DOI：10.16353/j.cnki.1000-7490.2019.05.024

● 周彦廷，白如江，王效岳（山东理工大学科技信息研究所，山东 淄博 255049）

# 主题对比视域下的科学前沿识别方法研究*
## ——以碳纳米管领域为例

摘 要：［目的/意义］以两种科技文本数据（科技规划文本和基金项目数据）为数据源，构建主题对比视域下的科学前沿识别方法，以期识别出碳纳米管领域科学前沿。［方法/过程］首先，获取碳纳米管领域科技规划文本和基金项目数据，面向科技规划文本提出了一种以触发词库为基础的规则匹配抽取研究主题的研究方法；面向基金项目数据利用主题模型对其进行研究主题识别；通过计算余弦相似度的方法对比研究主题，结合项目数、资助时长、资助强度等指标构建科学前沿识别模型，并对科学前沿的研究价值与意义进行综合评价。［结果/结论］实验结果表明该方法可以更有效地识别出科学研究前沿主题，科技规划文本的识别粒度为句子级，相比以词为最小识别单位的识别，结果较为宏观。

关键词：科学研究前沿；主题识别；信息抽取；多源数据

**Research on the Identification Method of Scientific Front from the Perspective of Topic Contrast: A Case Study of Carbon Nanotubes**

**Abstract:** [Purpose/significance] This paper uses the two data sources of science and technology planning text and fund project data to compare the topics and identify the research front topics. [Method/process] Obtaining the data of science and technology planning and fund project data in the field of carbon nanotubes, and proposing a research method of the topic of rule matching extraction based on triggering thesaurus for the text of science and technology planning; comparing research topics by calculating cosine similarity, the identification model of scientific front was established by combining the number of projects, funding duration, funding intensity and other indicators, and the research value and significance of scientific front was comprehensively evaluated. [Result/conclusion] The experimental results show that the method can identify the front of scientific research more effectively, and the recognition granularity of the scientific planning text is sentence level. Compared with the word as the smallest recognition unit, the result is more macroscopic.

**Keywords:** scientific research front; topic recognition; information extraction; multi-source data

科学技术的发展具有一定的规律：以科学领域的进步与技术的突破为先导，引发各学科领域系统性、自发性、群发性的创新突破[1]。研究前沿识别成为各领域科研人员竞相研究的科学重点。如何准确识别研究前沿是获取科技战略情报的重要基础，也是决策层制定发展战略、规划研究布局的智库保障[2]。科学研究前沿识别研究是科技战略情报研究的重要方向之一，它对于战略决策特别是在支撑重要领域科技创新和支持宏观科技决策的前瞻性、战略性的科技情报信息研究与服务有着重要的作用和影响[3]。

## 1 相关研究

Price 最早提出了“研究前沿（Research Front）”的概念，他认为在一个给定的研究领域，科学家积极引用的近期文献的集合所表征的研究领域便是研究前沿[4-5]。随后，许多学者从不同角度和层面定义了研究前沿，至今还没有形成统一的共识。虽然研究前沿至今没有统一的定义，通常被认作是某时期内最具发展潜力的新兴研究领域或主题，但被普遍认为是科学研究中最新、最先进、最有发展潜力的研究主题或领域[6]。它来自于科学发现，代表了科学发展的难点、重点以及发展趋势，具有高度的前瞻性[7]。

研究前沿的识别方法主要为基于专家知识识别（Expert-based）和基于计算机识别（Computer-based）两个方面[8-9]。基于专家知识识别的方法在过去的几十年中一直是被广泛地使用，主要是专家通过阅读大量文献或者通过交流识别、判断研究前沿。然而，基于专家知识识别的方法识别研究前沿耗费时间、效率低、主观性强，学者在这

* 本文为国家社会科学基金项目“未来新兴科学研究前沿识别研究”的成果，项目编号：16BTQ083。

方面的研究也越来越少[10]。基于计算机的识别方法是通过计算机进行科学计量或文本挖掘，高效地辅助科学家识别研究前沿。由于计算机分析具有更高的时间和成本效益，所以计算机的识别适用于从海量的信息中识别出研究前沿，相较于基于专家知识识别的方法更贴合时代的需求。

计算机技术不断进步为科研人员成功探测研究前沿提供了契机，诸多专家利用计算机技术实现了研究前沿的识别。侯剑华等以 Web of Science 数据库为研究对象，利用 Citespace 绘制纳米技术文献突现词共引混合网络图谱，分析纳米技术研究的前沿热点[11]。白如江等以科技规划文本为研究对象，基于自然语言处理技术，利用 Java 语言开发了主题抽取工具 Topic Extract Tool，准确的绘制出了碳纳米管领域的科学研究前沿地图[12]。刘自强等提出一种基于时间序列模型的研究热点分析预测方法，以 CNKI 收录的以竞争情报为主题的期刊论文为研究对象，运用社会网络分析、关键词群分析和时间序列模型分析预测其研究前沿的发展趋势[13]。

经过文献梳理发现科学研究前沿识别方法已经形成了相对成熟的方法体系，但仍有不足之处：①时滞性问题。由于科学论文为科研活动产出结果的主要表现形式，因此，相关专家开展的科学前沿识别方法研究多从期刊论文数据入手。然而，学术论文从审稿到发表再到产生具体的引用关系，这个过程使得论文数据本身在时间上存在一定的滞后性。②研究数据源单一问题。以某一种数据源来识别研究前沿具有局限性，并不能全面代表所有科学研究前沿信息。基金项目数据与论文数据相比，其研究主题皆为还未进行深入研究的研究主题，是一种规划和目标，而论文数据为研究的结果产出，从科技规划文本规划到基金项目申请再到论文成果可能需要 1～3 年的时间[14]。

由此可见，基于科技规划文本和基金项目数据的研究前沿识别相比论文数据，其前瞻价值更高[14]。因此，本文以科技规划文本和基金项目数据为数据源，运用文本数据挖掘的方法分别识别出其研究主题，并进行对比分析，以期更前瞻的识别出研究前沿，为研究前沿识别相关研究提供思路借鉴。

## 2 研究思路

为了识别出科技规划文本和基金项目数据中的研究前沿主题，本文以科技规划文本和基金项目数据作为研究对象。首先将两种数据源以 2 年为一子时期进行切片划分，分别运用触发词库匹配的方法和 LDA 主题识别模型两种数据源中的研究主题，然后将得到的研究主题做相似度计算，并结合基金项目数据研究主题的项目数、资助时长、资助强度构建研究前沿主题识别模型，识别出研究前沿，最后对研究前沿进行评价，具体流程如图 1 所示。

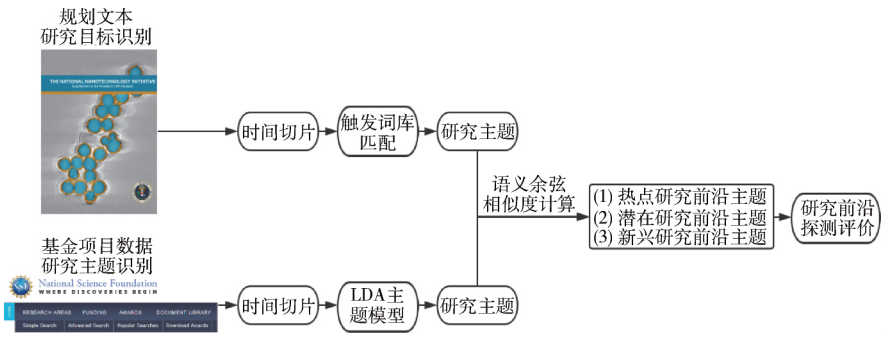



图 1 探测研究前沿主题方法流程

通过对比科技规划文本和基金项目数据的研究主题可能出现以下几种情况[15-17]：

1）热点研究前沿主题：与该科技规划文本研究主题对应的基金项目数据研究主题的研究数目、资金投入与时间投入较大，并随时间的推移呈平稳或增长的趋势。

2）潜在研究前沿主题：该科技规划文本研究主题既不是热点研究前沿，也不是新兴研究前沿，但未来存在成为新兴研究前沿或热点研究前沿的可能性与潜力。

3）新兴研究前沿主题：与该科技规划文本研究主题对应的基金项目数据研究主题的研究数目、资金投入与时间投入较少，但是整体趋势随时间的推移呈增长趋势的研究主题，未来可能会成为热点研究前沿。

### 2.1 基于科技规划文本的研究主题识别

科技规划文本为非结构化文本，且文本叙述较为零散，同领域的规划十分的分散，且文本中还有不少篇幅在叙述已完成的研究项目以及完成的效果是什么，还有大量学术会议的叙述。实现抽取科技规划文本中的研究主题并转为结构化的文本存在一定的难度。为此，本文提出了一种以触发词库为基础的规则匹配抽取研究主题的方法，以期得到理想的结果，具体研究目标识别流程如图 2 所示。

第一步：获取数据与预处理。登录发布科技规划文本的相关网站，下载数据，将下载后的科技规划文本进行文本格式转换和句子级抽取。

第二步：抽取领域相关文本。在前期人工判读基础上，确定与目标领域相关的关键词，构建正则表达式，通过正则匹配方法从科技规划文本中抽取含有关键词或时间的句子作为研究主题识别数据集。

第三步：构建规划触发词库。利用 Standford coreN-

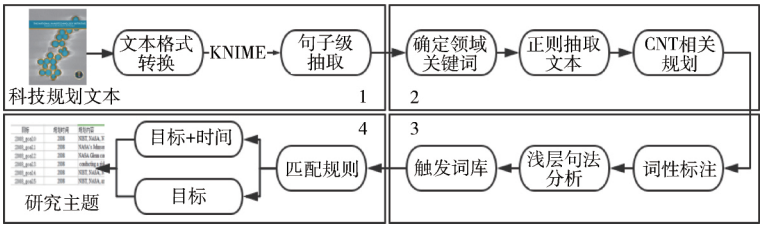



图2　科技规划文本的研究主题识别流程

LP[18]对上述抽取出的句子进行词性标注与浅层句法分析，分析含有规划语义句子的结构，寻找含有规划语义句子的特征，将发现的特征归纳为范式，并将范式描述出来。为了提高识别规划的准确度，继续利用正则表达式抽取含有特征范式的词组，剔除不含规划语义的词组，同时集合表达目标意思的单词，构建规划触发词库。

第四步：研究主题识别。利用构建好的规划目标触发词库，通过构建匹配规则识别出含有规划研究目标的句子及规划目标时间，若无时间信息，设定规划目标时间为此规划研究目标所在规划文本的年份。

### 2.2　基于基金项目数据的研究主题识别

基金项目数据本身就是结构化的数据源，本文通过LDA主题模型识别基金项目数据研究主题，具体识别主题流程如图3所示。

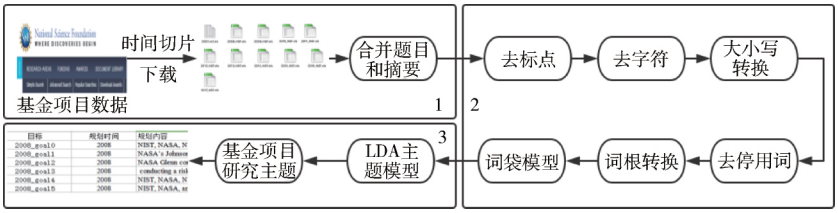



图3　基金项目数据的研究主题识别流程

第一步：获取数据。登录基金项目数据库的相关网站，构建检索式，下载基金项目数据，并对已下载的基金项目数据进行时间切片。

第二步：数据预处理。对基金项目数据进行文本格式转换，合并基金项目题目与摘要作为实验文本，对文本进行去标点、去字符、去数字，大小写转化、去除停用词、词根转换等预处理。

第三步：研究主题识别：利用LDA模型依次识别出不同子时期的基金项目数据的研究主题，构建文档—主题矩阵与主题—主题词矩阵，为研究前沿主题探测做基础。

计算基金项目数据研究主题的资助时长与资助强度，为构造研究前沿主题识别模型做基础，并对研究前沿主题进行评价，具体如下：

1）资助时长（Funding Time，FT）。基金项目的资助时长越长表明该研究主题需要更深入的研究，代表其研究意义更大。

$$FT_t = \frac{SumFT_t(X)}{PC_t(X)} \quad (1)$$

式中，$PC_t(X)$为在$t$子时期内与研究主题$X$相关的基金项目总数（Project Count，PC）；$SumFT_t(X)$为在$t$子时期内与研究主题$X$相关的基金项目的资助总时长（单位为年）；$FT_t$为在$t$子时期内与研究主题$X$相关的基金项目的平均资助时长。

资助时长，主要分析与同一研究主题相关基金项目的平均资助时长，反映了对该研究主题的重视程度，认为该研究主题值得投入时间进行长期探索和研究。

2）资助强度（Funding Amount，FA）。基金项目的资助强度主要由基金项目的资助金额反映，基金项目的资助金额越高说明其研究价值和研究难度越高。

$$FA_t = \frac{SumFA_t(X)}{PC_t(X)} \quad (2)$$

式中，$SumFA_t(X)$为在$t$子时期内与主题$X$对应基金项目的资助总金额；$FA_t$为在$t$子时期内与主题$X$对应基金项目的平均资助强度；$PC_t(X)$为在$t$子时期内与研究主题$X$对应基金项目总数（Project Count，PC）。

资助强度通过计算与同一研究主题对应基金项目的平均资助金额，反映该研究主题的研究困难程度（如研究设备投入等），但正是由于研究具有一定的难度，越可能成为科技创新突破口，主题的资助强度越高，前瞻价值越高[18]。

### 2.3　基于科技规划文本和基金项目数据的研究前沿识别

为了识别出研究前沿，将科技规划文本的研究主题与基金项目数据的研究主题进行语义余弦相似度计算，该指标数值越高则说明科技规划文本映射在基金项目数据研究主题的文本重合度越高，表明科技规划文本的研究主题在基金项目数据中有所布局，通过相余弦似度计算与研究前沿探测指标相结合，可以识别出研究热点主题、潜在研究前沿主题、新兴研究前沿主题，计算公式如下：

$$Topic\_cosin = Cos(Goal_i, Topic_f) \quad (3)$$

式中，$Goal_i$为科技规划文本的研究主题；$Topic_f$为基金项目数据的研究主题；$Cos(Goal_i, Topic_f)$为两个数据源研究主题的语义余弦相似度值。

计算出研究主题相似度后，结合项目数、资助时长以及资助强度等指标构建研究前沿主题识别模型，识别出的研究前沿主题分为热点研究前沿主题、新兴研究前沿主题以及潜在研究前沿主题，研究前沿主题识别模型如图4

所示。

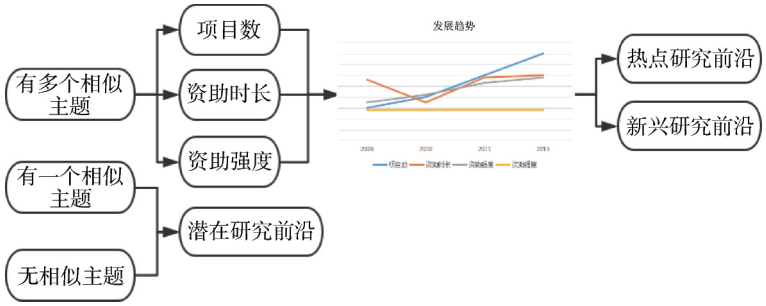



图4　研究前沿主题识别模型

找出与同一科技规划文本的研究主题有一定相似度的基金项目数据的研究主题，将其项目数、资助时长、资助强度的三项数值按时间先后顺序排列起来，识别规则如下：

1）热点研究前沿主题：存在多个与同一科技规划文本研究主题相似的基金项目数据研究主题，项目数、资助时长、资助强度的指标数值皆高于平均值，且随时间推移整体趋势非下降趋势。

2）新兴研究前沿主题：存在多个与同一科技规划文本研究主题相似的基金项目数据研究主题，项目数、资助时长、资助强度的指标数值随时间推移具有增长趋势，且存在某项指标数值低于平均值。

3）潜在研究前沿主题：因与同一科技规划文本研究主题存在一个或无相似的基金项目数据研究主题，代表目前未有围绕此研究主题的前沿研究，故无法根据对应的基金项目数据研究主题的各项指标数值进行判断对比，科技规划文本的研究主题本身具有前瞻性，是前沿研究主题，尽管目前研究人员对此关注较少，但具有一定的潜力在未来会获得更多的关注，发展成新兴研究前沿主题，甚至是热点研究前沿主题。

## 3　实证研究

### 3.1　实验平台

硬件环境：CPU（R）Core（TM）i5-3360M @ 2.80GHz；内存（RAM）——16.00GB

操作系统：Windows7 旗舰版 64 位

工具：KNIME，Python2.7，Standford coreNLP

### 3.2　研究领域与数据获取

美国在众多科研领域占据领先地位，美国纳米技术科技规划（The National Nanotechnology Initiative，NNI），本文皆简称为 NNI 科技规划，其中所规划的研究目标代表了相关领域未来的研究方向；美国科学基金会（National Science Foundation，NSF），本文皆简称为 NSF 基金项目，其资助的基金项目代表了相关领域的最高技术水平，因此，以 NNI 科技规划文本和 NSF 基金项目数据作为识别研究前沿的数据源是可行有效的[14,19]。

科技规划文本：从 NNI 科技规划文本相关网站（https://www.nano.gov/）上下载 2008—2017 年的 NNI 科技规划文本，共得到 10 年的文本。

基金项目数据：检索数据库：NSF 基金项目数据库；数据检索式：Keyword = " carbon nanotube* "；检索范围：基金项目名称；时间跨度：2008 年 1 月 1 日—2017 年 12 月 31 日；检索结果：8724 项。

### 3.3　科技规划文本的研究主题识别

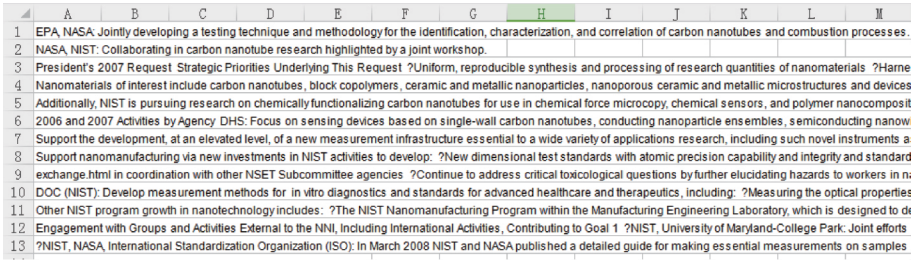
3.3.1　数据预处理　对 NNI 科技规划文本进行文本格式转换，利用 KNIME 平台的 Sentence Extractor 模块对文本进行句子级抽取，储存为 Excel 表格格式，每一个单元格存储一句话，共得到结果 11624 条。

3.3.2　抽取相关文本　通过前期人工判读，确定与碳纳米管领域相关的关键词有“carbon nanotube*”“CNT*”“single wall carbon nanotube*”“SWNT*”“double wall carbon nanotube*”“DWNT*”“mutilwall carbon nanotube*”“MWNT*”，利用 Python 通过正则匹配的方法，为了从 NNI 科技规划文本中抽取含有上述关键词或时间的句子作为碳纳米管领域规划文本数据源，其中“carbon nanotube*”“single wall carbon nanotube*”“double wall carbon nanotube*”“mutilwall carbon nanotube*”都重复包含“carbon nanotube”，“SWNT*”“DWNT*”“MWNT*”都重复包含“WNT”，因此以关键词重复包含的字段为特征，便可得到理想结果，所以构建的正则表达式为：

“(.*[Cc][Aa][Rr][Bb][Oo][Nn][Nn][Aa][Nn][Oo][Tt][Uu][Bb][Ee].*)|(.*[Cc][Nn][Tt].*)|(.*[Ww][Nn][Tt].*)|(.*20\d\d.*)”。共抽取得到碳纳米管相关句子 303 条，结果如图 5 所示。

3.3.3　构建规划触发词库　为了提高抽取结果的精准度，更加精确的定位到含有规划内容的句子，需要构建规划触发词库。为了寻找规化语义文本的特征，并对其进行正确的表达，本文利用 Standford coreNLP 对句子文本进行词性标注和浅层句法分析，标注结果如图 6 所示。

通过总结发现主要有三种特征范式：第一种范式为 be + VBD（动词过去式）；第二种范式为 VB（动词基本形式）\ VBG（动名词和现在分词）\ VBN（过去分词）\ VBP（动词非第三人称单数）\ VBZ（动词第三人称单数）+ IN（介词或从属连词）；第三种范式为 NN（常用名词单数形式）\ NNS（常用名词复数形式）\ NNP（专有名



| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | EPA, NASA: Jointly developing a testing technique and methodology for the identification, characterization, and correlation of carbon nanotubes and combustion processes. | | | | | | | | | | | | |
| 2 | NASA, NIST: Collaborating in carbon nanotube research highlighted by a joint workshop. | | | | | | | | | | | | |
| 3 | President's 2007 Request Strategic Priorities Underlying This Request ?Uniform, reproducible synthesis and processing of research quantities of nanomaterials ?Harne | | | | | | | | | | | | |
| 4 | Nanomaterials of interest include carbon nanotubes, block copolymers, ceramic and metallic nanoparticles, nanoporous ceramic and metallic microstructures and devices | | | | | | | | | | | | |
| 5 | Additionally, NIST is pursuing research on chemically functionalizing carbon nanotubes for use in chemical force microcopy, chemical sensors, and polymer nanocomposit | | | | | | | | | | | | |
| 6 | 2006 and 2007 Activities by Agency DHS: Focus on sensing devices based on single-wall carbon nanotubes, conducting nanoparticle ensembles, semiconducting nanow | | | | | | | | | | | | |
| 7 | Support the development, at an elevated level, of a new measurement infrastructure essential to a wide variety of applications research, including such novel instruments a | | | | | | | | | | | | |
| 8 | Support nanomanufacturing via new investments in NIST activities to develop: ?New dimensional test standards with atomic precision capability and integrity and standard | | | | | | | | | | | | |
| 9 | exchange.html in coordination with other NSET Subcommittee agencies ?Continue to address critical toxicological questions by further elucidating hazards to workers in n | | | | | | | | | | | | |
| 10 | DOC (NIST): Develop measurement methods for in vitro diagnostics and standards for advanced healthcare and therapeutics, including: ?Measuring the optical properties | | | | | | | | | | | | |
| 11 | Other NIST program growth in nanotechnology includes: ?The NIST Nanomanufacturing Program within the Manufacturing Engineering Laboratory, which is designed to de | | | | | | | | | | | | |
| 12 | Engagement with Groups and Activities External to the NNI, Including International Activities, Contributing to Goal 1 ?NIST, University of Maryland-College Park: Joint efforts | | | | | | | | | | | | |
| 13 | ?NIST, NASA, International Standardization Organization (ISO): In March 2008 NIST and NASA published a detailed guide for making essential measurements on samples | | | | | | | | | | | | |

图5　正则表达式抽取目标句子

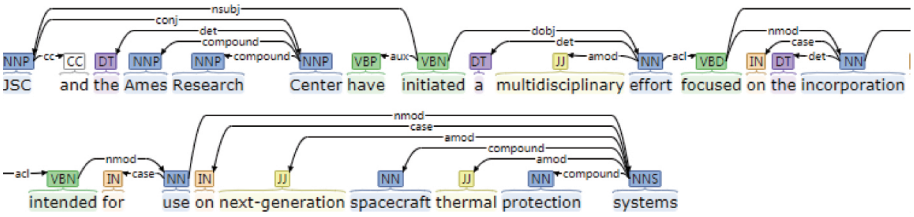



图6　浅层句法分析

词单数形式）\ NNPS（专有名词复数形式）+ IN（介词或从属连词）。

对上述的三种特征范式分别构建正则表达式，抽取对应词组构，剔除不含规划语义的词组，同时集合表达目标意思的单词，构建规划触发词库，结果如表1所示。

表1　触发词（部分）

| | 触发词 |
|---|---|
| 范式一 | be emphasized、be investigated、be used、be applyed … |
| 范式二 | focus on、aim to、plan to、continues to、research to … |
| 范式三 | the challenge to、the purpose of、the goals to、the development of … |
| 单词 | goal、address、develop、purpose、aim、highlight、application … |

3.3.4　研究目标识别　利用构建好的规划目标触发词库，通过构建匹配规则识别出含有规划研究目标的句子及规划目标时间，若无时间信息，设定规划目标时间为此规划研究目标所在规划文本的年份，具体匹配规则如下：

变量说明：

NNI_CNT_20XX：20XX 年 NNI 规划文本中与碳纳米管领域相关文本集合

Goal_dic：规划研究目标触发词库

Goal_sen：含有规划研究目标的句子

FORGoal_dic IN NNI_CNT_20XX

IF MATCH THEN

IF 20\d\d

STORE

(Goal_sen，20\d\d)

ELSE

STORE

(Goal_sen，20XX)

根据上述的匹配规则，共识别出研究目标58个，规划文本的规划领域范围较大，并不是只针对碳纳米管领域的规划，所以各年在碳纳米管领域规划的研究目标数差距较大，规划的具体内容及时间如图7所示。

| 目标 | 规划时间 | 规划内容 |
|---|---|---|
| 2008_goal_0 | 2008 | NIST, NASA, NIOSH: These agencies have initiated a coordinated effort to develop the first Reference Mat |
| 2008_goal_1 | 2008 | NASA's Johnson Space Center (JSC) is working with the State of Texas and the Nanoelectronics Research I |
| 2008_goal_2 | 2008 | NASA Glenn continues to work on taking advantage of enhancedelectromechanical properties in CNT nanoco |
| 2008_goal_3 | 2008 | conducting a risk assessment on carbon nanotubes |
| 2008_goal_4 | 2008 | NIST, NASA, NIOSH: Carbon nanotubes have the potential of leading to significant advances in microelectro |
| 2008_goal_5 | 2008 | NIST, NASA, and NIOSH have initiated a coordinated effort to develop the first Reference Material for resid |
| 2009_goal_0 | 2009 | Joint efforts by these institutions have resulted in significant progress in several projects on nanomaterials, inch |
| 2009_goal_1 | 2009 | This involves planning and conducting toxicology research across the classes of nanoscale materials, including |
| 2010_goal_0 | 2010 | Improved solar cell efficiency with inclusion of carbon nanotubes as the interface layers of a standard 3-layer |

图7　规划内容及规划时间（部分）

### 3.4　基金项目数据的研究主题识别

3.4.1　数据预处理　使用 KNIME 平台对获得的 NSF 美国国家基金项目数据进行数据预处理，首先将数据导入 KNIME 平台，转换文本格式，将项目名称和摘要文本合并处理，进行数据清洗，去掉标点、字符、数字和停用词、大小写转换和词根转换操作。

3.4.2　研究主题识别　基于 LDA 主题模型，通过分别对2008—2017年中的各年数据集进行主题建模，可以得到主题和文档以及主题词的矩阵，通过文档—主题矩阵和主题—主题词矩阵可以得到主题和基金项目的映射关系和主题表征结果，为下面实验中统计分析基金项目的资助时长和资助强度提供数据支撑。共识别出50个主题以及500个主题词，识别出的主题词结果如表2所示。

相关参数设置：No of topic 主题数 5；No of words per topic 每个主题的词数 10；Alpha 狄利克雷分布 0.5；Beta 狄利克雷先验参数 0.1；No of iteration 迭代次数 4000；No of thread 线程数 20。

根据公式（1）和公式（2），对 NSF 基金项目中的主题计算其 FA 与 FT 数值，为了方便数据的对比，对项目数和 FA、FT 指数进行 Min-max 标准化计算，分别得到 Num_ nor，FA_ nor，FT_ nor，各指标以其所有子时期研

究主题的平均数为阈值，判断其的研究意义与研究价值，Min-max 标准化结果如表 3 所示。

表 2　识别出的主题词结果（2008 年）

| Topic | Topic words |
|---|---|
| topic0 | regrowth | catalyst | seeds | effective | templating | chirality | diameter | applications | soil | growth |
| topic1 | energy | macro-films | properties | fabrication | graduate | composites | based | deformable | systems | undergraduate |
| topic2 | thermal | transfer | heat | devices | gas | transport | membranes | materials | membrane | performance |
| topic3 | growth | properties | techniques | science | nanowires | synthesis | graduate | development | structures | diameter |
| topic4 | polymer | sensing | applications | platform | performance | conducting | crystals | single | properties | polymers |

表 3　Min-max 标准化结果（部分）

| | Num_nor | FA_nor | FT_nor |
|---|---|---|---|
| 2008_topic_0 | 0.37 | 0.49 | 0.58 |
| 2008_topic_1 | 0.56 | 0.53 | 0.67 |
| 2008_topic_2 | 0.21 | 0.69 | 0.66 |
| 2008_topic_3 | 0.83 | 0.72 | 0.42 |
| 2008_topic_4 | 0.35 | 0.43 | 0.56 |
| 2009_topic_0 | 0.25 | 0.37 | 0.21 |
| 2009_topic_1 | 0.16 | 0.41 | 0.43 |
| … | … | … | … |

### 3.5　研究前沿识别与探测

分别对已经得到的 NSF 基金项目的研究主题和 NNI 规划文本的研究主题添加标签，例如 2008 年 NSF 基金的主题 topic_0 标记为 2008_topic_0，2009 年 NNI 科技规划文本的 goal_3 标记为 2009_goal_3；利用余弦相似度公式，对已标记的 NSF 基金项目的研究主题和 NNI 科技规划文本的研究主题进行相似度计算：①以 NNI 科技规划文本的研究主题为横坐标，以 NSF 基金项目的研究主题为纵坐标，得到一个 58* 50 的主题相似度矩阵。②以 NSF 基金项目的研究主题为横纵坐标，得到一个 58* 58 的主题相似度矩阵。③以 NNI 科技规划文本的研究主题为横纵坐标，得到一个 50* 50 的主题相似度矩阵；为了方便从结果中识别出相似的主题，将矩阵转化成热力图，如图 8 所示。

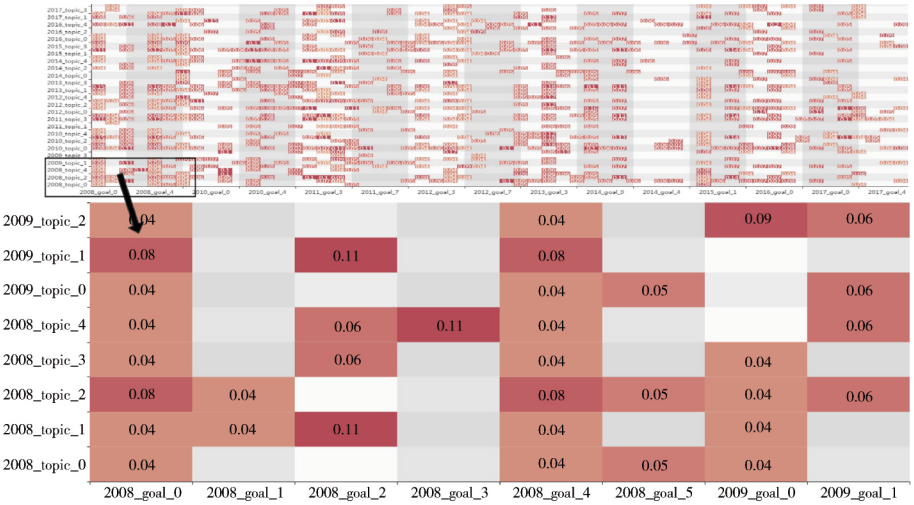



图 8　主题相似度热力图

通过分析 NNI 规划文本与 NSF 基金项目研究主题相似度热力图并结合 Num_nor，FA_nor，FT_nor 构造研究前沿主题识别模型，识别热点研究前沿主题、潜在研究前沿主题、新兴研究前沿主题。根据主题相似度热力图的热度颜色，找出与同一 NNI 科技规划文本的研究主题有一定相似度的 NSF 基金项目的研究主题，将其 Num_nor，FA_nor，FT_nor 的三项数值按时间先后顺序排列起来：①若其各项指标数值皆大于平均值且整体趋势非下降趋势，则判断该 NNI 科技规划文本的研究主题为热点研究前沿主题。②若其非热点研究前沿主题且具有增长趋势，则判断该 NNI 科技规划文本的研究主题为新兴研究前沿主题。③若只具有一个相似或无相似的 NSF 基金项目的研究主题，则判断该 NNI 科技规划文本的研究主题为潜在研究前沿主题。得到的具体结果如下：

1）热点研究前沿主题：识别出的热点研究前沿主题总计 21 个，年份分布较为均匀。热点研究前沿主题主要集中于传感器、材料制造、半导体、聚合物、物理化学性能等方面，举例说明，例如研究热点前沿主题 2009_goal_0 的内容为：“这些机构的共同努力在几个纳米材料项目将取得重大进展，包括定向碳纳米管的生长；这些项目在平板显示器，化学传感器和生物传感器的电极以及基于有机分子的结构化表面组装的纳米级器件的优化等领域有着长期的应用。”与该主题具有高相似度的几个主题中 2009_topic_2、2013_topic_3 为在化学传感器方面的相关研究，2012_topic_4、2014_topic_1、2017_topic_2 为在生物传感器方面的相关研究，相应的 Num_nor，FA_nor，FT_nor 都大于平均数，有较大的研究意义与研究价值。

2）新兴研究前沿主题：识别出新兴研究前沿主题共计 10 个，其中 2010 年和 2012 年分布较多，2008 年、2014 年、2016 年无新兴研究前沿主题分布。新兴研究前沿主题，其对应的高相似度主题主要集中于能源、电池、环境等相关研究方面，例如：新兴研究前沿主题 2010_

goal_0 的内容为：“为提高太阳能电池的效率，纳入碳纳米管作为标准 3 层太阳能电池的界面层。”与该主题相似度较高的主题只有一个，为 2017_goal_0，通过回溯原文得到其主要研究内容为提高太阳能电池的转化效率以及降低太阳能电池相关材料的成本，且 Num_nor，FA_nor，FT_nor 指数皆小于平均数，说明其相关研究目前较少、研究价值与研究意义还未被重视。

3）潜在研究前沿主题：识别出的潜在研究前沿主题总计 27 个，其中 2011 年、2014 年、2017 年分布较多，2009 年无潜在研究前沿主题分布。潜在研究前沿主题与热点研究前沿主题不同，其研究方面极为离散，例如：潜在研究前沿主体 2011_goal_1 的内容为：“工业合作伙伴继续致力于开发多模式碳纳米管传感平台。”潜在研究前沿主体 2011_goal_7 内容为：“将加速在碳纳米管风险评估领域的活动；开发出改进的方法来检测和测量大批量商业化的纳米材料，如纳米管和金属氧化物；评估呼吸保护的有效性；并开始对缓解人类暴露和环境释放的工程控制进行定量评估。”通过相似度热力图分析，2011_goal_1、2011_goal_7 都无相似的 NSF 基金项目研究主题，所以无法进行研究前沿探测评价。

## 4 结束语

本文使用科技规划文本和基金项目数据两种数据源，分别通过以触发词库为基础的规则匹配抽取和主题识别模型，识别出研究主题，综合两个数据源的结果，进行相似度计算，结合项目数、资助时长与资助强度构造研究前沿主题识别模型，识别出热点研究前沿主题、新兴研究前沿主题和潜在研究前沿主题，以碳纳米管为例进行了实证研究，实验结果证明，该方法可以有效的识别出碳纳米管领域的研究前沿。本文对科技规划文本的识别结果的最小单位为句子，没有以词为单位更能突出研究主题的内容，识别粒度较为宏观，接下来的研究将会以词为单位对科技规划文本进行研究，并将识别出的结果与本文结果进行对比，互相补充，以期更好的对研究前沿主题进行解读。□

## 参考文献


[1] 白春礼．把握新科技革命与产业革命机遇以创新驱动塑造引领型发展 [J]．时事报告（党委中心组学习），2017（5）：35-49.

[2] 周群，周秋菊，冷伏海．基于科技媒体视角的研究前沿识别方法研究与实证 [J]．现代情报，2018（2）：62-67.

[3] 白如江，冷伏海，廖君华．科学研究前沿探测主要方法比较与发展趋势研究 [J]．情报理论与实践，2017，40（5）：33-38.

[4] PRICE D J. Networks of scientific papers [J]. Science，1965，149（3683）：510-515.

[5] 杨国立，李品，刘竟．我国图书馆学研究知识图谱分析 [J]．国家图书馆学刊，2012，21（1）：52-59.

[6] 孙震．基于科学论文多源数据的研究前沿集成识别模型研究 [J]．情报杂志，2016，35（8）：95-100.

[7] 陈仕吉．科学研究前沿探测方法综述 [J]．现代图书情报技术，2009，25（9）：28-33.

[8] 佚名．硕士研究生论文文摘 [J]．现代图书情报技术，2007（10）：91-92.

[9] 许晓阳，郑彦宁，刘志辉．论文和专利相结合的研究前沿识别方法研究 [J]．图书情报工作，2016，60（24）：97-106.

[10] 郑彦宁，许晓阳，刘志辉．基于关键词共现的研究前沿识别方法研究 [J]．图书情报工作，2016，60（4）：85-92.

[11] 侯剑华，刘则渊．纳米技术研究前沿及其演化的可视化分析 [J]．科学学与科学技术管理，2009，30（5）：23-30.

[12] 白如江，冷伏海，廖君华．一种基于科技规划文本的研究前沿主题地图构建方法 [J]．图书情报工作，2017，61（23）：114-121.

[13] 刘自强，王效岳，白如江．基于时间序列模型的研究热点分析预测方法研究 [J]．情报理论与实践，2016，39（5）：27-33.

[14] 王效岳，刘自强，白如江，等．基于基金项目数据的研究前沿主题探测方法 [J]．图书情报工作，2017，61（13）：87-98.

[15] 白如江，冷伏海，廖君华．一种基于多数据源主题对比的科学研究前沿识别方法 [J]．情报理论与实践，2017，40（8）：43-48.

[16] 冷伏海，赵庆峰，周秋菊．中美科研实力比较研究：基于《2015 研究前沿》的分析 [J]．中国科学基金，2016（1）：8-19.

[17] 冷伏海，孙震，周秋菊．《2015 研究前沿》报告的研制实践与相关探讨 [J]．智库理论与实践，2016，1（2）：79-87.

[18] The stanford natural language toolkit [EB/OL]. http://nlp.stanford.edu/.

[19] 冷伏海，赵庆峰，周秋菊．中美科研实力比较研究：基于《2016 研究前沿》的分析 [J]．中国科学基金，2017（1）：48-65.



作者简介：周彦廷（ORCID：0000-0001-7624-2637），男，1994 年生，硕士生。研究方向：文本挖掘与情报分析。白如江（ORCID：0000-0003-3822-8484，通讯作者），男，1979 年生，博士，副研究馆员。研究方向：文本数据挖掘与科技情报。王效岳（ORCID：0000-0002-7100-7758），男，1961 年生，博士，教授。研究方向：数据挖掘与信息处理技术。



作者贡献声明：周彦廷，数据收集分析和论文撰写。白如江，拟定研究命题和思路设计。王效岳，论文框架确定和论文细节修改。

录用日期：2018－12－11