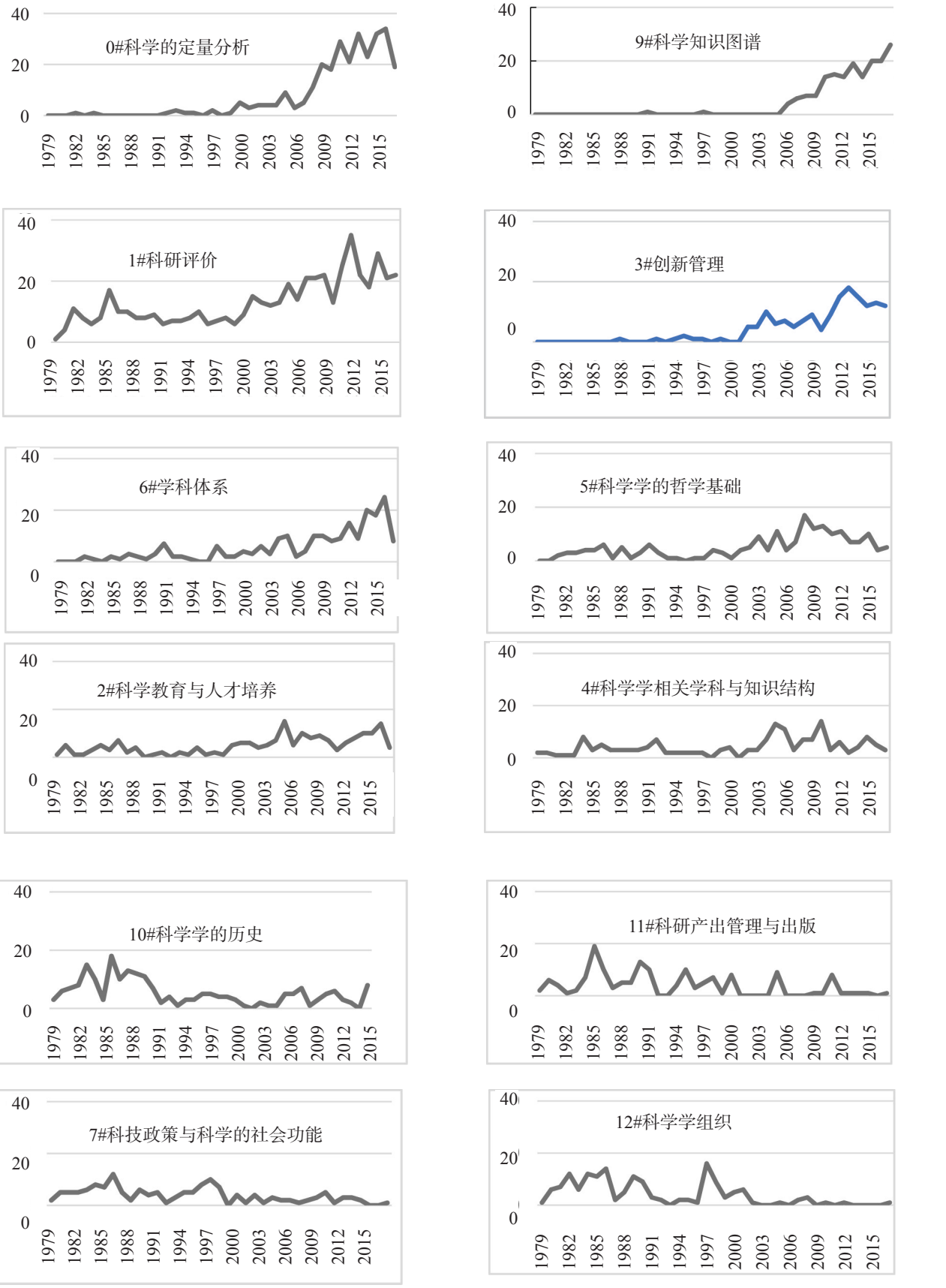
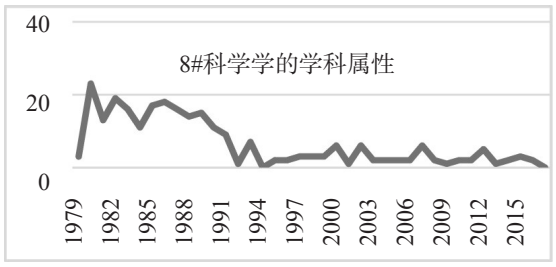


图10　13个科学学研究主题的活跃周期



科学学的定量研究、科学知识图谱、科技评价、创新管理、学科体系是近些年的新兴主题，也是当今的热门话题。

图10　13个科学学研究主题的活跃周期

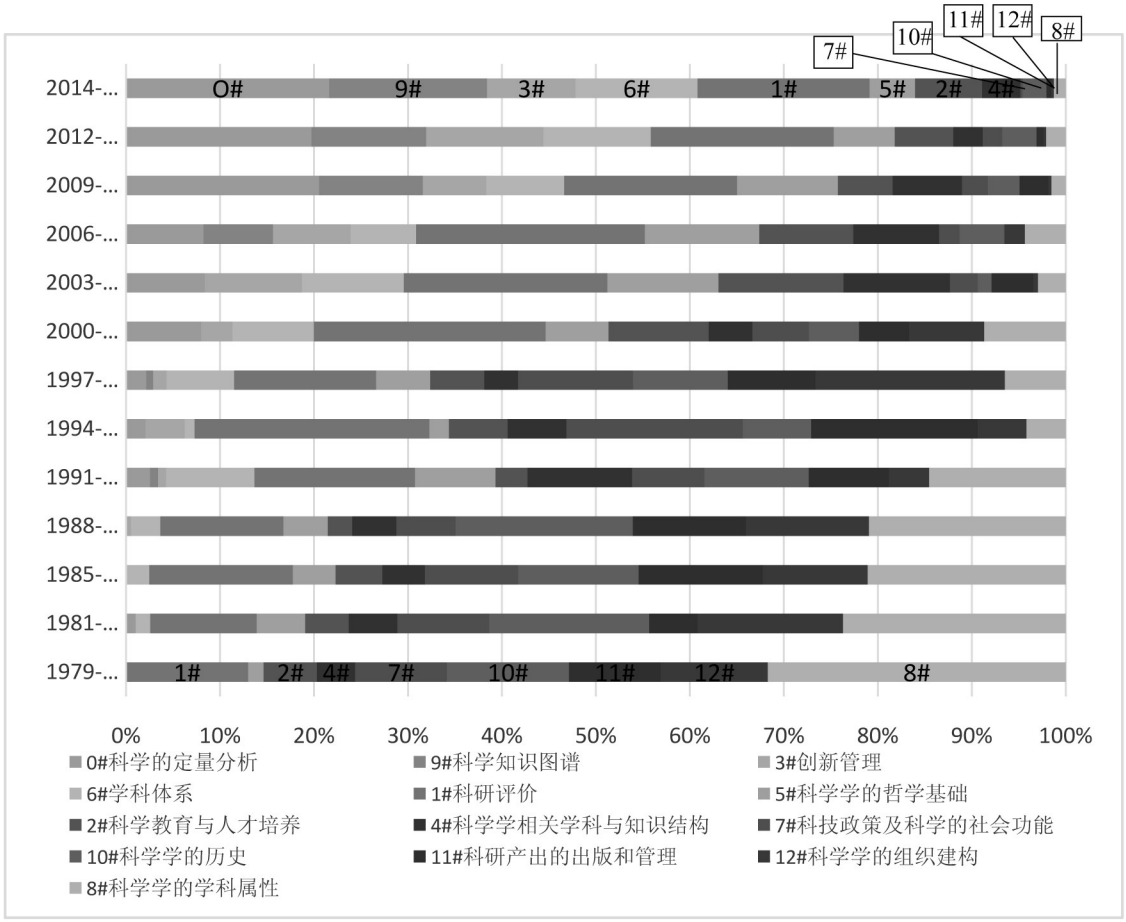


图11　科学学研究主题的流行趋势

学学是一门“外引内生”学科，即最早是从介绍并学习贝尔纳的科学学思想开始，并将其大线条地应用于中国的科技实践，而后随着定量分析及数据挖掘技术的介入与强调，科学学研究因量化基础而使得理论建构越发扎实，科学技术的社会功能也随之从创新的视角得以更深入的解构。具体表现为其研究从学科一般属性探讨转向相关学科与知识结构分析，从定性分析转向偏重于定量分析和可视化分析，从科学的一般社会功能研究转向更为具体的经济功能和战略功能研究。

特征最大化方法与无监督学习方法的结合，可以有效地揭示研究领域的主题变迁。本文提出一种*F*指标特征最大化和无监督学习聚类相结合的研究领域主题分析及可视化方法，它较之于以往基于文献共被引或主题词共现的聚类方法，以及LDA主题提取的方法，更适用于高维度的大规模数据分析。

这种方法在对中国科学学主题变迁的分析中，客观地反映了主题变化过程，同时这种变化也符合中国科学学的发展规律。

中国科学学的发展历经40年已经得到了长足的发展，但仍需加强基础理论研究。面对科学技术的迅猛发展和战略地位的提升，“科学和技术的本质”，“科学和技术的发展规律”，“科学技术知识体系”、“科学技术活动系统”、“创新的链条”等这些基本理论问题显得尤为重要。随着大数据技术的发展，科学大数据犹如一个宝藏，它亟待于用更先进的技术手段去进行科学技术发展的规律性研究，并将这些规律性研究应用于科学技术实践活动，指导科学技术正向地发挥杠杆作用。

参考文献

[1] 陈悦,张立伟,刘则渊. 世界科学学的序曲:波兰学者对科学学的重要贡献[J]. 科学学研究,2017,35(1):4-10.

[2] 赵红州,蒋国华. 格森事件与科学学的起源[J]. 科学学研究,1988,6(1):14-23.

[3] 赵红州,蒋国华. 伟大的事实,巨大的课题[J]. 科学学与科学技术管理,1983,卷首语.

[4] Bernal J D. The Social Function of Science[M]. London: George Routledge & Sons Ltd., 1939.

[5] 潜伟,李欣欣. 贝尔纳与中国[J]. 科学文化评论,2012,9(6):16-32.

[6] 钱学森. 现代科学技术[J]. 人民日报,1977-12-09.

[7] 钱学森. 科学学、科学技术体系学、马克思主义哲学[J]. 哲学研究,1979(1):20-27.

[8] 钱学森. 关于建立和发展马克思主义的科学学的问题:为《科研管理》创刊而作[J]. 科研管理,1980(1):3-8.

[9] 刘则渊,陈悦,朱晓宇. 普赖斯对科学学理论的贡献[J]. 科学学研究,2013,31(12):1762-1772.

[10] 浦根祥,狄仁昆. 科学社会学的认知转向[J]. 自然辩证法通讯,1998(5):29-34.

[11] Zeng A, Shen Z, Zhou J, et al. The science of science: From the perspective of complex systems[J]. Physics Reports, 2017(714/715):1-73.

[12] Fortunato S, Bergstrom C T, Borner K, et al. Science of science[J]. Science, 2018,359(6379):1-7.

[13] 和钰,陈悦,崔银河,刘则渊. 科学学的研究进路暨前瞻:基于贝尔纳奖的分析视角[J]. 科学学研究,2017,35(8):1121-1129.

[14] 王续琨. 科学学:过去、现在和未来[J]. 科学学研究,2000,18(2):19-23.

[15] 刘则渊,陈超美,侯海燕. 迈向科学学大变革的时代[J]. 科学学与科学技术管理,2009,30(7):5-12.

[16] 陈士俊. 科学学:对象解析、学科属性与研究方法:关于科学学若干基本问题的思考[J]. 科学学与科学技术管理,2010,31(5):28-35.

[17] 刘则渊. 论钱学森的科学学思想[J]. 科学学研究,2012,30(1):5-13.

[18] 冯之浚. 迎接中国科学学发展的新局面[J]. 科学学研究,1997,19(2):100-106.

[19] 刘则渊. 冯之浚之问:科学学的核心理论是什么?[J]. 科学学研究,2017,35(5):655-660.

[20] 刘则渊,胡志刚,王贤文. 30年中国科学学历程的知识图谱展现[J]. 科学学与科学技术管理,2010,31(5):17-23.

[21] Lamirel J C, Cuxac P, Chivukula A S, et al. Optimizing text classification through efficient feature selection based on quality metric[J]. Journal of Intelligent Information Systems, 2015,45(3):379-396.

[22] Fritzke B. A growing neural gas network learns topologies[C]. Denver: NIPS'94 Proceedings of the 7th International Conference on Neural Information Processing Systems, 1995.

[23] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(1):993-1022.

[24] Lamirel J C, Dugue N, Cuxac P. Performing and visualizing temporal analysis of large text data issued for open sources: Past and future methods // Kozielski S, Mrozek D, Kasprowski P, et al. Beyond Databases, Architectures and Structures: Advanced Technologies for Data Mining and Knowledge Discovery [M]. London: Springer, 2015.

[25] Lamirel J C, Dugue N, Cuxac P. New efficient clus-

tering quality indexes[C]. Vancouver: Neural Networks (IJCNN), 2016 International Joint Conference on IEEE, 2016.

[26] Macqueen J. Some methods for classification and analysis of multivariate observations[C]. Oakland: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967.

[27] Lamirel J C, Mall R, Cuxac P, et al. Variations to incremental growing neural gas algorithm based on label maximization[C]. San Jose: Neural Networks (IJCNN), the 2011 International Joint Conference on IEEE, 2011.

[28] Kassab R, Lamirel J C. Feature-based cluster validation for high-dimensional data[C]. Anaheim: Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications, ACTA Press, 2008.

[29] Dunn J C. Well-separated clusters and optimal fuzzy partitions[J]. Journal of Cybernetics, Taylor & Francis, 1974,4(1):95-104.

[30] Davies D L, Bouldin D W. A cluster separation measure[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE, 1979(2):224-227.

[31] Rousseeuw P J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational and Applied Mathematics, Elsevier, 1987(20):53-65.

[32] Calinski T, Harabasz J. A dendrite method for cluster analysis[J]. Communications in Statistics-Theory and Methods, Taylor & Francis, 1974,3(1):1-27.

[33] Xie X L, Beni G. A validity measure for fuzzy clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991,13(8):841-847.

[34] Cuxac P, Lamirel J C. Analysis of evolutions and interactions between science fields: The cooperation between feature selection and graph representation[C]. Tartu: 14th COLLNET Meeting, 2013.

[35] Kobourov S G. Spring embedders and force directed graph drawing algorithms[J]. Computer Science, 2012(1): 1-23.

[36] Attik M, Al Shehabi S, Lamirel J C. Clustering quality measures for data samples with multiple labels[C]. Innsbruck: IASTED International Conference on Artificial on Databases and Applications, 2006.

An Overview on 40 Years Science of Science Research Topic Evolution in China: A Novel Approach Based on Clustering and Feature Maximization

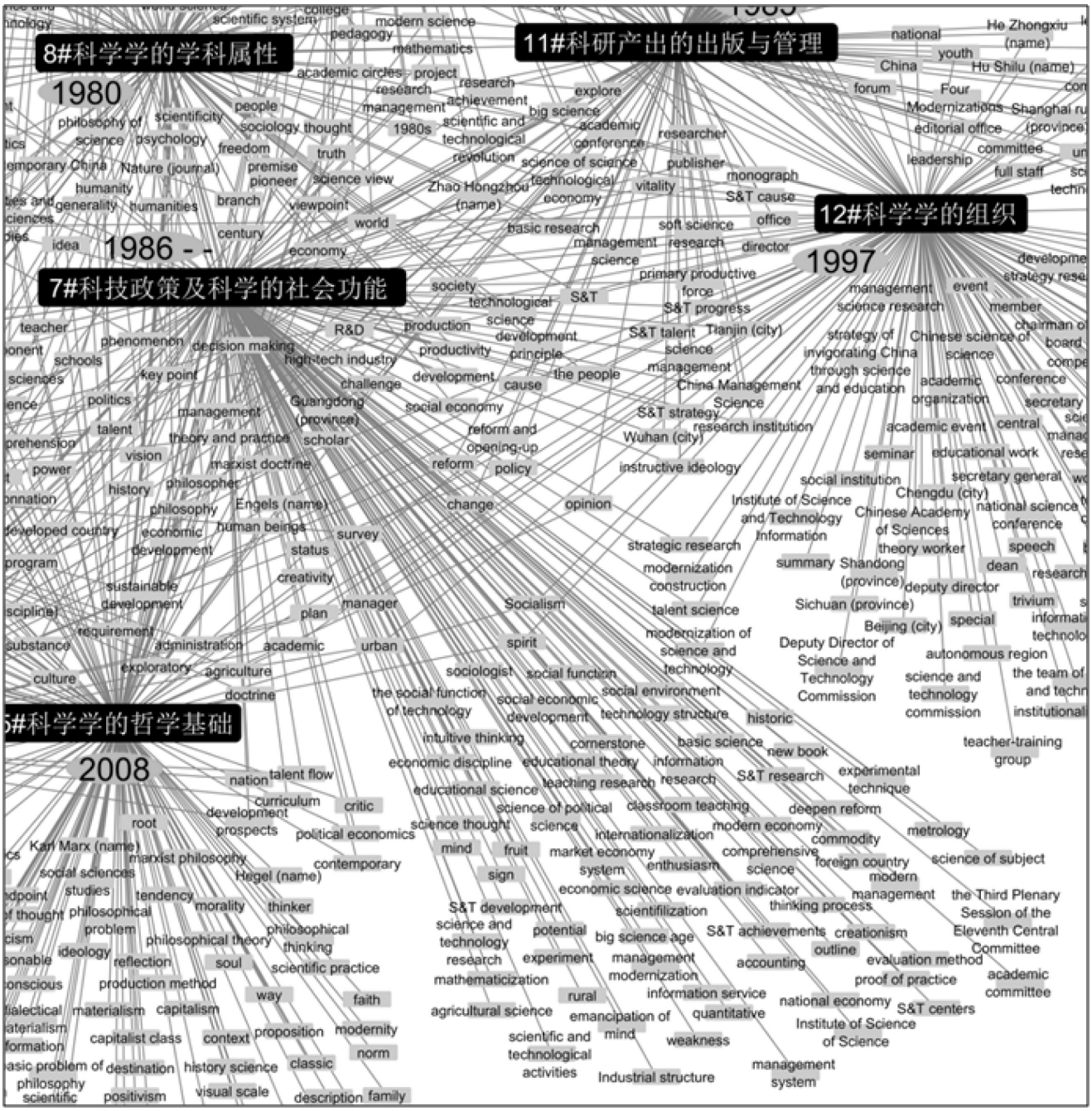
CHEN Yue[1], Jean-Charles Lamirel[1,2], LIU Zeyuan[1]

(1. Institute of Science of Science and S&T Management & WISE Lab, Dalian University of Technology, Dalian 116085, China;2. Synalp-Team-LORIA, University of Strasbourg, Strasbourg 67000, France)

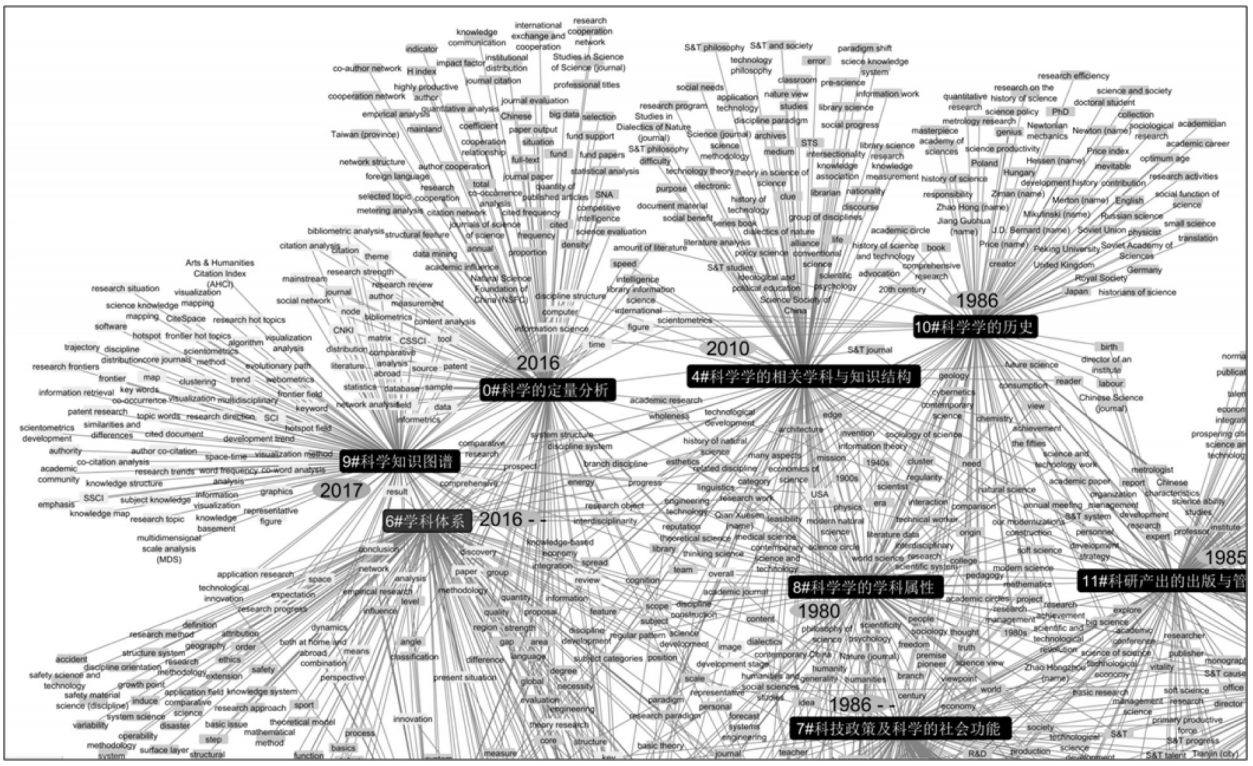
Abstract: Based on the unsupervised combination of GNG clustering with feature maximization, this paper analyses the contents of the academic journal papers in Science of Science in China, and constructs the map of the research topic structure in the last 40 years. Furthermore, it highlights the topic evolution by the exploitation of the publication time and makes use of the author's information for the sake of clarifying topics content. The obtained results interestingly show that the Chinese Science of Science has gradually become mature in the last 40 years, turning from the general nature of the discipline to the relative disciplines and knowledge structure analysis, from the qualitative analysis to the quantitative and visual analysis, and from the general social function research of science to more specific economic function and strategic function studies.

Key words: Science of Science in China; topic evolution; feature maximization; unsupervised learning

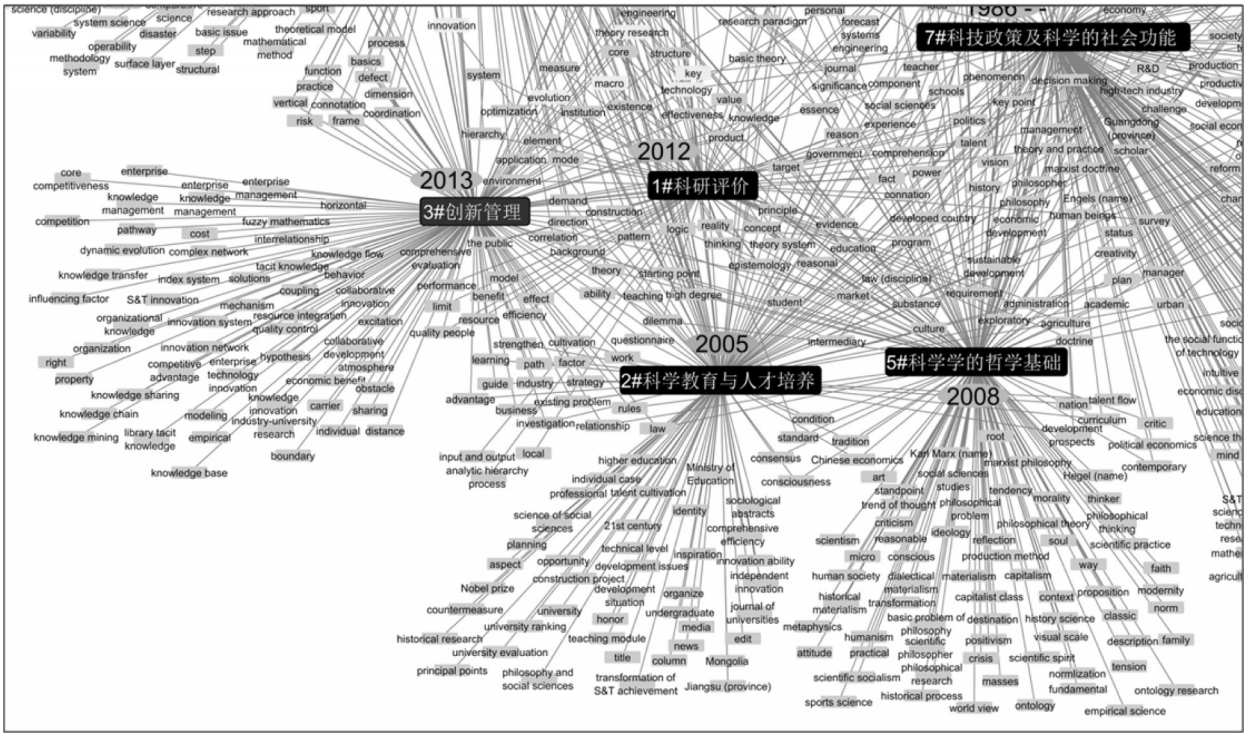
附录：



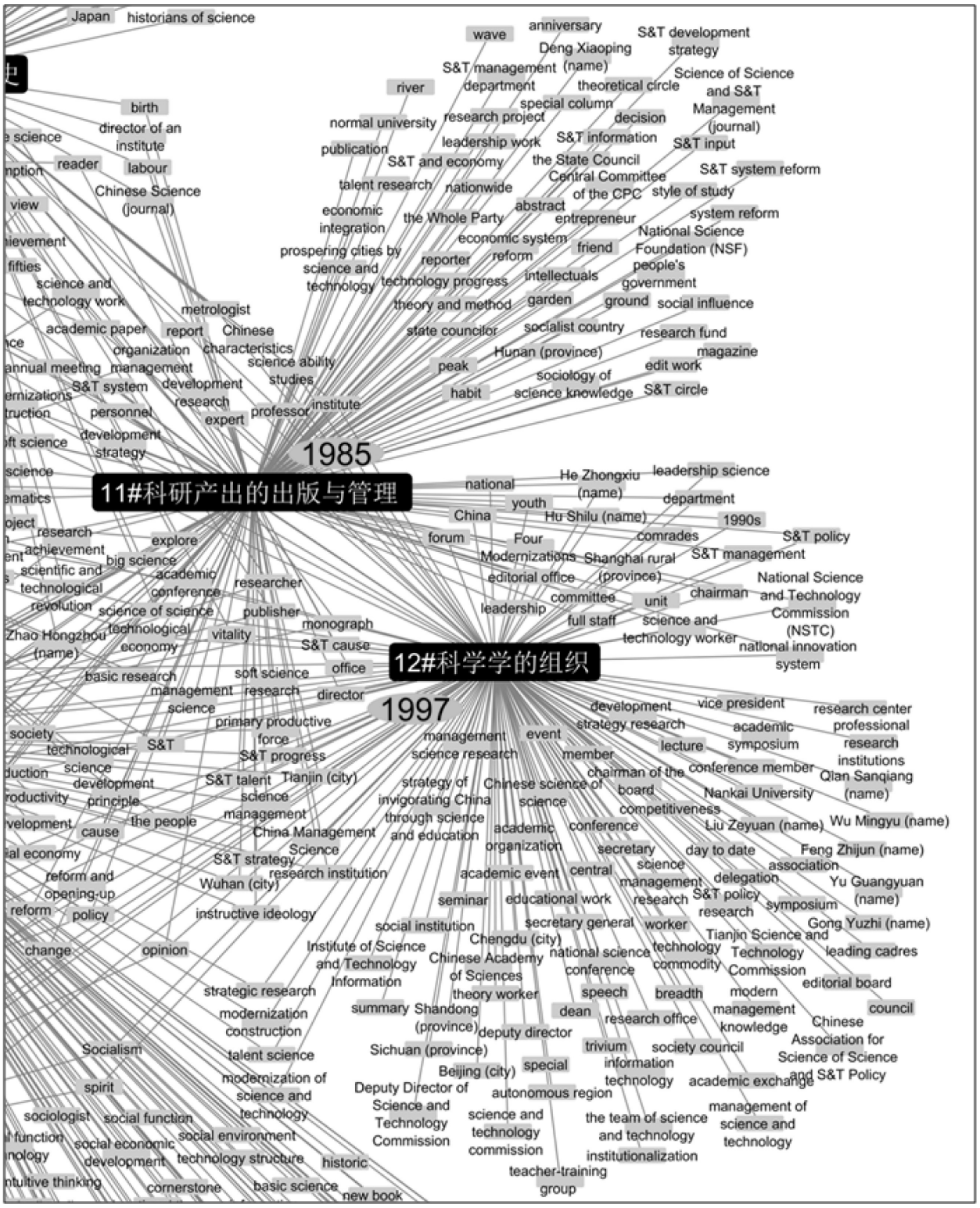
科学学研究主题结构图谱核心部分



科学学研究的“科学知识体系——认知层面”局部放大图



科学学研究的“科学活动系统——实践层面”局部放大图



科学学研究的“科学活动保障——制度层面”局部放大图