DOI：10.13998/j.cnki.issn1002-1248.2019.12.12-1085

# 基于文本语义理解的学科发展趋势分析

余　丽[1,2]

（1. 中国科学院文献情报中心，北京　100190；2. 资源与环境信息系统国家重点实验室，北京　100101）

**摘　要：**［目的 / 意义］学术论文是科技创新发展的重要战略资源，是反映学科研究动态的一手资料；为后续研究者提供了宝贵的方法论和创新基础。目前，学术论文的知识组织还缺乏细粒度知识的结构化描述，阻碍了科技情报服务向计算化和精准化的转型升级。［方法 / 过程］首先提出一种深入文本内容的语义分析框架，半自动化从论文摘要中识别出“研究主题”和“关键技术”；然后设计了一种短语级多层次聚类方法，水平方向上的聚类融合了同义词语，垂直方向上的聚类构建了层次关系；最后以地理信息科学领域的代表性期刊论文摘要为实验数据，运用文献计量分析方法，分析了地理信息科学领域近 10 年的热点研究主题和关键技术，及其随时间发展的脉络。［结果 / 结论］研究方法可为面向文本内容理解的情报分析提供算法与数据支撑。

**关键词：**人工智能；语义标注；神经网络；短语聚类；文献计量分析

**中图分类号：**G251　　**文献标识码：**A　　**文章编号：**1002-1248（2020）03-0029-08

**引用本文：**余丽. 基于文本语义理解的学科发展趋势分析[J]. 农业图书情报学报, 2020, 32(3):29-36.

## Discipline Development Trend Analysis based on Text Semantic Understanding

YU Li

(1. National Science Library, Chinese Academy of Sciences, Beijing 100190;

2. State Key laboratory of Resources and Environmental Information System, Beijing 100101)

**Abstract:** [Purpose / Significance] Academic papers are the important strategic resources for the development of scientific and technological innovation. They are also the primary data that reflect the research trends of one subject, which provide the valuable methodological and innovative basis for the follow-up researchers. Recently, the knowledge organization of academic papers still lack of the fine-grained knowledge, which hinders the upgrading of scientific and technological information services to computerization and precision. [Method / Process] Firstly, this paper provides a framework of analyzing the semantic of article content: the "research topics" and "key technologies" are extracted from papers by using a semi-automatic model. Secondly, a multi-level clustering method for phrases are

收稿日期：2019-12-12

基金项目：国家自然科学基金青年基金项目“中文网络文本的地理实体语义关系标注与评价”（项目编号：41801320）；资源与环境信息系统国家重点实验室开放基金

作者简介：余丽（ORCID：0000-0002-4374-8743）（1986-），女，博士，馆员，研究方向：知识图谱与文献计量研究。

DOI：10.13998/j.cnki.issn1002-1248.2019.12.12-1085

designed. The synonymous phrases are merged by clustering in the horizontal direction, and the hierarchical relations are built by clustering in the vertical direction. Finally, the experiments are carried out by using the massive abstracts from the core journals in the discipline of geographic information science. Based on the bibliometric analysis, we analyzed the top N of "research topics" and "key technologies", and their development trajectories over time. [Results / Conclusions] The proposed method can provide technologies and datasets for the intelligent service of the scientific and technological information.

**Keywords:** artificial intelligence; semantic annotation; neural network; phrase clustering; bibliometric analysis

## 1 引言

学术论文是科研人员了解、发表和传播科研成果的主要平台。在开展一项具体的科研工作之前，研究者们会事先从海量文献数据库中搜索出感兴趣的可能相关的论文，逐篇阅读后，凝练出关键科学问题，再进一步追踪该问题的国内外研究现状，从而确定自己的研究内容[1]。然而，每年都有大量的学术论文发表。2018 中国科技论文统计结果显示，全球平均每年在地理信息科学相关领域（地学、农业科学、空间科学、环境与生态学、计算机科学）有超过 18 万篇论文被发表[2]。事实上，人类的阅读能力几乎是不变的。2012 年美国科学家估计，人类平均每年只能阅读 264 篇论文[3]。可见，科研人员的文献追踪能力很难跟上井喷式的论文增长速度。如何快速掌握该领域的现状，定位前沿研究问题，追踪先进技术，已经成为每位研究者开展科学研究面临的首要难题。

传统的学术搜索系统一般是基于论文元数据，例如以标题、关键词、作者等为检索条件，无法面向指定的研究问题全面准确地搜集研究现状，难以满足研究者实际的检索需求：指定研究方向的关键科学问题是什么？核心技术方法是什么？某个问题的研究水平如何？因此，学术搜索系统需要深入文本内容进行语义理解，才能实现专业化、精准化的信息检索与知识服务。目前，国际上已经涌现出多个深入内容理解的学术搜索系统。例如，Nano[4]是目前全球最大的纳米科技知识服务平台，能够基于检索结果进行分类与汇总，为用户提供结构化的信息；Elsevier 提出语义出版思想[5]，从论文中抽取药物名称、材料属性、化学品等，打造深入内容的问题解决方案；科睿唯安建立的全球药物研发信息平台 Cortellis[6]，用户能够通过搜索化学结构式、药物靶标、临床试验等来获取相关的药物研发信息,支持制药公司进行药物发现与临床试验，为实验室到诊所提供了更可信的决策依据。大数据与人工智能时代，深入文本内容的语义理解技术已成为追踪研究动态、了解学科发展、认知科学问题的制胜法宝。

笔者将探讨如何从学术论文中识别领域相关的知识，用于扩展学科发展趋势分析的维度。从论文中抽取知识，普遍采用基于模式的方法[7]。该方法借助于词法、句法或依赖模式，无需标注语料库。然而，有限的预定义模式无法覆盖多样化的自然语言表达形式，方法的精度虽高但召回率低。为了提升基于模式方法的召回率，自动模式挖掘方法[8]应运而生。例如经典的 Bootstrapping 技术[9]，基于人工定义的少量模式，通过迭代挖掘，自动生成大规模的模式库。然而，Bootstrapping 存在“语义漂移”问题，如何有效避免多次迭代过程中引入的错误模式成为研究难点[10]。另一方面，基于假设“语义相似的知识出现在相似的语境中”，分布式方法将文本映射到低维稠密的向量空间，通过分类或者聚类将语义相似的知识划分到同一个簇中[11]。基于分布的方法放松了模式匹配的限制，扩大了方法的适用范围。分布式方法中，基于监督机器学习模型的分

类方法[12]，需要高质量大规模的标注语料来优化模型的性能，然而构建标注语料的成本极高。目前，大量的研究集中于无监督的知识抽取技术[13]：通过各种聚类算法，将相同语义的知识自动划分到同一个簇中。

笔者将从论文摘要中自动识别出两类领域知识：研究主题和关键技术。研究主题定义为论文中讨论和解决的问题，包括所有与“Problem、Task、Question”等概念相关的短语，例如“Animated Mapping、Block Correction、Border Matching”等。关键技术定义为论文中解决问题时使用的手段，包括所有与“Method、Model、Technique”等概念相关的短语，例如“Digital Elevation Model、Entity Relationship Approach、Entropy Coding”等。

笔者首先设计一种半监督领域知识抽取框架，从论文摘要中自动识别“研究主题”和“关键技术”。然后通过统计分析与人工校验，运用频数统计、共现矩阵和关联矩阵，来揭示近 10 年来地理信息科学领域研究主题的变化趋势与关键技术的更新规律，为科研人员掌握学科的发展态势、追踪领域研究热点、遴选科研课题提供便捷的技术手段。

## 2 数据来源

笔者使用的实验数据，首先由领域专家遴选出地理信息科学领域的 8 种代表性国际学术期刊，再以期刊名称为查询条件从 NSTL 成员单位加工的印本数据中检索论文，提取“作者、机构、发表时间、期刊名称、标题、摘要、关键词”等元数据，共计 3 653 条记录，经统计时间跨度为 2008 年—2018 年。各期刊的论文年度发表情况如表 1 所示。

## 3 研究方法

面向科技文献语义理解，笔者设计的半监督领域知识抽取框架包括 3 个模块：构建标注语料库、基于神经网络模型的语义标注、短语聚类。基于海量的科技文献摘要，通过上述 3 个模块，构建领域的研究主题分类树和关键技术分类树，为学科发展趋势分析提供精细尺度的数据资源。

### 3.1 建立标注语料库

监督的机器学习模型能有效识别出文本蕴含的知识，但是需要大规模的标注语料库来训练模型的参数。笔者采用[14]提出的基于自动回标的语料库构建方法，半自动化建立“研究主题—关键技术”的大规模标注语料库。

如图 1 所示：首先从百度文库中下载地理信息科学的领域词表[15]，同时补充论文的关键词，再通过人工分类，形成“研究主题”词库和“关键技术”词库；然后将上述的词库作为种子，利用基于评价模型的改进 Bootstrapping 方法迭代挖掘新的研究主题和关键技术，直到词库规模不再增大；最后，基于字典匹配法在论文摘要中查找词库中的实例，生成“研究主题—关键技术”的标注语料库。

表 1 2008 年—2018 年地球信息科学领域 8 种国际期刊论文发表数量

| 期刊名称 | 发表时段 | 论文数/篇 |
|---|---|---|
| Annals of GIS | 2014 年—2016 年 | 80 |
| Computers, Environment and Urban Systems | 2009 年—2018 年 | 644 |
| Cartography and Geographic Information Science | 2009 年—2018 年 | 336 |
| Cartographic Journal | 2009 年—2018 年 | 319 |
| Geoinformatica | 2009 年—2018 年 | 257 |
| International Journal of Digital Earth | 2008 年—2018 年 | 462 |
| International Journal of Geographical Information Science | 2008 年—2018 年 | 1 124 |
| Transaction in GIS | 2009 年—2017 年 | 431 |

DOI：10.13998/j.cnki.issn1002-1248.2019.12.12-1085

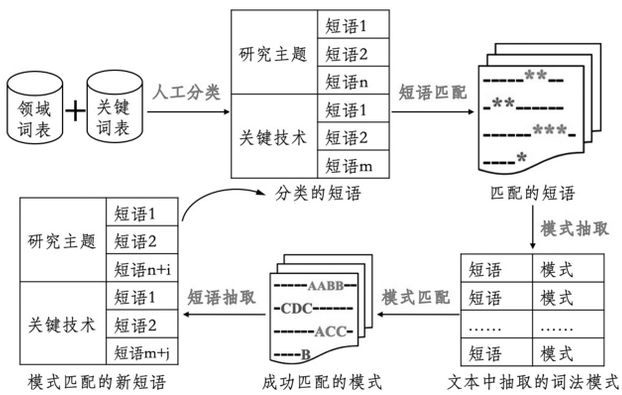


图 1　半自动构建标注语料库

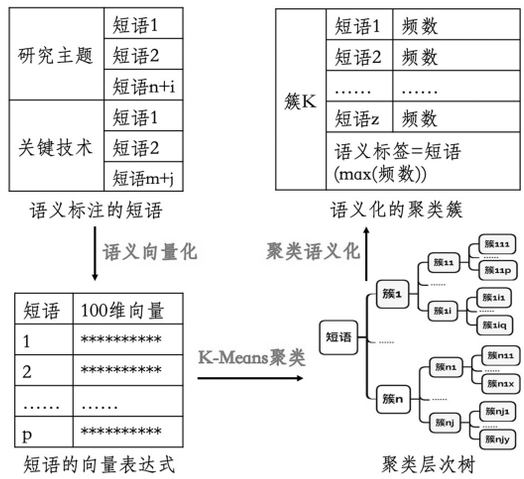


图 2　短语多层次聚类分析

## 3.2　基于神经网络模型的语义标注

神经网络模型使用低维稠密的向量空间来表达离散的语义，克服了传统特征表示模型的局限性[16]。其中，长短时记忆模型（Long-Short Term Memory，LSTM）因其有效避免了“梯度消失”问题，具有长时间范围内的记忆功能，被广泛应用于自然语言处理任务。同时，统计机器学习模型条件随机场（Conditional Random Field，CRF）能够从训练数据中学习到上下文的约束规则，保证了预测结果的合理性。笔者采用[17]提出的面向细粒度知识元抽取的深度学习方法，基于 LSTM-CRF 的神经网络模型学习框架，预测论文摘要中蕴含的“研究主题”和“关键技术”，以挖掘出标注语料库中未覆盖的新实例。

## 3.3　短语聚类

从论文摘要中识别出的“研究主题”和“关键技术”包含很多噪声。为了聚合语义相似的短语并删除无意义的短语，需要对上述的语义标注结果进行聚类分析。如图 2 所示：（1）笔者利用[18]提出的 Doc2vec 机器学习模型，将识别出的“研究主题”和“关键技术”短语转换为向量表达式。在选择最优向量维度时，分别设置为 50、100、200，再基于分布式表示学习模型的公开测试数据集[19]，决定了 Doc2vec 模型效果最佳的向量维度为 100。（2）使用 Cosine 函数构建短语的相似度矩阵。为了筛选出高相似性的短语对，笔者仅保留 Cosine 相似值大于 0.9 的短语对。（3）重复使用 K-Means 聚类算法得到一层或多层聚类结果。由于笔者仅关注排名前 10 的“研究主题”和“关键技术”，默认设置每层的聚类总数为 20。既能够为聚类结果的语义标注提供足够的候选集，又不会过多增加人工判读的成本。（4）基于词频统计结果，人工判读每个聚类簇的语义标签，并删除无意义的聚类簇。

# 4　实证分析

## 4.1　语义标注结果

笔者从表 1 所述的 3 653 篇论文摘要中，通过建立标注语料库、基于神经网络模型的语义标注、短语聚类 3 个步骤，自动识别出研究主题和关键技术。考虑到目前尚缺乏一个权威的标注有研究主题和关键技术的大规模语料库，笔者采用了人工评价来检验方法的性能。首先，分别针对“研究主题”和“关键技术”两种知识元，随机抽样 100 个结果，共计 200 个样本组成评价集合。然后，分别由 2 名图书馆学的硕士研究生同时对每一个样本进行判读，识别正确的样本标注为 1，识别错误的样本标注为 0。最后，两名同学标注完成后得到两份评价结果，如果两人对同一个样本的判读结果不一致，则两人同一名博士生讨论后得到最终判读结

果。基于该份人工标注的评价数据集，针对 200 个样本，计算预测结果的正确率。

语义标注及评价结果如表 2 所示。方法将原始词表提供的百余个领域词汇扩展到上千个，为快速获取领域知识提供了一种便捷手段。而且，研究主题和关键技术的平均识别正确率为 81%，能够为深入文本内容的学科发展趋势分析提供较为优质的数据资源。虽然标注语料中“研究主题”的数量仅为“关键技术”的一半，但语义标注模型识别出了更多的“关键技术”且正确率更高。分析采样数据发现，描述“关键技术”的词汇较为固定，语义模糊性相对较弱，使得 Bootstrapping 方法在迭代过程中能更快速捕捉到大多数实例的词法表达模式，缓解了“语义漂移”现象。

## 4.2 短语聚类结果

通过短语聚类，语义相似的实体划分到同一簇中。根据每个实体在论文摘要中出现的总次数，计算出 2008 年—2018 年累计排名前 10 的“研究主题”和“关键技术”。按照频数由高到低，前 10 名研究主题分别是：空间分析、城市模型、数字制图、土地使用、交通模型、主题地图、城市增长、数据管理、数据检索、航空摄影测量。前 10 名关键技术分别是：数字高程模型、元胞自动机、神经网络、遗传算法、回归模型、K 近邻算法、蒙特卡罗算法、模拟退火、支持向量机、随机森林。

统计结果显示，在地理信息科学领域，“空间分析”和“城市模型”是两个最热门的研究主题，其数量总和几乎占据了总量的一半；“数字高程模型”模型因其在地面建模过程中发挥着重要作用，影响范围极广，位于关键技术榜首；传统的“元胞自动机”模型则紧随其后；近几年逐步兴起的“神经网络”也跻身关键技术前 3 强之列。

## 4.3 研究主题和关键技术的时序分析

为了揭示研究主题和关键技术的发展规律与趋势，笔者以年份为横坐标，实体在当年所有论文中所占的频数百分比为纵轴（反映了研究主题或关键技术在当年的研究热度），绘制时间变化曲线。通过人工综合分析所有实体的频数百分比，选择 3% 为阈值，将研究趋势划分为 3 类：（1）爬升期：最近 3 年的研究热度持续上升，且极值差大于 3%；（2）衰落期：最近 3 年的研究热度持续下降，且极值差大于 3%；（3）稳定期，连续 3 年或以上的小幅波动，且极值差不超过 3%。

图 3 展示了前 10 名研究主题的热度随时间的变化。依据上述的判断条件，“空间分析”经过 2012 年的狂热之后虽有冷却，但目前正处于爬升期，有望成为新的研究热点。“数字制图”经过长期的稳定发展在 2015 年之后研究热度大幅下滑，陷入衰落期。剩下的 8 个研究主题的热度略有小幅波动，保持稳步发展。

图 4 展示了前 10 名关键技术的热度随时间的变化。依据上述的判断条件，“K 近邻算法”经过两次大起大落之后，自 2014 年开始其研究热度稳步攀升，成为研究热点。“数字高程模型”虽然早期非常火热，但近些年的研究热度骤然下降，目前正陷入衰落期。“遗传算法”、“回归模型”和“蒙特卡罗算法”的研究已进入稳定期。值得注意的是，传统的智能优化模型“元胞自动机”、“模拟退火”、新兴的机器学习方法“神经网络”、“支持向量机”、“随机森林”都存在多个波峰，曲线呈现出剧烈的抖动，其发展局势尚不明朗，这也预示着更多的研究机会。

## 4.4 研究主题与关键技术的相关关系

为了进一步探索“研究主题”与“关键技术”

表 2 语义标注结果

| 短语类型 | 原始词表数量/个 | 标注语料数量/个 | 语义标注模型识别的数量/个 | 识别正确率/% |
|---|---|---|---|---|
| 研究主题 | 83 | 667 | 6 989 | 83 |
| 关键技术 | 224 | 1 015 | 4 361 | 79 |

DOI：10.13998/j.cnki.issn1002-1248.2019.12.12-1085

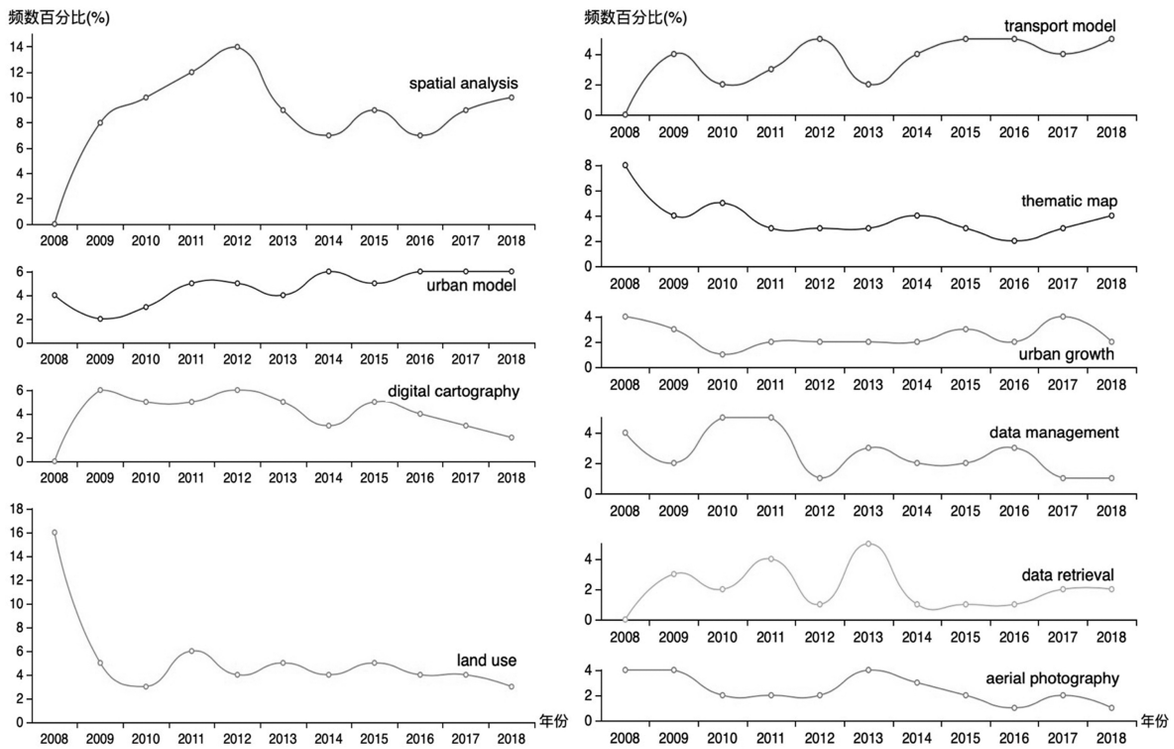


图 3　Top10 研究主题的时间变化曲线

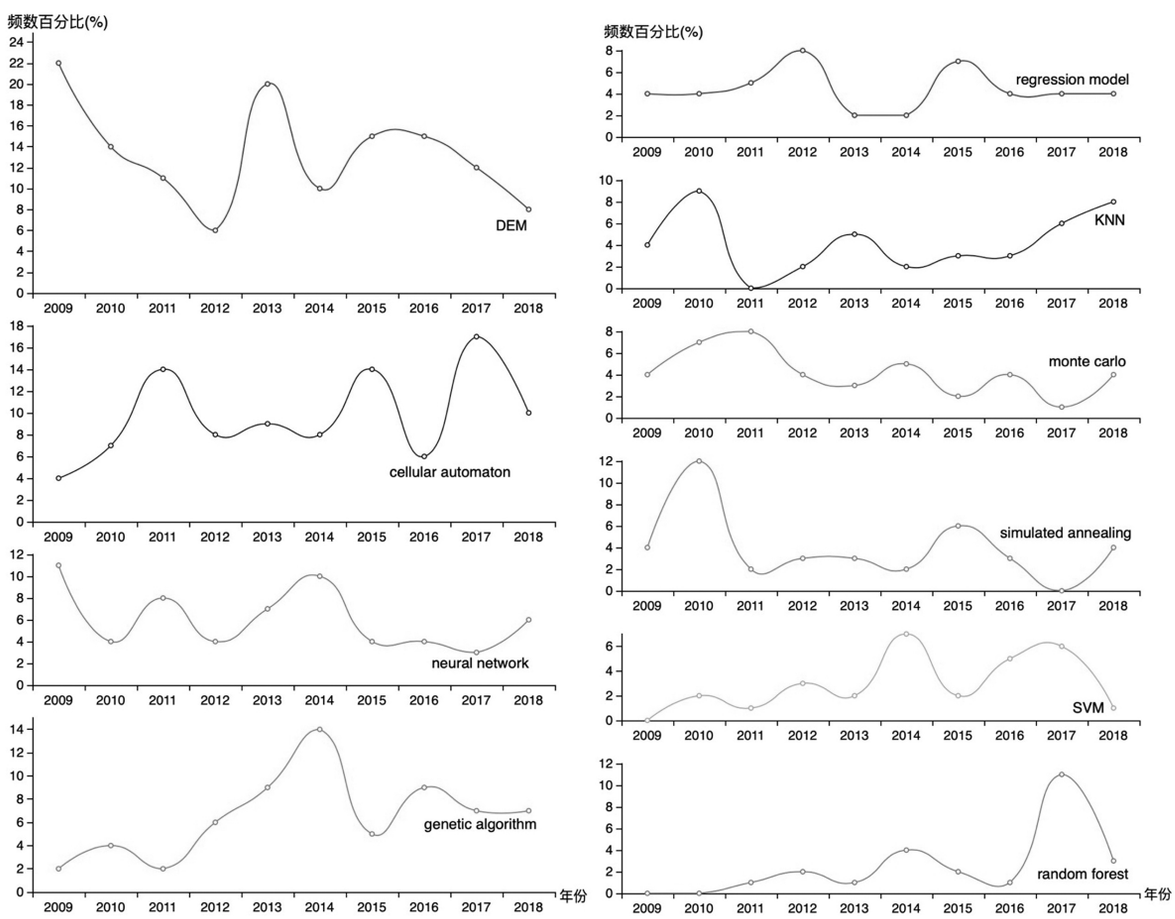


图 4　Top10 关键技术的时间变化曲线

之间的相互作用，笔者基于假设“同一篇论文摘要中识别出的研究主题和关键技术之间存在着共现关系”，构建了“主题—技术”共现矩阵。基于共现频次，笔者建立了排名前 3 的关键技术与研究主题之间的关联关系，且利用二次聚类分析展示子主题的分布情况，按照总频数由大到小排列，如表 3 所示。

笔者发现，不同的关键技术可用来解决相同的研究主题，例如“元胞自动机”、“神经网络”和“数字高程模型”均可以用于“空间分析”，这为科研工作提供了同一问题的不同解决方案。值得注意的是，仅使用“数字高程模型”研究的特有主题包括：“图像分割”、“太阳辐射”、“空间查询”、“空间表达”；仅使用“元胞自动机”研究的特有主题包括：“城市仿真”、“空间模式”、“空间变化”、“空间数据挖掘”；仅使用“神经网络”研究的特有主题包括：“城市环境”、“空间分配”。这为科研工作者遴选创新性研究课题提供了新思路，例如通过“神经网络”研究“太阳辐射”。

表 4 展示了排名前 3 的研究主题与关键技术之间的关系。结果显示，研究主题“空间分析”与“城市模型”使用的主流技术非常相似，原因是两者的研究主题具有高度的相关性。但是，“空间分析”更多使用机器学习模型，例如“回归模型”、“K 近邻”、“最大熵”、“决策树”；而“城市模型”中更多地使用智能优化算法，例如“遗传算法”、“模拟退火算法”、“粒子群优化算法”、“蚁群算法”等；研究主题“数字制图”则是兼容并包，综合使用机器学习模型、智能优化算法、计算几何、图理论等。

## 5 结论

笔者基于深入文本内容的语义分析框架，通过

表 3 top3 的关键技术对应的研究主题

| 关键技术 | 研究主题 | 子研究主题 |
|---|---|---|
| DEM | aerial photography | airborne lidar, image segmentation, light detection, satellite image, solar radiation |
| | spatial analysis | spatial distribution, spatial relation, spatial query, spatial interpolation, spatial representation |
| | facility network | geosensor network, drainage network, road network, river network |
| cellular automaton | urban model | urban growth, urban simulation, urban planning, urban modelling |
| | land use | land classification, land planning |
| | spatial analysis | spatial distribution, spatial pattern, spatial variation, spatial clustering, spatial data mining |
| neural network | urban model | urban growth, urban environment, urban modelling, urban planning |
| | spatial analysis | spatial distribution, spatial relation, spatial allocation, spatial clustering, spatial interpolation |
| | aerial photography | satellite image, airborne lidar, light detection |

表 4 top3 的研究主题使用的关键技术

| 研究主题 | 关键技术 |
|---|---|
| spatial analysis | regression model, cellular automation, DEM, neural network, monte cario, kernel density estimation, SVM, KNN, maximum entropy, decision tree |
| urban model | cellular automation, genetic algorithm, neural network, DEM, monte cario, simulated annealing, SVM, particle swarm optimization, random forest, ant colony |
| digital cartography | DEM, regression model, neural network, greedy algorithm, computational geometry, simulated annealing, monte carlo, SVM, rough set, Voronoi |

DOI：10.13998/j.cnki.issn1002-1248.2019.12.12-1085

语料库构建、语义标注、短语聚类3个步骤，运用频数统计、共现矩阵和关联矩阵等分析方法，对2008年—2018年地球信息科学领域8种国际核心期刊论文进行了科学计量分析，挖掘出近10年地理信息科学领域的研究主题和关键技术。然而，还存在如下不足：（1）由于实验数据仅包含8种领域核心期刊，无法全面覆盖作者的科研成果。今后工作中，需要增加数据源的种类和数量（论文、项目、专利等），以便尽可能真实地反映出学科整体的科研水平。（2）人工判读获取短语聚类的语义标签，耗时费力且结果容易受人的主观意识的影响，会导致后续科学计量分析的偏见。今后工作中，需要研究自动化的聚类算法，例如通过关系抽取技术自动构建短语之间的上下位关系[20]。

## 参考文献：


[1] H. D. Ribaupierre and G. Falquet. Extracting Discourse Elements and Annotating Scientific Documents Using the SciAnnotDoc Model: A Use Case in Gender Documents [J]. International Journal on Digital Libraries,2018,19(2-3):271-286.

[2] 中国科技论文统计结果—第二卷:中国国际科技论文产出状况[R],中国科学技术信息研究所,2018.

[3] R. V. Noorden, Scientists May be Reaching a Peak in Reading Habits [R],Nature|News,3 Feb,2014.

[4] What is Nano? Nature Nanotechnology[R],2016,11(8):575.

[5] H. Bell, R. Kwakkelaar. Making Research Easier and Smarter with Semantic Publishing Technologies[R],15 September,2014.

[6] Clarivate Analytics Introduces Next-generation Pre-clinical Drug Research Platform with Cortellis Drug Discovery Intelligence[R],Clarivate Analytics Plc,1 October,2019.

[7] R. Snow, D. Jurafsky, Y. Ng Andrew. Learning Systactic Patterns for Automatic Hypernym Discovery [C]. Advances in Neural Information Processing Systems 17 (NIPS 2004),2004.

[8] M. Singh, S. Dan, S. Agarwal, et, al. App TechMiner: Mining Applications and Techniques from Scientific Articles [C]. Proceedings of the 6th International Workshop on Mining Scientific Publications. 2017:1-8.

[9] T. Siddiqui, X. Ren, A. Parameswaran, et al. FacetGist: Collective Extraction of Document Facets in Large Technical Corpora [C]. Proceedings of the 25th ACM International on Conference on Information and Knowledge Management,2016:871-880.

[10] V.T. Phi, J. Santoso, M. Shimbo, et al. Ranking-Based Automatic Seed Selection and Noise Reduction for Weakly Supervised Relation Extraction [C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,2018(2):89-95.

[11] A. Gupta, R. Lebret, H. Harkous. Taxonomy Induction Using Hypernym Subsequences[C]. Proceedings of the 2017 ACM Conference on information and Knowledge Management,2017:1329-1338.

[12] W. Ammar, M. Peters, C. Bhagavatula, et. al. The AI2 System at SemEval-2017 Task 10 (ScienceIE): Semi-supervised End-to-End Entity and Relation Extraction[C]. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017),2017:592-596.

[13] J. Shen, Z. Wu, D. Lei, et. al. HiExpan: Task-Guided Taxonomy Construction by Hierarchical Tree Expansion[C]. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,2018:2180-2189.

[14] 王姬卜,陆锋,吴升等.基于自动回标的地理实体关系语料库构建方法[J].地球信息科学学报,2018,20(7):871-879.

[15] GIS专业词汇表[EB/OL].[2010-06-07].https://wenku.baidu.com/view/327c6c0216fc700abb68fc42.html.

[16] 刘知远,孙茂松,林衍凯等.知识表示学习研究进展[J].计算机研究与发展,2016,53(2):247-261.

[17] 余丽,钱力,付常雷等.基于深度学习的文本中细粒度知识元抽取方法研究[J].数据分析与知识发现,2019,25(1):38-45.

[18] Q.Le,M.Mikolov.Distributed Representations Sentences and Documents[C]. Proceedings of the 31st International Conference on Machine Learning,2014:1188-1196.

[19] text8[EB/OL].[2006-06-09].http://mattmahoney.NET/dc/text8.zip.

[20] 汪诚愚,何晓丰,宫学庆等.面向上下位关系预测的词嵌入投影模型[J].计算机学报,2019:1-17.