DOI：10.13998/j.cnki.issn1002-1248.2019.12.16-1097

基于 t-SNE 算法的双一流大学基金立项关键词降维的可视化建模研究

曹　祺

(珠海市横琴新区灰砚信息科学研究院，珠海　519030)

摘　要：[目的 / 意义]国家自然科学基金的立项资助是科研能力的重要体现，分析双一流大学的科研基金的立项数据有助于为大学建设提供战略支持。[方法 / 过程]研究国家自然科学基金委员会在 1998 年—2017 年资助项目的关键词数据，先对双一流大学进行预处理，然后利用 MATLAB 中 t-SNE 算法对结果进行数据降维和可视化。从时间维度和依托单位维度进行建模，研究过去 20 年内，双一流大学项目的关键词分布。[结果 / 结论]方法比传统基于结构化分析的方法更直观，为大学建设战略制定的提供参考。另外，相关学者也可以在笔者研究基础上，进一步建模和编程，尝试例如进行交互式的可视化建模，对海量项目数据进行快速定位，以提高科研效率。

关键词：科研基金；数据挖掘；科技情报分析；网络可视化；t-SNE

中图分类号：G350　　**文献标识码：**A　　**文章编号：**1002-1248（2020）02-0047-11

引用本文：曹祺. 基于 t-SNE 算法的双一流大学基金立项关键词降维的可视化建模研究[J]. 农业图书情报学报, 2020, 32(2): 47-57.

Visual Modeling of Keyword Dimension Reduction in Double First-Class University Funds Based on t-SNE Algorithm

CAO Qi

(Greysh Academy of Information Sciences, Hengqin New Area, Zhuhai City, Zhuhai 519030)

Abstract: [Purpose/Significance] The National Natural Science Foundation's project funding is an important indicator of scientific research capabilities. Analysis of the data of the establishment of the research funds of double first-class universities is helpful to provide strategic support for university construction. [Purpose/Significance] This article studies the keyword data of the National Natural Science Foundation of China from 1998 to 2017. At first we

收稿日期：2019-12-16

作者简介：曹祺（1988-），男，ORCID：0000-0001-6337-345，武汉大学管理学博士，中国科学技术信息研究所情报学博士后，副研究馆员，高级工程师，研究方向：情报科学。

DOI：10.13998/j.cnki.issn1002-1248.2019.12.16-1097

preprocess double first-class universities' data, and then use the t-SNE algorithm in MATLAB to reduce the dimension of the data and visualize the results. This paper models from the time dimension and the unit-dependent dimension, and studies the keyword distribution of double first-class universities' projects in the past 20 years. [Results/Conclusions] The method in this paper is more intuitive than the traditional method based on structured analysis and provides a reference for the formulation of Chinese universities' construction strategies. In addition, other scholars can further model and program based on this research for such purposes as interactive visual modeling and fast and positioning of massive project data to improve scientific research efficiency.

Keywords: research fund; data mining; scientific and technological intelligence analysis; network visualization; t-SNE

1 引言

1998 年 5 月 4 日，为了响应“为了实现现代化，中国要有若干所具有世界先进水平的一流大学”的号召，国务院批转了教育部《面向 21 世纪教育振兴行动计划》,标志着中国高校高水平建设的“985 工程”正式启动。2017 年 9 月 21 日，教育部、财政部、国家发展改革委联合发布《关于公布世界一流大学和一流学科建设高校及建设学科名单的通知》，标志着双一流大学建设的正式启动，双一流大学分为 A 类和 B 类，在之前 985 大学的基础上增加了新疆大学、云南大学和郑州大学形成了 36 所 A 类双一流和东北大学、湖南大学、西北农林科技大学 B 类双一流。

回溯 1998 年—2017 年中国高水平大学建设成果，科研能力是高水平大学建设的重要评价指标。而国家自然基金委员会（以下简称“基金委”）资助的国家自然科学基金（以下简称“自科基金”）是衡量一个大学的科研能力的重要标尺之一，系统研究自科基金的资助情况和双一流大学建设之间的关联关系有着重要理论意义和实践价值：微观层面，如对双一流大学来说，分析自科基金资助情况有助于大学本身发现自身优势并确定重点科研方向，服务优势学科和行业；宏观层面有助于国家行政管理部门对双一流大学的科研情况有更清晰的全景认识，更好地进行科学决策，提高科研经费投入的产出效益。

鉴于此研究意义，有国内外学者就国家自科基金可视化与双一流大学之间的关联关系，进行了研究，主要分为两类：

（1）研究国家自科基金可视化：该类别主要是研究国家自科基金资助后，发表的论文的可视化[1]、自科基金的评价指标体系[2]和资助后效果评价[3,4]，并且更多是基于美国的国家自然科学基金。如邓方基于关键词统计分析了过去 30 年自动化学科基金的分布[5]，陈挺基于 t-SNE 对美国自然科学基金（NSF）的正文数据进行了分析和可视化[6]，Mejia 分析了基金赞助后论文的引用量用来分析基金赞助的效果[7]。可以对出版物的情况来预测科研基金资助情况[8]，他们的主要研究是处理这些基金数据的方法，如先对大量的基金数据进行建模然后进行可视化分析，先基于关键词的组合来提取主题后分析[9]，然后基于深度学习的方法来提取主题[10]，或可针对元数据进行网络可视化分析[11]。在处理关键词的时候，一般可以采用 R 编程语言或者 MATLAB，如通过 R 分析期刊关键词，来判断文章之间共现关系[12]，也可以通过 R 编程语言进行全文分析来研究科研产出[13]。这些利用 R 编程语言基于关键词的分析（如标题关键词），可以用来衡量科研基金及财政政策资助的有效性[14]。

（2）研究国家自科基金与双一流大学的关联关系：该类别主要研究时双一流大学申报国家自科的情况，如张然研究了“双一流”背景下加强国家自然科学基金组织申报工作探讨[15]。张品慧研究了科学基金对“双一流”建设学科的前期资助研究（2012—2016）[16]，马晓萌研究了双一流高校自然科学基金面上项目资助

特点探析[17]。

因此，目前行业对自科基金的关键词的量化研究较少。同时自然科学基金的关键词专业术语很多，立项数量很大，导致利用计算机建模时矩阵维度很大（稀疏矩阵），不利于可视化，不利于分析其中的规律。笔者基于情报学的理论方法，从数据角度，提出了新的研究思路，以期望对研究问题的提供更好的信息和更高的参考价值。

2 数据来源及研究方法

2.1 数据来源

研究对象为 36 所双一流 A 类大学，由于国防科技大学属于军事涉密单位，笔者不进行研究，数据来源是 A 类大学中除国防科技大学外的剩余 35 所，同时笔者研究的实验数据来自 1998 年 1 月 1 日—2017 年 12 月 31 日 20 年的自科基金数据，基本涵盖了教育部 985 计划实施以来 20 年和双一流计划开始的所有数据。

研究方法主要包含两个模块，即数据源管理模块和数据分析模块。数据源管理模块的主要作用是对公开获取的自科基金数据进行数据清洗，数据分类等预处理，并生成可供 MATLAB 分析的数据源。数据分析模块是对数据源管理模块提供的数据进行数据降维和可视化分析，以期望能更清晰更直观的寻找双一流大学和自科基金之间的关联关系。

实验采用的技术手段为 Java 编程语言和 MATLAB，开发工具主要为 JDK 1.8 和 MATLAB R2018b，为便于其他研究者验证，生成的实验数据和实验代码可以从 Github[18]地址处免费获取。

2.2 常见数据处理方法

由于从基金委的网站上获得的数据存在大量噪音，需要进行数据清洗及分类等预处理。

对于结构化数据，如立项日期、立项编号等，进行直接归类存储，分析的时候利用数据字典将内容映射成对应的整型变量；对于结构化数据集合，将其转化为对应的整型变量后并构造成向量数据；多个向量数据作为一组数据时，将该组数据转化为对应的二维矩阵，如数据不止一组，则转化为高维矩阵。

结构化数据的优势在于基金委已经做了大量的数据规范工作，一般可以直接提供给程序分析使用，如趋势研究，数据筛选和演化网络分析等。但是结构化数据的劣势在于数据之间的关系往往是数据首次存储时已经定好数据范式和关联关系，这类数据一般只能进行简单的数据统计研究，如科技成果统计，却很难进行有效的数据挖掘。

非结构化数据的优势刚好弥补结构化数据的劣势，非结构化数据往往是具备语义属性的，很难直接提供给程序进行分析使用。如果需要转化为程序可以处理的矩阵数据，必须先定义好映射规则。不同的研究人员站在不同的研究角度上，会提出不同的映射规则，解决不同的研究问题。

基于传统的自然语言处理技术是对非结构化数据处理的常见方法，其主要原理都是基于词袋模型（Bag of Words），词袋模型的核心思想将文本视作一系列单词构成的向量空间。利用词袋模型计算不同文本的相似度，一般是先用分词工具将文本分词，分词的时候先过滤掉停用词，然后利用分词工具的主题词库对文本进行文本向量化，这样一个文本会基于主题词库生成一个文本向量。然后计算两个的文本相似度的问题则转化为计算两个向量的夹角的余弦值，也就是余弦相似度，但是词袋模型中的难度在于分词问题和向量维度问题。

2.2.1 分词问题的处理策略

基于通用分词工具的词库，往往对特定领域的文本分词效果差，这是由于不同语言环境中的上下文语境的适用性导致的。语言本身的复杂导致很难建立一个普遍适用的词库，如词语在语义上存在上位词，下位词的分类问题；如词语在语法上存在歧义问题，存在同义词反义词分类的问题。如果消除歧义考虑文本上下文需要进行词语的推断处理，如基于二元词（2-gram），三元词（3-gram）和词组句群的推断处理。更复杂的问题是存在词语跨语言的固定表达，这一点

DOI：10.13998/j.cnki.issn1002-1248.2019.12.16-1097

在科研文献中尤为突出，如“CRISPR/Cas9 基因编辑”往往是固定表达，但是“CRISPR”属于英语，基因编辑属于汉语。

尽管常见的词库都是采用的基于贪婪模式的最大匹配算法，但是也存在不同科研文献中的正向最大匹配和逆向最大匹配适用性问题。如果不基于通用词库，则需要自己建立主题词库，基于人工阅读建立主题词库的效果最好，但是人工成本花费很大。如果基于机器学习的方法建立主题词库，如采用 word2vec 模型[19]，但是训练的样本数量，训练模型，迭代次数的优化则是一个很复杂的研究问题。因此很难有一个绝对适用的词库，需要基于不同的场景设计不同的策略。

2.2.2　维数灾难问题处理策略

如果希望利用词库对文本进行向量化时尽可能减少信息的损失，则应该尽可能减少词语的合并。但是这样会带来维数灾难问题（Curse of Dimensionality）。这一点在汉语中表现的尤为严重，对于汉语语言，本来缺乏类似英文的空格分隔符，汉语的分词复杂性导致词库合并时很难有一个标准解，带来的文本矢量化之后的矩阵过于稀疏。稀疏矩阵的最大问题在于分析时的计算量为指数级上升，同时矩阵进行计算时，相关数据要加载到内存中，内存中如果驻留大量的数据导致计算机运行变慢。并且内存数据和硬盘数据不同，内存计算时的局部性原理，矩阵数据的不可分割问题导致内存数据很难像硬盘数据的分布式存储和分布式计算。

对于内存计算时的维数灾难问题，主流的方法就是数据降维。数据降维有多种方法，一般利用词频逆文本频率算法（Term Frequency - Inverse Document Frequency，TDIDF）建立向量空间模型（Vector Space Model，VSM）[20]。即 TFIDF 算法对文本数据进行词频统计（TF）建立词库，然后统计逆文本频率指数（IDF），得到 VSM 模型后，利用 VSM 模型的向量进行计算。如果要进行进一步向量降维，会进行两种策略，特征选择（Feature Selection，FS）和特征抽取（Feature Extraction，FE）。FS 和 FE 的本质都是对 VSM 模型进行奇异值分解（Singular Value Decomposition，SVD）生成 SVD 矩阵，如果是进行 FS 策略，是通过 SVD 后的左奇异变量代表整体特征，这类分析策略如潜在语义分析（Latent Semantic Analysis，LSA），如果是 FE 策略，则是通过 SVD 的右奇异变量的计算，将 VSM 的维度，如 m 维度向量空间转化（如旋转）到 n 维度向量空间，使 n 维向量空间是 VSM 的向量样本方差最大的空间，减少矩阵稀疏程度，这类策略如主成分分析（Principal Components Analysis，PCA）。由于词袋模型没考虑词语词之间的顺序，一般还可以基于主题的隐含狄利克雷分布（Latent Dirichlet Allocation，以下简称 LDA）对 VSM 的向量进行处理，用于主题分类进行简化。

2.3　数据处理方法

对于分词问题：笔者选取的结构化数据的字段主要包含基金编号、双一流大学名、基金立项时间和基金资助类别。笔者选取的非结构化的字段主要每个资助基金对应的基金关键词。为了简化问题，采用基金关键词作为实验数据是因为关键词列表本身是有关键词词组间的分隔符，尽管自科基金系统在不断升级，但是关键词之间的分割符在一定的时间内是不变的，利于计算机处理。考虑科研基金的行业特点，不采用通用分词工具，而是从自然基金的立项项目的关键词列表进行数据清洗生成面向自然科学基金领域的主题词库。

对于在建立主题词库时，需要考虑立项基金的关键词的停用词表和关键词列表分割词问题。同时考虑到上文提到的“CRISPR/Cas9 基因编辑”等类似词语情况，采取的策略是如果关键词词组是中文开头构成的，进行关键词组分割，如果一个关键词词组是英文开头的，如“Cas9 基因编辑”则视作词库中的一个完整单词，不需要对关键词组进行分割。

对于维数灾难问题：研究词库来自现有立项自科基金的数据清洗，因此可以做到对关键词列表的矢量化。但是由于不同科研工作者对项目关键词的理解不一样，关键词的数量很多，导致向量化的维度很高，需要向量维度降维。笔者对关键词向量的降维主要采

用一种新的方法用于解决维度问题，即 t- 分布领域嵌入算法（t-SNE，t-distributed Stochastic Neighbor Embedding），该方法由图灵奖得主 Geoffrey Hinton 提出。该方法和 PCA 方法的区别在于能更好的可视化，PCA 主要用于压缩数据，但是不能解决线性不可分问题，但是 t-SNE 主要用来解决线性不可分问题。通过 t-SNE 的降维可以有效对数据分类，能够很清晰的看到数据全景。但是他的局限性主要是只能用来做投影，将高维度向量进行映射，但是不反应聚类关系，也不能用于预测,而一般为了降低向量的维度也一般在运行 t-SNE 前先用 PCA 进行初步的数据降维。尽管 t-SNE 有这些局限性，但是并不影响数据分析，而笔者是需要 t-SNE 对双一流大学的所有立项数据进行分析。

3 实验

3.1 数据清洗

采集完成后，数据存到数据库文件，并清除清空关键词为空的项目数据。笔者所有的数据和实验代码可以在 Github[18]下载。

“35 所大学国家自科基金立项表”（1998 年 1 月 1 日—2017 年 12 月 31 日）文件共有 132 899 条记录，其中第一条为表头。代表了这 35 大学 1998 年 1 月 1 日—2017 年 12 月 31 日的所有国家自然科学基金立项项目。以一条数据演示为例，其中字段如表 1 所示。

上图中的 YEAR 为立项时间，grantNo 为国家自然基金委员会分配的基金编号，orgn 为基金依托单位，orgnid 为基金依托单位对应的编码，projectkeyword 为基金关键词。为了方便 MATLAB 处理，上图中的基金标题和基金关键词为非结构化数据，需要将文本变为向量。以基金依托单位编码代表基金依托单位，删除 grantNo 字段，可以直接利用 MATLAB 进行处理。得到数据如表 2 所示。

表 2 35 所大学国家自科基金立项简化表

YEAR	Orgnid	Projectkeyword
1998	100027	枚系统经济学.可持续发展.知识经济

需要对表 2 的数据字段进行数据清洗，建立自然科学基金的主题词库。先将表 2 的 projectkeyword 保存为文本文件 3.txt，然后对于 projectkeyword 字段直接利用分隔符进行拆分，拆分主要分割符号主要规则如表 3 所示。

表 3 分隔符

字符	类别
：	全角冒号
，	全角逗号
、	全角顿号
；	全角分号
:	半角冒号
,	半角逗号
;	半角分号
.	半角句号
.	半角句号

然后删除停用词，如表 4 所示。

特别需要注意的是如果是一个基金中的全英文词组间的空格不视作分隔符，全中文词组间的空格视作分隔符。最后得到 projectkeyword 的生成的清洗词库文件，共有 230 001 条词典记录，生成的清洗词库如表 5 所示。

然后用“清洗词库”对“35 所大学国家自科基金立项简化表”的数据文件进行文档向量化，得到向量化之后的文件和“35 所大学国家自科基金立项简化表”文件对比并合并得到“35 所大学国家自科基金立项简化清洗表”，此处 132 899 条记录中仅有一条记录异常并进行了跟踪（序号为第 2 条），考虑到异常数据占全部数据很低，因此笔者的研究删除了第三条数据确保数据的一致性，示例数据如表 6 所示。

接下来需要对表 6 的数据进行可视化绘制和进行试验，每次筛选实验数据为二维数据结构，实验的标

表 1 35 所大学国家自科基金立项表

YEAR	grantNo	orgn	orgnid	Projectkeyword
1998	79870052	上海交通大学	100027	枚系统经济学.可持续发展.知识经济

DOI：10.13998/j.cnki.issn1002-1248.2019.12.16-1097

表 4　停用词

字符	类别
“	全角左引号
”	全角右引号
“	半角引号
/	斜线
**	双星号
（	全角左括号
）	全角右括号
(	半角左括号
)	半角右括号
《	左书名号
》	右书名号
？	全角问号
?	半角问号
。	全角句号
-	半角连字符

表 5　清洗词库

字符	出现次数/次
数值模拟	806
自噬	725
凋亡	637
信号通路	626
分子机制	601
转移	591
信号转导	579
石墨烯	570
稳定性	561
自组装	523

签变量为时间（YEAR）或依托单位编码（orgnid）。实验的数据变量为关键词对应字典编码。

由于自科基金的系统存储的数据也是在不断完善和规范化，为了判断关键词选取的列数。由于 2017 年是实验数据中最新的年份，以 2017 年作为关键词维度选取的关键词维度选取的数量。

对 2017 年的不同双一流大学的基金项目关键词，则根据表 6 的数据筛选得到“2017 年的 35 所大学国家自科基金立项简化清洗表”文件，如表 7 所示（只显示其中 7 个关键词），共 13 170 项（包含表头）。

笔者对“2017 年的 35 所大学国家自科基金立项简化清洗表”数据进行数据筛选，发现以下结果。

从表就可以看到，2017 年 98.5%自然科学基金的关键词包含 3 项，而采用同样办法筛选 1998 年的数据，和 2017 年的数据进行对比研究，其结果如下：

从表 8 和表 9 的对比，可以看过去 20 年，自然科学基金申报立项书的关键词从 3 个关键词升至 4 个关键词，如果从 1998 年的自科基金数据选取 4 个关键词建立矩阵则矩阵过于稀疏，如果对于 2017 年的自科基金基金数据选取 4 个关键词，则关键词不满足条件的

表 6　35 所大学国家自科基金立项简化清洗表

YEAR	orgnid	projectkeyword	对应字典编码	备注
1998	100027	枚系统经济学. 可持续发展. 知识经济	[枚系统经济学],699,157367	异常数据
1998	100027	非线性控制系统. 观测器设计	14213,17550	正常数据
1998	100027	场强预测. 射线跟踪. 移动通信网规划	212675,91451,161162	正常数据
1998	100027	心室纤颤. 螺旋波. 混沌控制	62250,5824,3115	正常数据
1998	100027	大数量物体. 波浪荷载. 递推算法	215440,10710,133913	正常数据

表 7　2017 年的 35 所大学国家自科基金立项简化清洗表

orgnid	keyword1	keyword2	keyword3	keyword4	keyword5	keyword6	keyword7
100027	198245	796	469	60275	88565	94420	
100027	811	10806	34689	157	3999		
100027	285	15	251	1117	106178		
100027	16118	9	5312				
100027	2308	2994	51153	1463	11400		

表 8 2017 年双一流大学 13 169 项自科基金关键词维度统计结果

关键词列	空白关键词数/个	非空关键词占比/%（不算表头）
Keyword1 列	14	99.89
Keyword2 列	67	99.49
Keyword3 列	198	98.50
Keyword4 列	1 035	92.14
Keyword5 列	3 401	74.17
Keyword6 列	12 277	6.77

表 9 1998 年双一流大学 1 772 项自科基金关键词维度统计结果

关键词列	空白关键词数/个	非空关键词占比/%（不算表头）
Keyword1 列	1	99.94
Keyword2 列	32	98.19
Keyword3 列	219	87.64
Keyword4 列	1 762	0.56
Keyword5 列	1 170	0.11
Keyword6 列	1 772	0.00

不到 10%。但是相关学者如果研究当年的数据，尤其是最近几年的数据，则可以选取 4 个关键词。

在实验中，为了考虑的数据的兼容性，通过 1998 年和 2017 年的数据对比，对每个立项基金采用 3 个关键词标签。

3.2 词库选择

另一方面，对于词库文件共有 230 001 条数据，如果直接将 1998 年和 2017 年的数据进行转化，每一条记录代表在一个 230 001 维度的向量投影，如果该记录包含词典顺序为 N 的关键词，则该向量在 N 维度的内容为 1，反之则为 0。但是即便这样，对于 2017 年的关键词数据是一个[13169 * 230001]的矩阵，如果包含标签和表头，则是一个[13170 * 230002]的矩阵，并且该矩阵相当稀疏。

为了简化实验，在原理不变的前提下，对词库文件进行优化，减少关键词数量，用来降低维度。通过统计词库，发现共有 230 001 个关键词，共出现 582 148 次，即平均每个词出现 2.53 次，发现以下统计结果，如表 10 所示。

从表 10 发现，前 2 万的关键词占了自然科学基金关键词综述一半以上，但是前 2 万关键词占总词库不到 10%，因此选取词频高的关键词进行降维有利于降低矩阵的维度。

基于前 20 000 关键词，对“35 所大学国家自科基金立项简化清洗表”的数据进行降维。如果每个科研基金包含前 20 000 的关键词中的关键词，则保留记录，对于完全不包含前 20 000 的关键词的科研基金暂不研究，经程序处理后的得到文件“包含前 20 000 的关键词的 35 所大学国家自科基金立项简化清洗表”，简称“包含前 20 000 关键词的数据文件”。该文件共包含 7 162 项科研基金。由于该文件中只保留前 20 000 关键词的向量列表，因此会因为没有处理前 20 000 关键词外的关键词，导致有重复序列，删除重复序列后得到 6 958 项(不含表头) 科研基金。这 6 958 项科研基金来自前 20 000 的关键词，得到的文件为“1998 年—2017 年所有基于前 20 000 关键词的实验数据”，其中数据如表 11 所示。

表 10 双一流大学自科基金关键词维度统计结果

关键词序列	关键词条数占总词库条数比例/%	出现次数/次	出现次数占总出现次数比例/%
前 500	0.22	81 362	13.98
前 1 000	0.43	111 650	19.18
前 2 000	0.87	148 103	25.44
前 5 000	2.17	206 608	35.49
前 10 000	4.35	257 125	44.17
前 20 000	8.70	311 780	53.56
前 200 000	86.96	552 147	94.85

DOI：10.13998/j.cnki.issn1002-1248.2019.12.16-1097

表 11　1998 年—2017 年所有基于前 20 000 关键词的实验数据

YEAR	orgnid	keyword1	keyword2	keyword3
1998	100027	3		
1998	100404	10	177	5689
1998	100112	1	5054	
1998	200654	1		
1998	201016	11		

笔者主要对着 6 958 项数据进行分布研究。

3.3　数据降维可视化

为了方便描述，将 year 和 orgnid 做为 label 的变量，并且标记 label，具体分为 label(year)和 label(orgnid)，同时将 keyword 列表标记为变量因子 factor，其中 factor(1)表示 keyword1，factor(1,2)表示 keyword1 和 keyword2，factor(1,2,3)代表 keywords1，keywords2 和 keywords3。

对于 6 958 项实验数据，主要分为两类分析，第一种分析维度是基于时间分析过去 20 年的不同年基金的总成果，建立 label(year)和 factor(1,2,3)的关系并可视化。另一种分析维度是基于单位角度分析过去 20 年不同机构的总成果，建立 label(orgnid)和 factor(1,2,3)的关系并可视化。但是不论是基于时间角度分析还是基于单位角度分析，其目的都是为了对高维数据进行降维，在二维空间进行更清晰直观的可视化并可以寻找其中隐藏的规律。

对于这 label 的分析又分为两种分析策略：

（1）对角矩阵：如果均只包含 keyword1，则构成 label(year)和 factor(1)的对角矩阵或者 label(orgnid)和 factor(1)的对角矩阵，对于对角矩阵的分析可以直接进行 Excel 的统计分析，这类分析最为简单，但是最直观；

（2）非对角矩阵：但是如果分析 label 和 factor(1,2)的关系或者 label 和 factor(1,2,3)的关系，则需要将 keyword 映射到 dict-20000.txt 的高维空间，然后进行降维。

对于以上两种分析策略，主要生成不同的分析目标文件，如表 12 所示。

其中序列 3 为序列 2 的真子集，序列 2 为序列 1 的真子集，同时其中序列 6 为序列 5 的真子集，序列 5 为序列 4 的真子集，但是由于删除了重复项，因此序列 1 和序列 4 的数据条数最少。

如上表所示，对于“二维关键词时间文件”、“三维关键词时间文件”、“二维关键词机构文件”、三维关键词机构文件”文件这些数据文件，根据“前两万关键词库”词库进行投影，并且生成维度为 20 000 的向量，得到矩阵文件，去掉重复项之后，分别为上表所示的“向量化后的二维关键词时间文件”，“向量化后的三维关键词时间文件”，“向量化后的二维关键词机构文件”，“向量化后的三维关键词时间文件”。笔者暂不分析对角矩阵的情况，对角矩阵指“一维关键词时间文件”的数据和“一维关键词机构文件”的数据。

用 t-SNE 算法对“二维关键词时间文件”进行 2 维分析，如图 1 所示。

图 1 反映过去 20 年的自科基金立项数据投影到二维关键词的分布，每个点代表一个基金，不同年份采用不同颜色，即 label(year)~factor(1,2)分布关系。

表 12　不同分析策略的分析文件

序列	数据文件	数据条数/条（含表头）	分析策略	矩阵文件
1	一维关键词时间文件	311	label(year)～factor(1)	
2	二维关键词时间文件	5 619	label(year)～factor(1,2)	向量化后的二维关键词时间文件
3	三维关键词时间文件	5 401	label(year)～factor(1,2,3)	向量化后的三维关键词时间文件
4	一维关键词机构文件	471	label(orgnid)～factor(1)	
5	二维关键词机构文件	5 651	label(orgnid)～factor(1,2)	向量化后的二维关键词机构文件
6	三维关键词机构文件	5 372	label(orgnid)～factor(1,2,3)	向量化后的三维关键词时间文件

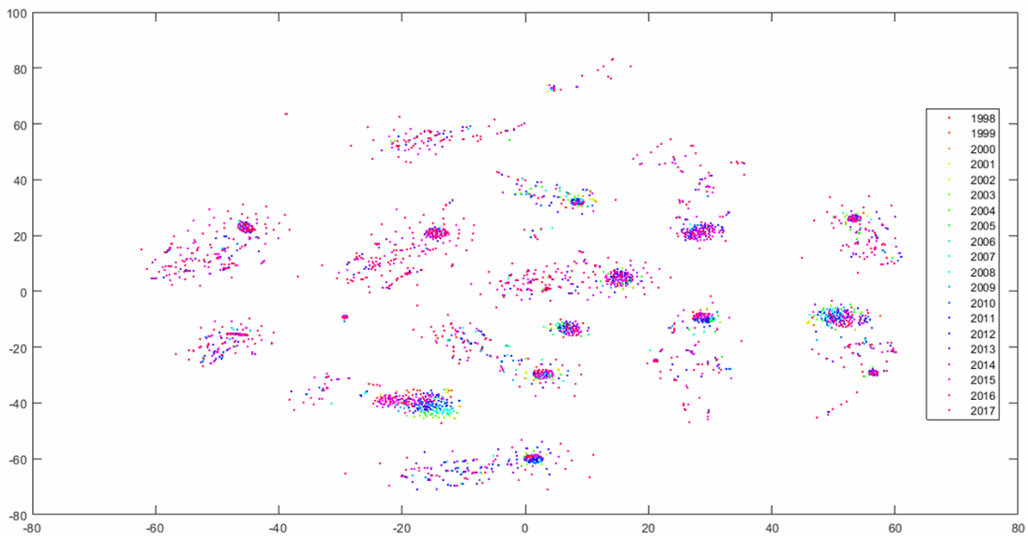

图 1　二维关键词时间文件分析结果

但是需要注意的时候，笔者采用的 t-SNE 算法并不能用于聚类的研究，也不能用于推演预测的研究。不同点之间的距离并不反应聚类关系。上图的意义在于一方面能把图 1 中 5 618 项基金（不含表头）投影到平面，能有效可视化，另一方面要注意颜色重叠部分，要对这边进行分析，发现其中是否有共通的规律。

研究 label(year)~factor(1,2,3)，生成的图如图 2 所示。

图 2 反映过去 20 年的自科基金立项数据投影到三维关键词的分布，每个点代表一个基金，不同年份采用不同颜色，即 label(year)~factor(1,2,3)分布关系。

类似的，研究 label(orgnid)~factor(1,2) 和 label(orgnid)~factor(1,2,3) 生成图 3 和图 4，图 3 代表对 5 650 项二维关键词进行投影，图 4 代表对 5 371 三维关键词进行投影，如图 3 所示。

图 3 和图 4 代表不同科研机构过去 20 年的立项关键词的分布，图 3 为二维关键词，图 4 为三维关键词。图中每一个点代表一项基金，不同的颜色代表不同单位，左边的编码为基金依托单位编码。

4　结语

通过笔者的预处理、词库建模及可视化处理，可以将过去 20 年的自科基金立项数据数据进行绘制，并

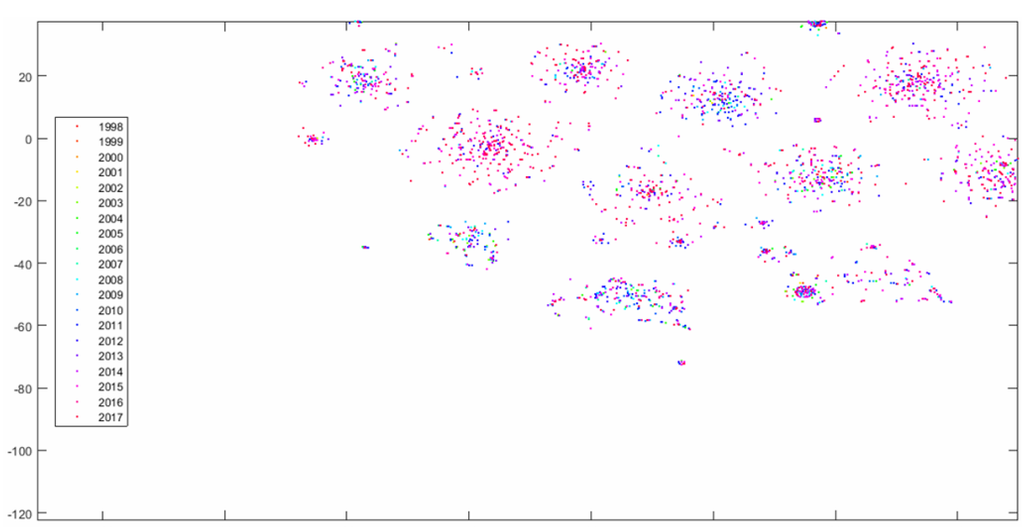

图 2　三维关键词时间文件分析结果

DOI：10.13998/j.cnki.issn1002-1248.2019.12.16-1097

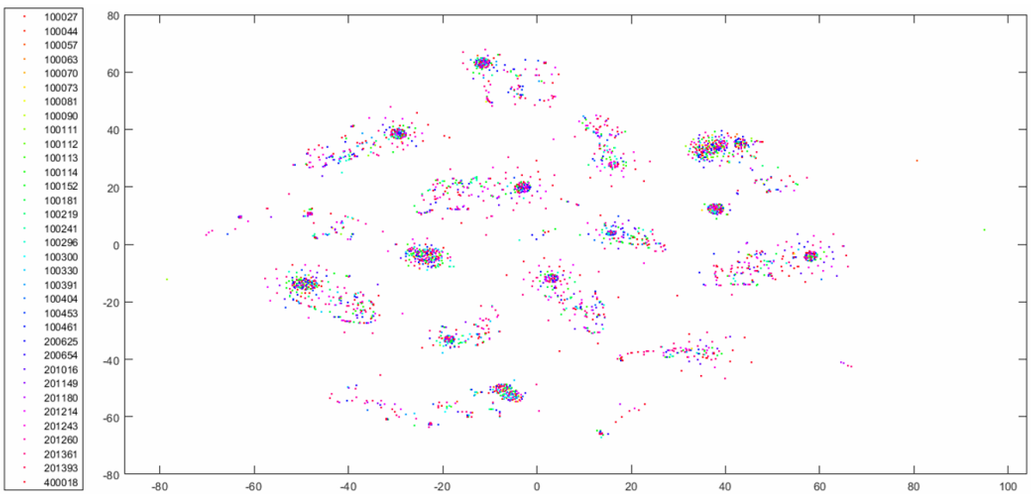

图 3　二维关键词机构文件分析结果

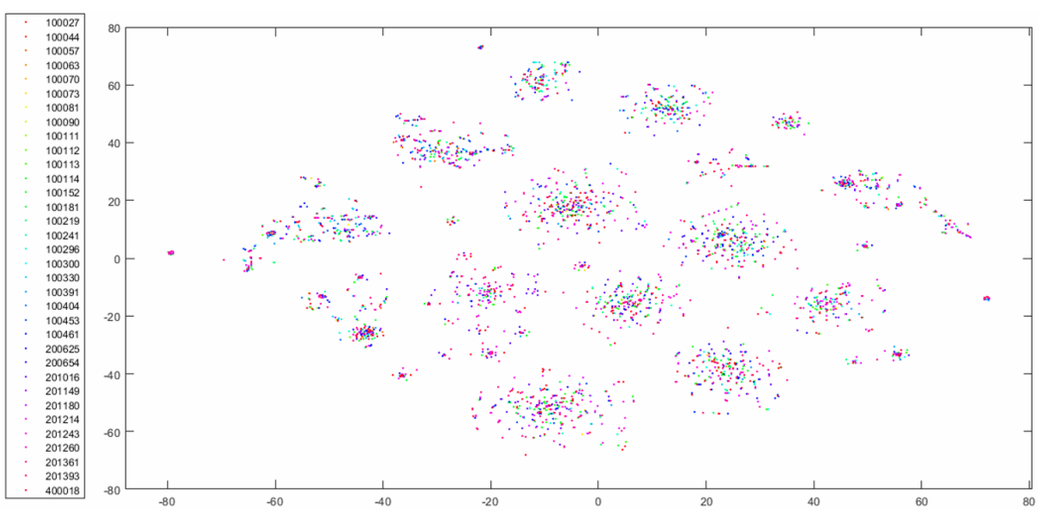

图 4　三维关键词机构文件分析结果

可以采用时间维度和依托单位维度进行分类。不论从时间维度还是依托单位维度，建立起关键词映射关系都是一个巨大的稀疏矩阵，但是可以通过预处理过滤掉无效词，并且基于词频建立起领域词库，最后利用 t-SNE 算法对高维度矩阵进行降维。降维之后，实现了自动分类，通过自动分类，有利于在微观层面为双一流大学提供学科建设的科学决策支撑，同时在宏观层面为科技行政主管部门提供宏观决策支持及总结经验，这是笔者的主要研究成果。

但是笔者的方法也存在一定的局限性，局限性主要原因一方面是国家自然科学基金委的网站在不断升级，很多历史数据并不合规。在词库建立上，鉴于中文词义的复杂性，本身很难提供一个足够通用的词库。在降维处理上，笔者只研究了二维关键词和三维关键词，并且采用的分析数据是基金立项时的关键词列表，但是如果是对基金的标题或者全文进行词库生成和分析，则研究的信息量比纯基金申报关键词精准。在词库处理上，如果对词语本身的上位词和下位词分析，则有利于减少词库的条数和降维。

笔者研究可视化建模的学科意义与主要贡献在于：国家自然科学基金项目关键词反映了国内科研的趋势，是有重要学术研究价值的目标数据源。但是，直接分析国家层面资助的海量项目会带来“维度灾难”。数据可视化的降维方法是众多降维研究中的最清晰直观研究点。

笔者基于 t-SNE 的研究方法是对海量自然科学基金数据的可视化分类方法的研究。而同行主要基于 VOSViewer 或者 CiteSpace 进行研究。VOSViewer 或

者 CiteSpace 的研究国家自然科学基金数据的方法是需要人工干预的交互式研究。而笔者采用 t-SNE 方法研究是一种无监督学习的研究方法。

笔者在处理海量数据时，能降低研究者进行数据分类的工作量，让研究者能更关心业务而是不是技术。另外，相关学者也可以在笔者研究基础上，进一步建模和编程，尝试例如进行交互式的可视化建模，对海量项目数据进行快速定位，提高科研效率。

参考文献：

[1] 万华. 基于项目论文引文关联的协同研究关系分析——以国家自然科学基金图书情报类研究项目为例[J].情报科学,2013(6):53-59.

[2] 范云满,马建霞,刘静.国家自然科学基金的评估指标体系与指标的分析研究[J].图书情报工作,2013,57(16):100-106.

[3] 刘多, 宋敏, 谢亚南等.2009—2015 年国家自然科学基金资助产出 ESI 高被引论文分析[J].中国科学基金,2017(4):353-358.

[4] 冯磊,朱宇华,吕相征等.国家自然科学基金资助产出 SCI 医药卫生论文的计量分析[J].科技与出版,2017(3):112-118.

[5] 邓方,宋苏,刘克等.国家自然科学基金自动化领域数据分析与研究热点变化[J].自动化学报,2018,44(02):377-384.

[6] 陈挺,李国鹏,王小梅.基于 t-SNE 降维的科学基金资助项目可视化方法研究[J].数据分析与知识发现,2018,2(08):1-9.

[7] MEJIA C,KAJIKAWA Y.Using acknowledgement data to characterize funding organizations by the types of research sponsored: the case of robotics research[J].Scientometrics,2018,114(3):883-904.

[8] LI K, YAN E. Are NIH-funded publications fulfilling the proposed research? An examination of concept-matchedness between NIH research grants and their supported publications[J]. Journal of Informetrics,2019,13(1):226-237.

[9] YANG C, HUANG C, SU J. An improved SAO network-based method for technology trend analysis: A case study of graphene[J]. Journal of Informetrics,2018,12(1):271-286.

[10] ABRISHAMI A, ALIAKBARY S. Predicting citation counts based on deep neural network learning techniques[J]. Journal of Informetrics, 2019,13(2):485-499.

[11] FENG F, ZHANG L, DU Y, et al. Visualization and quantitative study in bibliographic databases: A case in the field of university - industry cooperation[J]. Journal of Informetrics,2015,9(1):118-134.

[12] 袁润,李莹,王琦等. 用 R 语言分析关键词集共现网络研究[J].现代情报,2018,38(07):88-94.

[13] LI K, YAN E, FENG Y. How is R cited in research outputs? Structure, impacts, and citation standard[J].Journal of Informetrics, 2017, 11(4):989-1002.

[14] 张永安,马昱.基于 R 语言的区域技术创新政策量化分析[J].情报杂志,2017,36(03):113-118.

[15] 张然.“双一流” 背景下加强国家自然科学基金组织申报工作探讨——以吉林大学电子科学与工程学院为例[J].办公室业务,2019,306(1):186-187.

[16] 张品慧,张瑜婷,赵星.科学基金对“双一流”建设学科的前期资助研究(2012-2016)[J].图书与情报,2018,182(4):10-16.

[17] 马晓萌,徐峰.双一流高校自然科学基金面上项目资助特点探析[J].情报工程,2018,4(06):63-75.

[18] tsne[EB/OL].[2019-06-05].https://github.com/greysh/paper-tsne.

[19] HU K, WU H, QI K, et al. A domain keyword analysis approach extending Term Frequency-Keyword Active Index with Google Word2 Vec model[J].Scientometrics,2018,114(3):1031-1068.

[20] CHEN G, XIAO L.Selecting publication keywords for domain analysis in bibliometrics: A comparison of three methods[J].Journal of Informetrics,2016,10(1):212-223.