Hans Journal of Soil Science 土壤科学, 2022, 10(2), 78-90
Published Online April 2022 in Hans. http://www.hanspub.org/journal/hjss
https://doi.org/10.12677/hjss.2022.102011

Hans汉斯

基于文本挖掘技术识别土地管理领域未来方向的研究

董　浩[1,2,3,4,5]，孟婷婷[1,2,3,4,5]

[1]陕西省土地工程建设集团有限责任公司，陕西　西安
[2]陕西地建土地工程技术研究院有限责任公司，陕西　西安
[3]自然资源部退化及未利用土地整治工程重点实验室，陕西　西安
[4]陕西省土地整治工程技术研究中心，陕西　西安
[5]自然资源部土地工程技术创新中心，陕西　西安

收稿日期：2022年3月14日；录用日期：2022年4月17日；发布日期：2022年4月24日

摘　要

立足生态文明时代和新型城镇化建设背景，围绕国家经济社会高质量发展和治理体系现代化要求，企业、研究机构和政府部门积极寻求新兴趋势和问题的方法，这些趋势和问题可能会影响其未来的运营环境。本文利用文本挖掘技术来调查土地管理部门的未来信号。通过系统回顾有关文本挖掘来检测未来信号的文献后，本研究建议使用隐含狄利克雷分布模型来增强对未来信号的解释。这项研究的发现突出了与土地利益及其记录相关的广泛问题，确定了17个未来信号主题，从缓解气候变化和使用卫星图像进行数据收集到标准化和参与式土地整理。研究表明，在使用自动化过程时，区分弱信号与潜伏信号，较强信号和强信号是具有挑战性的。本研究总结了土地管理领域的当前论述，并指出了当前哪些主题正在蓬勃发展。

关键词

土地管理，地籍系统，未来信号，文本挖掘，隐含狄利克雷分布模型

Research on Identifying the Future Direction of Land Management Based on Text Mining Technology

Hao Dong[1,2,3,4,5], Tingting Meng[1,2,3,4,5]

[1]Shaanxi Provincial Land Engineering Construction Group Co., Ltd., Xi'an Shaanxi
[2]Institute of Land Engineering and Technology, Shaanxi Provincial Land Engineering Construction Group Co., Ltd., Xi'an Shaanxi
[3]Key Laboratory of Degraded and Unused Land Consolidation Engineering, Ministry of Natural Resources, Xi'an

文章引用: 董浩, 孟婷婷. 基于文本挖掘技术识别土地管理领域未来方向的研究[J]. 土壤科学, 2022, 10(2): 78-90.
DOI: 10.12677/hjss.2022.102011

Shaanxi
[4]Shaanxi Provincial Land Consolidation Engineering Technology Research Center, Xi'an Shaanxi
[5]Land Engineering Technology Innovation Center, Ministry of Natural Resources, Xi'an Shaanxi

Received: Mar. 14th, 2022; accepted: Apr. 17th, 2022; published: Apr. 24th, 2022

Abstract

Based on the background of the era of ecological civilization and new urbanization construction, and around the requirements of high-quality national economic and social development and modernization of the governance system, enterprises, research institutions and government departments are actively seeking ways to address emerging trends and issues that may affect their future operating environment. This paper uses text mining technology to investigate the future signal of Land Management Department. After a systematic review of the literature on text mining to detect future signals, this study suggests using the Latent Dirichlet allocation model to enhance the interpretation of future signals. The findings of the study highlighted a wide range of issues related to land interests and their records, and identified 17 future signal themes, ranging from climate change mitigation and the use of satellite imagery for data collection to standardization and participatory land consolidation. The research shows that it is challenging to distinguish weak signal from latent signal, strong signal from strong signal in the process of automation. This study summarizes the current discussion in the field of land management, and points out which topics are currently developing vigorously.

Keywords

Land Management, Cadastral Systems, Future Signal, Text Mining, Latent Dirichlet Allocation Model

1. 引言

土地管理是国家为调整土地关系，组织和监督土地的开发利用，保护和合理利用土地资源，而采取的行政、经济、法律和技术的综合性措施，是一个国家经济和社会发展的基础[1]。2020 年 3 月 12 日，国务院发布《国务院关于授权和委托用地审批权的决定》中明确强调赋予省级人民政府更大用地自主权。要求在严格保护耕地、节约集约用地的前提下，进一步深化“放管服”改革，改革土地管理制度。在最近 20 年中，国家在社会经济发展中的产业结构布局，以及新型城镇化建设进程对耕地数量变化产生了重要的影响[2]。土地在社会和人民中的意义和作用随着时间的推移已经发生了变化[3]。法律地籍领域模型的历史发展表明，人地关系的变化导致法律地籍领域模型作为土地管理系统核心组成部分的功能发生了匹配性的变化[4]。因此，为了使法律地籍领域模型今后也能满足国家土地管理信息系统建设的需要，必须注意新出现的问题和变化的驱动因素。

2011 年 8 月美国情报高级研究计划署(IARPA)启动了 FUSE (Foresight and Understanding from Scientific Exposition)项目，旨在通过数据驱动挖掘最新研究动态和热点，系统、持续、全面地对新兴技术进行评估，预测潜在的新兴技术。地平线扫描(Horizon scanning)是 FUSE 项目的主要应用场景[5]。地平线扫

描是指在组织或机构所处的政治、经济、社会、科技、生态环境中，系统而广泛地收集与未来问题、发展趋势、观念和事件有关的信息和证据(未来信号)，一种全面扫描挑战和机遇并对未来进行预测的方法[6]。地平线扫描主要有预警功能和创造性功能，预警功能有助于政府或组织决策者更早、更准确地预测新出现的问题，而创造性功能则激活新出现问题的产生[7] [8] [9]。国内外学者在进行地平线扫描研究中经常使用诸如弱信号[10] [11] [12]、新出现问题[13]、通配符[14]和大趋势[15]等术语。由于这些术语有许多定义，以及对他人定义的批评，因此术语可能会令人困惑。有时从批判的角度看待定义可能更容易：本研究可以定义一个概念不是什么，或者它与另一个概念有何不同。除了术语，地平线扫描的方法也是一个有争议的问题。一些学者支持使用参与式方法，如访谈回顾、德尔菲问卷或研讨会，而其他人则倾向于使用非参与式和(半)自动化方法，如搜索引擎和文本挖掘。陈美华和王延飞认为地平线扫描方法的选择取决于上下文和内容问题[6]。无论选择何种信号检测方法，除了大趋势之外，还是值得研究关注未来信号。如果前瞻性的工作和新兴的技术仅仅依赖于大趋势，那么存在很高的风险认为未来只随趋势发展。此外，感知特定领域的过去、现在和未来环境有助于将预期工作聚焦在相关问题上[16]。随着文本数据量的不断增加，探索文本挖掘技术的可能性也变得越来越重要。自动化的趋势或信号检测方法已经被使用，例如，在公安情报预警模式构建[15]，技术预测[16]，预测犯罪心理[17]等。不同的应用环境突出了文本挖掘技术的广泛可用性，唯一的先决条件是能够获得相关主题的大规模(数量多且质量好)文本数据。

因此，本研究的目的是通过使用文本挖掘工具来识别土地管理领域的未来信号。此外，本研究旨在测试在基于文本挖掘的未来信号检测过程中添加语义元素，以便将术语组而不是单个术语识别为未来信号，以减少信号解释的歧义和抽象。本研究使用科学论文的摘要作为学习材料，并展示出通过半自动信号检测过程，可以确定许多合理的未来信号主题。主题范围从面向技术的主题到环境和社会主题。因此，本研究的结果强调了应该以一种整体的方式来理解土地管理及其构建土地管理信息系统，而不仅仅是一个登记系统。

2. 文献综述

在这一部分，本研究回顾了关于未来信号的文献。重点在于弱信号，因为从理论上讲，未来的其他信号类型可以通过它来理解要么这种信号比弱信号更弱，超越了本感知范围且没有能力接收；要么弱信号已经增强为一种众所周知的信号或一种强信号，在本研究感知范围内，但就本研究自身的心智模式无法识别。下面本研究通过回顾未来研究信号检测是如何在文本挖掘应用中进行的。

2.1. 未来研究中的弱信号

首先，对微弱信号的概念进行了较为详细的阐述，即微弱信号的特征及其与未来趋势等问题的关系。此外，还简要讨论了通配符等其他问题。使用的术语很广泛；因此，本研究应该在本文中定义如何处理弱信号。本研究使用 Hiltunen 的定义，尽管有各种未来学家讨论定义[10]。根据她的说法，微弱的信号是新问题出现的第一个迹象，但是能见度很低。必须强调，弱信号不是愿景，而是现有问题。现有文献承认其他几个弱信号的定义，其中最早的定义是由 Ansoff 在 1970 年代提出的[18]。弱信号研究的困难在于对这一概念有各种各样的定义，而且在今后的研究中一直在进行辩论，讨论弱信号是否是一个独立的概念，还是新出现问题的同义词[18]或预警信号[16]。在以前的文献中，弱信号和通配符有时被用作同义词[19]。Kuosa 认为弱信号是未来发展的最模糊的信号之一，这些信号可能完全令人忽视，也可能暗示未来的变化。弱信号作为变化的早期预警，当与其他信号结合时，信号增强。这种对弱信号行为的解释符合 Hiltunen 的观点，他将弱信号描述为拼图游戏的一部分，对未来的变化有一个整体的看法[20]。弱信号的特征可以从六个方面进行探讨[21]。这些方面将弱信号描述为过渡现象、弱信号的持续时间、弱信号的客

观性和主观性、观察者对同一信号的各种解释方式、信号的增强以及与信号的接收者和分析者有关的问题。这与 Kuusi 和 Hiltunen 强调弱信号不需要解释器存在，它本身就是一种现象相一致。Hiltunen 描述了在弱信号为信号、对象和解释的情况下的三个维度。此外，在这种表现中，客体与解释的二重性也是存在的。这些维度可以在一个三维空间中呈现出来，空间中维度位于轴上。当信号离开原点时，信号强化并最终成为一个强大的信号。Kuusi 和 Hiltunen 讨论了未来信号的动态性及弱信号到强信号的意义。这种意义的形成过程与解释密切相关。因为，只有当弱信号发展成为强信号或有意义的未来信号时才足够可见以便观察。

2.2. 文本挖掘应用中的未来信号

利用文本挖掘工具识别趋势和信号的文献[22] [23] [24]虽然数量不多，但数量一直在上涨，它们利用文献中的出现频率一词来衡量未来信号的变化和强度。Yoon 提出了两个指标用于区分信号和未来标志的发布维度：1) 可见度(Degree of Visibility, DoV)，用于测量一组文献中定义的关键词的频率的程度，用作信号的代理；2) 扩散度(Degree of Diffusion, DoD)，用于测量与文献总数相关的每个关键词的文献频率，用作发布的代理。这两个指标把更多的权重通过时间权重系数将文献中多次提出的主题连接在一起[22]。Yoon 进一步建议，当可见度和扩散度指标分值与平均频率计数(分别为平均术语频率和平均文献频率)一起映射时，可以形成关键词涌现图和关键词问题图，这些图可用于从一组文献中检测未来的信号。旨在识别四种信号(如图 1)：1) 潜在信号，是具有低频率和低变化率的词 2) 弱信号，是具有低频率但变化率较高的词 3) 较强信号，是具有高于平均频率但变化率较低的词；4) 强信号，是具有很高的频率和变化率的词。

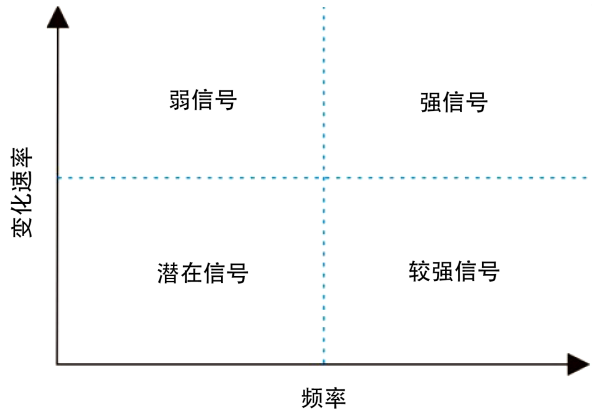

Figure 1. Future signal type recognition
图 1. 未来信号类型识别

从文本数据集中提取这类信息相对来说比较简单，但是如何解释(Hiltunen 框架中的第三个维度)新出现的问题和信号就更成问题了。正如 Lee 和 Park 研究指出，Yoon 提出的方法在这方面含有模糊性，因为关键词涌现图和关键词问题图不可避免地彼此不同，而且它们的比较和整合是一项主观任务。此外，结果是一个没有给定上下文的关键词列表，其中存在某种较高抽象性[24]。非常具体的关键词也造成问题，因为它们可能涉及个别事件或地点，因此不能被视为未来信号。因此，Lee 和 Park 考虑了上述缺点，他们提出在未来信号检测框架中引入语义分析。使用语义元素，目标是从一组词而不是单个词中找到意义，以便更好地理解未来信号。在文本挖掘应用程序中，关注主题而不是单个词是一种常见的做法[25]。主题建模算法是有助于发现文献集合主题的统计学方法，这些主题之间的联系以及它们随着时间的推移从原

始文本开始变化，这是当前研究的一种趋势。

3. 研究方法和数据来源

3.1. 研究方法

地平线扫描组织或机构所处的政治、经济、社会、科技、生态环境中，利用人类注意力或机器感知和采集分析系统广泛地收集与未来问题、发展趋势、观念和事件有关的信息和证据(未来信号)，按一定流程全谱地系统感知获取信息、存储分析处理信息、描绘出该领域实际全貌的过程。该过程从问题定义开始，在这种情况下，这是对土地管理领域的未来信号的识别，然后进行文本挖掘练习。在这一点上，应该强调的是，文本挖掘虽然是一种定量方法，但很少是线性且完全自动化的过程，所以也需要文本挖掘者采取行动和修改[26]。毕竟，文本挖掘与内容分析具有相似的功能，因为目标是从文本语料库中提取常见的主题和关键词，这可能需要基于手动选择添加和/或删除类别。执行文本分析需要建立一个文本挖掘框架，如图 2 所示。通常，应包括以下四个步骤。1) 是将文本导入计算环境；2) 是组织和构造导入的文本，以统一的方式访问它们；3) 是文本语料库整理和预处理，整理包括删除停用词，标记文本等。预处理的文本还必须转换为结构化的格式，以使分析成为可能。4) 是使用适当的方法进行分析，在文本分析部分完成之后，将结果整理成见解。在本研究中，分析包括三个主要步骤：关键词选择，关键词出现和问题分类的构建以及未来信号主题的识别。在执行前两个步骤时，本研究遵循先前的文献[22] [23] [24]，进行了少量修改。在第三步中，本研究将主题建模应用于将关键词集合分组在一起，以使信号更易于解释。因此，将主题建模包括在内，以应对先前框架中对未来信号解释的歧义。

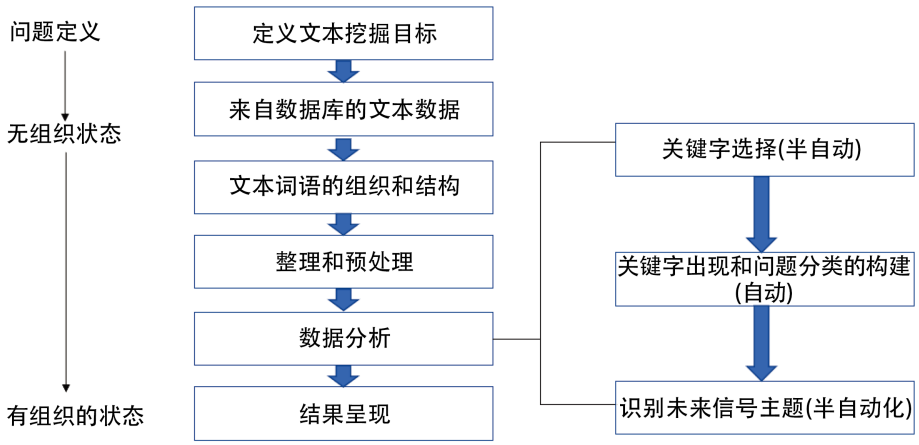

Figure 2. Summary of the study
图 2. 研究过程总结

3.2. 文本挖掘工具

在这里，本研究将简要介绍分析中使用的文本挖掘工具。在回顾以往文献的基础上，主要目标是 1) 识别术语和文献频率的时间变化，2) 研究术语的共现现象，并将其归类为合理的主题，并据此选择相应的工具。使用 r 语言统计软件进行分析。

3.2.1. 术语和文献频率

TF-IDF 索引是许多文本挖掘练习的最常用的数据结构。它通过将文献中一个词的出现频率与该词出现的文献的出现频率结合起来，表达了该词在一组文献中的重要性。因为在本研究中，本研究使用的是

摘要而不是完整的文献，使用 TD-IDF 索引可能会产生误差，因为文献中的词频在摘要中提供的信息较少。相反，本研究只是用下面的简单公式计算专业术语在总体词汇中的频率：

$$TF_x = \frac{x}{y} \tag{1}$$

其中：TF_x 表示术语在总体词汇中的频率，x 表示术语项数，y 表示收集文献的术语总数。

本研究很想知道术语 x 出现在文献中的频率。计算文献频率是发生术语的文献数量与文献总数之间的关系。

$$DF_x = \frac{M_x}{N} \tag{2}$$

其中：DF_x 表示含术语的文献数在文献总数中的频率，M_x 表示出现术语 x 的文献数，N 表示文献总数。

3.2.2. 可见度(*DoV*)和扩散度(*DoD*)

除了术语的频率，还有时间方面，即术语出现率的变化率在未来信号检测中起着关键作用。DoV_{ij} 和 DoD_{ij} 两种变化率在相关文献中已经确立了它们的地位。DoV_{ij} 和 DoD_{ij} 都是在周期 j 期间为项 i 计算的(总分是周期值的总和)。前者总结了一个术语在一段时间内的使用量，而后者则测量了一个术语在文献集合中的分布是如何随时间变化的。这些指标是根据以下方程式计算的：

$$DoV_{ij} = \left(\frac{TF_{ij}}{NN_j}\right) \times \left\{1 - tw \times (n - j)\right\} \tag{3}$$

$$DoD_{ij} = \left(\frac{DF_{ij}}{NN_j}\right) \times \left\{1 - tw \times (n - j)\right\} \tag{4}$$

其中 NN 表示文献总数，n 表示时间周期的长度，tw 表示预先确定的时间权重。

由于时间权重系数的影响，指标权重近期出现了比以往更重的术语。在这种情况下，这是直观的，因为在未来信号检测中，本研究并不特别热衷于识别过去的模式并根据它们做出解释。相反，本文的目的是检测那些正在新兴的、潜在的、有发展潜力的术语，以及那些正在从主流中消失的术语。

3.2.3. 主题建模

在本研究中，采用概率主题模型隐含狄利克雷分布模型(Latent Dirichlet Allocation, LDA)来寻找同时出现的术语集(主题集)。LDA 被描述为最简单的主题模型[26]。其逻辑是：词语库中的每篇文献都是主题的概率混合物，同样，每个主题也是术语的概率混合物。主题的数量由分析人员设置，因此 LDA 不会给出主题名称。因此，需要手动确定主题名称，通常是基于顶部的主题词[27]。

3.3. 数据收集与资料

本研究使用中国知网、万方、维普、Elservier、Springer、Scopus 等科学论文数据库作为科学论文摘要的来源。尽管完整的论文比摘要包含更多的信息，但仅使用摘要还是有很多优点的。首先，它们应该包含最重要和最简洁的关键词，使其成为识别未来信号的更相关的来源。其次，较小的文本词语库使分析速度更快。摘要能够反映整篇论文的内容，而不会给分析带来太多的“干扰”。本研究使用“土地管理”、“地籍系统”、“地籍管理系统”和“土地管理系统”等作为搜索关键词，并将搜索范围限于 2010 年 1 月 1 日至 2020 年 7 月 1 日期间出版或接受出版的文献。共检索到的文献总数为 6252 份。对于每篇文献，本研究能够识别标题、作者、关键词、期刊名称、出版年份和摘要。从图 3 可以看出，在过去的十年中，涉及土地管理问题的文章数量稳步增加。图 3 还显示了在这个数据集中，研究论文是主要的文

献类型，其次是会议论文类型。为了进一步验证所收集文件的相关性，本研究通过简单地计算关键词的总数来检查关键词(如表 1)。“土地管理”是最常见的关键词，其次是“土地规划”和“土地征收”。此外，更具体的关键词，如“土地利用”、“气候变化”、“土地管理领域模型”属于表中的中部。总的来说，最受欢迎的关键词与当前土地管理领域内的论述[28]是一致的。在分析之前，进行必要的预处理任务，比如去掉最通用的词。如“土地”、“管理”、“作者”、“出版”、“定性”、“基础”等等。

Table 1. The 20 most common keywords in the Dataset
表 1. 数据集中最常见的 20 个关键词

序号	关键词	数量	序号	关键词	数量
1	土地管理	621	11	三维城市模板的通用数据模型	125
2	土地规划	382	12	地理信息系统	123
3	土地征收	233	13	土地登记	123
4	土地使用权	204	14	土地科学	123
5	耕地保护	185	15	农业	123
6	土地所有权	155	16	新型城镇化	121
7	土地利用	143	17	土壤湿度	111
8	气候变化	136	18	城市规划	92
9	土地管理领域模型	132	19	遥感	90
10	土地整治	131	20	卫星	88

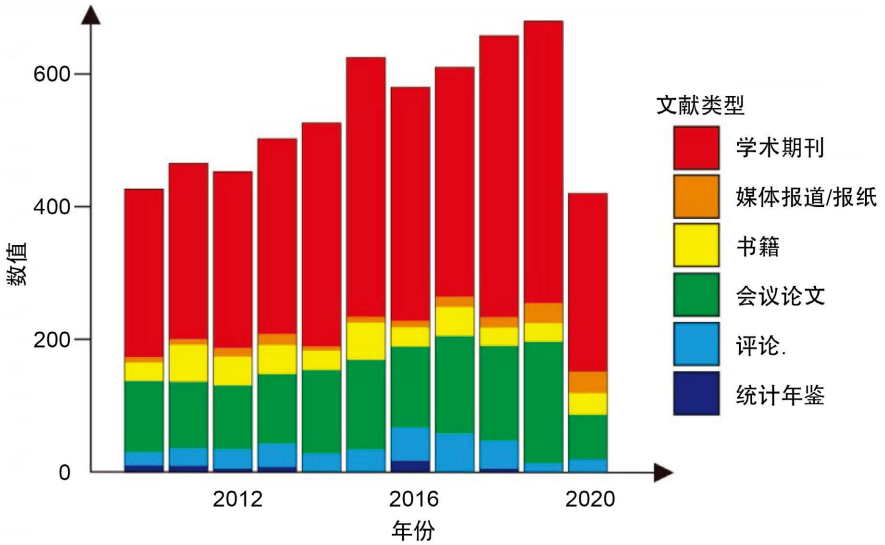

Figure 3. Number of articles and distribution of literature types related to land management by year
图 3. 按年份分类的与土地管理相关的文章数及其文献类型分布

4. 结果

文本挖掘练习的结果。首先，本研究使用术语和文献频率标准以及手工选择的组合创建了一个包含 641 个关键词的列表。在此之后，使用 DoV 和 DoD 得分和平均频率，包含的术语被划分为四个关键词涌

现和问题组：潜在的、弱的、较强的和强的(如图 1 所示)。最后，利用主题建模技术对主题下的关键词进行分组。

4.1. 关键词选择

首先，文本词语库中每个词的总体术语频率和文献频率被测量。非常一般的术语，如权力、农业、制度改革、可持续发展等，在本数据集中出现得最频繁(表 2)。为了引导对信号的关注，需要提取一个更加具体和土地管理相关关键词的列表。以前的研究要么选择使用预先确定的关键词列表，要么使用他们自己对关键词选择的主要标准的判断手动构建列表。本文使用了自动选择和手动选择的结合。首先使用以下筛选标准：1) 该术语必须在文本语料库中至少出现十次，2) 该术语必须在至少五篇文献摘要中出现。使用这些标准，本研究可以将词的数量从 38,435 个减少到 7708 个。为了完成关键词的选择，使用了手动选择。最后，选取了 635 个关键词进行进一步分析。

Table 2. Visibility and diffusion of the 10 most common keywords
表 2. 最常见的 10 关键词的可见度和扩散度

关键词	总学期频率	总文件频率	可见度	扩散都
权利	1535	1135	1.65	1.275
农业	1247	545	1.373	0.577
制度改革	650	396	0.657	0.406
规划	663	409	0.647	0.405
气候变化	693	377	0.722	0.394
土地整治	492	363	0.533	0.388
可持续发展	1557	333	1.724	0.363
三维建模	742	354	0.767	0.36
自然资源保护	477	303	0.555	0.35
战略	525	330	0.545	0.35

4.2. 未来信号

接下来，根据以前的研究[22] [23] [24]提出的 TF_x 和 DF_x 值以及 DoV 和 DoD 分数，将关键词分类在四个象限中。表 3 总结了四类关键词的划分。第一个观察结果是，当用平均值作为分类的阈值时，关键词不会非常均匀地分为四类。高达 81%的关键词被分类为强关键词或潜在关键词，而且在问题维度中的份额更高达到 84%。为了更好地了解不同类别下的关键词类型，基于结果绘制了关键词涌现图(图 4)。关键词问题图与其相似。

Table 3. Keyword counts and percentages for each category
表 3. 每个类别的关键词计数和百分比

类别	信号规模		问题维度	
	数量	百分比(%)	数量	百分比(%)
潜伏信号	236	37	273	43
弱信号	94	15	54	9
较强信号	27	4	45	7
强信号	278	44	263	41

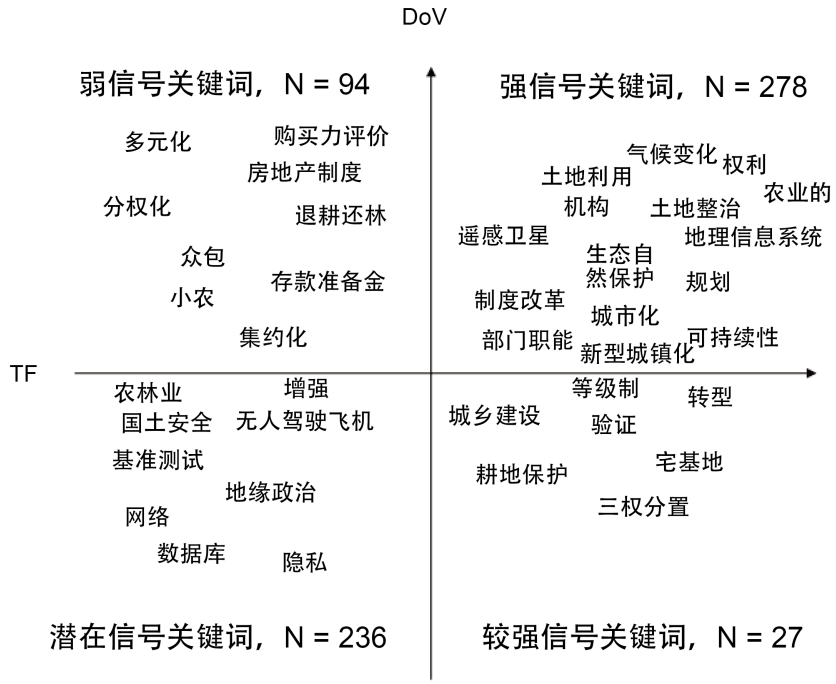

Figure 4. Keyword occurrence diagram (random sort in quadrant)
图 4. 关键词出现图(象限中随机排序)

然后，使用 LDA 对 25 个主题下的关键词进行分组。选择这个数字是一个迭代过程的结果，其中特别有两个目标需要加权：主题的一致性和新兴主题的出现。本文将主题的数量固定到一个结果或连贯和易于解释的水平，同时，将数量设置得足够高，以产生一些潜在的弱信号主题。因此，这里使用的主题数量并不适用于所有上下文和所有词语库。由于假设关键词在不同类别中同时出现，并且强制基于关键词出现或问题状态的分组会产生人为的结果，因此对整个关键词集进行了 LDA 分析。对话题分配结果进行如下处理：回顾了每个话题中最常见的 15 个术语，并检查了它们的关键词出现和问题状态。在此之后，主题被给予一个名称和基于一组较小的关键词的解释。以达到所包含的关键词将形成一个连贯的解释。在这个阶段，由于关键词的重叠，一些主题被合并了。最后，保留了十七个主题(表 4 和表 5)。最后，在四个信号类别下分配已确定的主题。遵循一个简单的规则：类别是根据该主题最重要的关键词所在的象限来确定的。三个潜在的信号主题和一个弱信号主题是这样识别的(表 3)。除了一个例外，其余的主题被归类为强信号(表 4)。第一个潜在信号的主题是所谓的“非普遍增强的定位措施”，它围绕差分全球定位系统，一个系统，提供定位修正的 GPS 信号。第二个潜在信号“社会经济变量的新数据来源”也是基于一个技术缩写，因为气象卫星 NPP (国家极地轨道伙伴关系)是这个话题的核心关键词。对这一问题的解释是，天气和气候监测数据来源越来越多地被用于其核心功能中，作为经济活动代理。第三个潜在信号主题叫做“国土安全”，由武器、条约和通量等关键词组成。从国家安全的角度来看，土地是一种重要的资源，本研究从这个角度来解读这个话题。唯一的弱信号专题称为“数字形式的计划”，它涉及到 GIS 软件的改良和从纸质文件向数字化转变的能力。

较强信号的主题是：参与式土地整合。土地整合是解决土地零碎化问题的常用土地调查程序，反映了相关关键词出现的频率较高。这个主题的一些关键词的变化率是中等的，这使它区别于强信号。公认的强信号清单要更加的广泛，主题范围从以技术为导向的主题，如“三维城市模型可视化”、“摄影测量和激光扫描的进展”、通过“减缓气候变化”、“生物多样性”和“与环境有关的威胁”等生态和数

据收集卫星地图，到与土地管理的核心职能密切相关的主题，如“土地使用协调”、“土地权”和“土地冲突”。每个主题的解释见表 5。

Table 4. Potential and weak signal themes identified in land management
表 4. 土地管理中确定的潜在信号和弱信号主题

主题	关键词	解释	类别
非普遍加强定位措施	DGPS；船舶；北极；性能；拓扑；制图；协同合作	地面基准站发送(非通用)修正	潜在信号
社会经济变量的新数据来源	NPP；城市化；旅游；气象；海域；沿海城市	天气和气候监测数据源(例如 NPP)越来越多地被用作经济发展和实时测量社会经济变量的代理	潜在信号
国家安全	武器；法规；法律；财产；土地	从国家安全的角度来看，土地是至关重要的战略资源	潜在信号
数字形式的计划	地理数据库；CAD；地理信息系统；精确性；软件；遥感；卫星	GIS 的改良和优化使数字化成为可能	弱

Table 5. Identified mainstream and priority themes for land management
表 5. 已确定的土地管理的主流主题和重点主题

主题	关键词	解释	类别
参与式土地治理	流程化；农村；协同合作；体制；碎片化；参与性	参与式土地整理作为解决土地碎片化问题的工具	较强信号
三维城市模型可视化	建筑物；可视化；城市规划；交互性；可行性	三维城市模型利用了土地和建筑信息的可视化	强信号
运输与安全	交通运输；安全；海事机构；公路铁路；应急	土地管理在综合安全管理体系中的作用日益增强	强信号
土地冲突	土地使用权；冲突；重建；城镇化建设；“三农问题”	农村土地制度三项改革、宅基地“三权分置”造成土地冲突	强信号
摄影测量和激光扫描的进展	精度；自动化；无人机；地形；扫描；摄影	摄影测量学和激光扫描技术的进步产生了更高精度的数据	强信号
图像传感器的进展	传感器；成像；雷达；卫星；坐标；光谱	图像传感器的质量迅速提高贡献在新的应用领域	强信号
土地使用协调	城市；治理；权力；灌溉；培育；竞争；农民；开发	权力下放、分散土地管理作为支持当地土地使用的工具	强信号
促进土地可持续利用	植被；降雨；退化；土壤；可再生；常态化差值植被指数；障碍	在植被指数(NDVI)等的支持下，促进土地可持续利用的努力	强信号
减缓气候变化	气候变化；气体排放；污染；不确定性因素；减缓；植被破坏	提高对气候变化影响和土地利用在减缓气候变化方面的作用的认识	强信号
生物多样性	物种；组织；复杂性；多样性；珍稀动物；有毒性；许可证	生物多样性的减少成为一个日益关注的问题	强信号
清洁能源	容量；太阳能；风能；管理；郊区；升级优化；资金	太阳能电池板可作为在郊区的一种清洁能源	强信号
卫星图像数据收集	住宅区；卫星；安全；人口普查；档案馆；资源	卫星图像为城市扩张提供了数据	强信号
与环境有关威胁	洪水；网络；可用性；脆弱性；水文地质；灾害	洪水和水文地质灾害影响了土地收益	强信号
土地产权	保护；可持续性；法律；法规；政策；制度；整合；土地所有者	为原住民和农民土地拥有者提供土地权益保障	强信号

5. 讨论

本研究证实了以前的研究已经注意到的未来信号的解释，这将对未来信号检测提供实证资料。在本研究中，个别关键词被分组在主题之下，以方便解释任务。除了难以对解释部分进行量化和自动化之外，未来的信号通常是高度主观的；对一个人来说可能是新兴信号，对另一个人来说可能主流信号。因此，即使用最谨慎的信号检测过程，本研究也不能得到所有人都支持的结果。因此，本文与 Yoon 研究结果一致，他期望自动化方法和基于专家的方法将可能在信号检测中继续相辅相成[22]。尽管进行了一些主题的手动合并，但本研究不希望过多地干预 LDA 的结果。因此，一些重叠的信号主题可以观察到，例如，在使用卫星地图和 NPP 收集有关社会经济变量的信息。本研究结果最突出的特点是大量的强信号主题。其原因可能是研究方法和使用的文本挖掘工具，也可能仅仅是土地管理领域的话语倾向于围绕某些主题。无论如何，本研究发现已确定的信号主题非常可信并且多元化，因为它们从气候、生物多样性、运输到传感器技术、土地冲突和标准化都有所不同。这恰恰强调了土地是如何连接本研究今天所面临的“三农”问题、新型城镇化建设问题和蓝天白云金山银山政策等。同时，土地管理职能(土地保有权、土地价值、土地开发、土地使用)为实现“十四五”规划中可持续发展目标起到关键作用。

本文研究中挖掘国外文献时在一组潜在的信号和弱信号的主题中发现(表 4)。土地在国家安全中的战略作用，在土地管理文献中并不经常被提及。例如，在芬兰为保护国土安全，新的法案从 2020 年 1 月开始执行，外国居民购买房产需要获得许可。此外，唯一被归类为微弱信号的主题是平面图数字化，其应是一个热门主题，在澳大利亚、新西兰和新加坡，地籍系统正在改革以支持数字地籍数据[29]，在北欧五国，公共机构努力实现数字土地利用规划。

最后，我们注意到本研究中识别的未来信号主要反映了学术论文，因为我们使用中国知网、万方、维普、Elservier、Springer、Scopus 等数据库作为文本数据的来源。尽管科学成果的数量在不断增长，我们建议这种方法应该通过更大、更全面的文本词语库进一步测试。例如，利用社交网络信息推特、微博、知乎进行拓展分析，可能是这项研究的一个有趣的后续行动。尽管如此，我们认为，在土地管理领域，学术话语与非学术话语紧密交织，使得利用科学资源成为本研究的合理选择。Hiltunen 也认为学术和科学期刊是探测弱信号的最佳来源[30]。该框架也可以在未来扩展，例如，涵盖关键词和未来信号主题的网络方面，因为可视化哪些术语和主题相互连接可以大大加强解释。最后，我们注意到本研究使用了一种主题建模技术 LDA。除了 LDA 之外，还应该探索其他技术，以增加对未来信号解释自动化的理解。

6. 结论

在研究中，使用文本挖掘工具来展示如何在土地管理的背景下探索、组织和分析未来信号。未来的信号主题分为四类：潜在的、弱的、较强的和强的信号。研究结果表明，未来土地管理领域的信号从自然灾害和可持续土地利用到测绘技术的灵活标准化和优化等诸多主题方面。本研究的结果强调了从整体角度探讨未来土地管理的重要性。研究结果也支持地籍系统在支持可持续发展目标实现中的核心作用。除了经验证明，研究通过提出一种半自动的方式来解释未来信号，为未来信号检测文献做出了贡献。也就是说，本文展示了如何将主题建模作为最初由 Yoon 开发的未来信号检测框架的一部分，进而挖掘出提高未来信号解释质量的潜力。

总的来说，使用文本挖掘进行预见练习显示了巨大的潜力，随着可用数据量的增加和语言处理工具的发展，我们期望在未来增加这种潜力。与此同时，我们注意到，特别是弱信号检测需要结合自动化的定量观点(减少人们在接收信号和主题方面的偏见)和参与式的定性观点(确保更高程度地展示新兴的因素)的方法。

基金项目

2022 年集团公司内部科研项目－基于 InVEST 和 CA-Markov 模型的西北农牧交错带碳储量时空变化研究-DJNY2022-21。

参考文献

[1] 刘俊, 孟鹏, 龚暄杰. 应加快推进《土地管理法》新一轮修改完善工作——推进依宪修改完善《土地管理法》专题研究座谈会综述[J]. 中国土地科学, 2015, 29(5): 16-21.

[2] 王静怡, 李晓明. 近 20 年中国耕地数量变化趋势及其驱动因子分析[J]. 中国农业资源与区划, 2019, 40(8): 171-176.

[3] 商冉, 曲衍波, 姜怀龙. 人地关系视角下农村居民点转型的时空特征与形成机理[J]. 资源科学, 2020, 42(4): 672-684.

[4] 徐忠国, 李冠, 郑红玉, 吴次芳, 卓跃飞. 法律地籍领域模型研究动态及对中国地籍概念模型建构的理论启示[J]. 中国土地科学, 2019, 33(9): 19-27.

[5] Intelligence Advanced Research Projects Activity (2018) Foresight and Understanding from Scientific Exposition (FUSE). http://iarpa.gov/index.php/research-programs/fuse

[6] 陈美华, 王延飞. 科技管理决策中的地平线扫描方法应用评析[J]. 情报理论与实践, 2017, 40(12): 63-68.

[7] Cuhls, K.E. (2020) Horizon Scanning in Foresight—Why Horizon Scanning Is Only a Part of the Game. *Futures and Foresight Science*, **2**, e23. https://doi.org/10.1002/ffo2.23

[8] Amanatidou, E., Butter, M., Carabias-Hütter, V., Könnölä, T., Leis, M., Saritas, O., et al. (2012) On Concepts and Methods in Horizon Scanning: Lessons from Initiating Policy Dialogues on Emerging Issues. *Science and Public Policy*, **39**, 208-221. https://doi.org/10.1093/scipol/scs017

[9] 张志娟, 刘萍萍, 王开阳, 郄海拓. 国外科技创新治理的典型政策工具运用实践及启示[J]. 科技导报, 2020, 38(5): 26-35.

[10] 董尹, 刘千里, 宋继伟, 赵小康. 弱信号研究综述: 概念、方法和工具[J]. 情报理论与实践, 2018, 41(10): 147-154.

[11] 董尹, 刘千里, 章蕾. 弱信号传递博弈研究: 基于传信博弈视角[J]. 情报理论与实践, 2019, 42(10): 21-28.

[12] 苗红, 郭鑫, 吴菲菲. 产业融合弱信号识别研究——以老年智能家居领域为例[J/OL]. 情报杂志: 1-8. http://kns.cnki.net/kcms/detail/61.1167.G3.20200731.1031.006.html, 2020-08-15.

[13] Sutherland, W.J., Broad, S., Butchart, S.H., Clarke, S.J., Collins, A.M., Dicks, L.V., *et al*. (2019) A Horizon Scan of Emerging Issues for Global Conservation in 2019. *Trends in Ecology and Evolution*, **34**, 83-94. https://doi.org/10.1016/j.tree.2018.11.001

[14] 王乐, 王水, 刘胜蓝, 王辉兵. 基于索引树的带通配符序列模式挖掘算法[J]. 计算机学报, 2019, 42(3): 554-565.

[15] 司谨源. 基于地平线扫描的公安情报预警模式构建[J]. 情报杂志, 2020, 39(1): 56-62.

[16] 罗威, 武帅, 田昌海. 数据驱动的技术预测之研究评析——以 FUSE 项目为例[J]. 情报理论与实践, 2019, 42(7): 15-19+34.

[17] 陈世伟. 犯罪学的绿色视角:西方绿色犯罪学的发生、发展及借鉴[J]. 国外社会科学, 2016(3): 97-109.

[18] Ansoff, H.I. (1975) Managing Strategic Surprise by Response to Weak Signals. *California Management Review*, **18**, 21-33. https://doi.org/10.2307/41164635

[19] Hiltunen, E. (2006) Was It a Wild Card or Just Our Blindness to Gradual Change. *Journal of Futures Studies*, **11**, 61-74.

[20] Kuosa, T. (2011) Different Approaches of Pattern Management and Strategic Intelligence. *Technological Forecasting and Social Change*, **78**, 458-467. https://doi.org/10.1016/j.techfore.2010.06.004

[21] 邓胜利, 林艳青, 王野. 企业竞争弱信号的特征提取与定量识别研究[J]. 图书情报工作, 2016, 60(10): 67-75.

[22] Yoon, J. (2012) Detecting Weak Signals for Long-Term Business Opportunities Using Text Mining of Web News. *Expert Systems with Applications*, **39**, 12543-12550. https://doi.org/10.1016/j.eswa.2012.04.059

[23] Kim, H., Han, Y., Song, J. and Song, T.M. (2019) Application of Social Big Data to Identify Trends of School Bullying Forms in South Korea. *International Journal of Environmental Research and Public Health*, **16**, Article No. 2596. https://doi.org/10.3390/ijerph16142596

[24] Lee, Y. and Park, J. (2018) Identification of Future Signal Based on the Quantitative and Qualitative Text Mining: A

Case Study on Ethical Issues in Artificial Intelligence. *Quality & Quantity*, **52**, 653-667. https://doi.org/10.1007/s11135-017-0582-8

[25] Kim, H., Ahn, S. and Jung, W. (2018) Horizon Scanning in Policy Research Database with a Probabilistic Topic Model. *Technological Forecasting and Social Change*, **146**, 588-594. https://doi.org/10.1016/j.techfore.2018.02.007

[26] 何伟林, 谢红玲, 奉国和. 潜在狄利克雷分布模型研究综述[J]. 信息资源管理学报, 2018, 8(1): 55-64.

[27] 张亮. 基于 LDA 主题模型的标签推荐方法研究[J]. 现代情报, 2016, 36(2): 53-56.

[28] 朱道林, 程建, 张晖, 戚渊. 2019 年土地科学研究重点进展评述及 2020 年展望——土地管理分报告[J]. 中国土地科学, 2020, 34(1): 92-101.

[29] Olfat, H., Jani, A., Shojaei, D., Darvill, A., Briffa, M., Rajabifard, A., *et al.* (2019) Tackling the Challenges of Visualising Digital Cadastral Plans: The Victorian Cadastre Experience. *Land Use Policy*, **83**, 84-94. https://doi.org/10.1016/j.landusepol.2019.01.037

[30] Hiltunen, E. (2008) The Future Sign and Its Three Dimensions. *Futures*, **40**, 247-260. https://doi.org/10.1016/j.futures.2007.08.021