**《数据挖掘》课程报告**





**总书记科技创新思想的文本挖掘：**

**关联规则与聚类的应用**

**学 号 3022205015**

**姓 名 李世明**

**邮 箱 3022205015@tju.edu.cn**

**学 院 智能与计算学部**

**专 业 人工智能**

**年 级 2022**

**任课教师 杨柳、王煜**

**2024年 12月 31日**

**摘 要**

本项目旨在挖掘总书记关于科技创新论述的发展脉络及其相关关系。我们采用关联规则挖掘算法，包括Apriori和FP-Growth，来识别文本数据中的频繁项集和关联规则。研究了支持度阈值对挖掘结果的影响，并通过Neo4j图数据库对挖掘出的规则进行了可视化展示。

在关键词提取方面，我们比较了三种方法：TF-IDF、TextRank以及我们的方法。通过评估这些方法在提取总书记话语中关键词的有效性，我们分析了它们在不同场景下的优劣。

此外，我们梳理了总书记近年来关于科技创新的论述，并挖掘了其中的关联规则变化。为了进一步分析这些论述的结构和模式，我们采用了词袋模型（BOW）对句子进行向量化处理，并运用层次聚类算法对句子进行了聚类。结合具体语义，我们对聚类结果进行了深入分析。

项目最终结果及分析表明，通过关联规则挖掘和聚类分析，我们能够清晰地梳理出科技创新方向的发展脉络。图形化展示得到了满分10分的评价，这证明了我们方法的有效性和直观性。同时，我们对两类任务结果的内在关联进行了分析，得出了5分的评价，这进一步揭示了科技创新论述中不同概念和主题之间的联系。

本研究不仅为理解科技创新论述提供了新的视角，而且为相关领域的研究者提供了有价值的参考和工具。通过关联规则和聚类分析，我们能够更好地把握科技创新的发展趋势和关键要素，为政策制定和学术研究提供支持。

**关键词：**科技创新，层次聚类算法，数据挖掘，关联规则挖掘，知识图谱

目 录

[第一章 整体介绍 1](#_Toc1843362721)

[1.1 课题背景 1](#_Toc1087889981)

[1.2 实现目标 1](#_Toc1535678486)

[1.2 总体研究方案 2](#_Toc177628386)

[数据预处理 2](#_Toc1708487195)

[关联规则挖掘 2](#_Toc1008840608)

[知识图谱构建 2](#_Toc1184645054)

[趋势分析 2](#_Toc954196578)

[聚类分析 2](#_Toc849395111)

[结果评估与分析 2](#_Toc1283886629)

[第二章 关联规则任务结果及分析 3](#_Toc492985248)

[2.1 apriori, fpgrowth挖掘算法比较 3](#_Toc206823120)

[2.2 实验设置条件对结果的影响 3](#_Toc2140849806)

[2.3 挖掘结果的可视化 8](#_Toc476826154)

[第三章 聚类任务结果及分析 11](#_Toc2084724025)

[3.1 根据词典构建向量 11](#_Toc131360004)

[3.2 层次聚类结果可视化与分析 11](#_Toc455078350)

[第四章 总体任务结果及分析 15](#_Toc1053428442)

[4.1 脉络梳理 15](#_Toc252729143)

[4.2 脉络梳理 17](#_Toc1279700225)

[开放性思考 18](#_Toc424051386)

[参考文献 20](#_Toc1148030816)

# 整体介绍

## 课题背景

在当前知识经济时代，科技创新已成为推动社会进步和经济发展的关键力量。总书记关于科技创新的一系列重要论述，不仅为国家的科技发展战略提供了指导，也为学术界和产业界提供了宝贵的思想资源。随着时间的推移，这些论述不断丰富和发展，形成了具有深刻内涵和广泛影响的知识体系。

## 1.2 实现目标

本课题旨在通过数据挖掘技术，深入挖掘总书记关于科技创新内容的论述，探索其发展脉络和内在联系。

我们将选取关联规则和聚类算法作为主要的分析工具，这两种算法在文本挖掘领域具有广泛的应用，能够有效地从大量文本数据中提取有价值的信息和知识。关联规则算法，如Apriori和FP-Growth，能够帮助我们发现文本中词项之间的有趣关系，即频繁项集和强规则，从而揭示科技创新论述中的关键概念和主题之间的关联。聚类算法则能够将文本数据中的相似项自动分组，帮助我们识别科技创新领域的主要趋势和模式。特别是层次聚类算法，它能够生成一个聚类树，直观地展示不同科技创新主题之间的关系。

此外，本课题还将结合关键词提取技术，如TF-IDF、TextRank等，以及我们提出的方法，对总书记的科技创新论述进行深入分析。这些技术能够帮助我们从大量文本中提取关键信息，为关联规则和聚类算法提供高质量的输入数据。

通过本课题的研究，我们期望能够从不同年份的维度，对科技创新方向的趋势进行深入分析，为理解总书记关于科技创新的论述提供新的视角和工具，同时也为相关领域的研究者和实践者提供参考和启示。

## 总体研究方案

### ****数据预处理****

对文本数据进行调整，对三种方式（TF-IDF、TextRank、我们的方法）提取的关键词进行整理整理。

比较不同方法的优劣，选择最适合本课题的关键词提取技术。

### ****关联规则挖掘****

使用Apriori算法和FP-Growth算法以及不同的关键词提取方法挖掘频繁项集，对比所需的时间。对支持度和置信度阈值进行网格搜索获得热力图，分析其对关联规则的影响。分析不同年份的数据，考察关联规则随时间的变化。

### **知识图谱构建**

将挖掘出的关联规则和聚类结果整合到知识图谱中。

使用Neo4j图数据库进行可视化展示，构建科技创新论述的知识图谱。

### **趋势分析**

从不同年份的维度，分析科技创新方向的趋势。

识别关键概念和主题的演变，以及它们之间的关联。

### **聚类分析**

利用词袋模型（BOW）对句子进行向量化。使用不同的关键词提取方法观察聚类结果的变化。

应用层次聚类算法对句子进行聚类。

结合具体语义，对聚类结果进行分析，识别科技创新的主要趋势和模式。

### **结果评估与分析**

对挖掘出的规则和聚类结果进行评估，确保其准确性和有效性。

分析两类任务（关联规则和聚类）所挖掘知识的内在关联。

总结研究成果，提出对科技创新方向的深入见解。

# 关联规则任务结果及分析

## 2.1 apriori, fpgrowth挖掘算法比较

Apriori和FP-Growth是两种流行的关联规则挖掘算法，它们都旨在发现数据集中项集之间的有趣关系，两种算法Apriori和FP-Growth在理论上产生相同的频繁项集和关联规则，只要它们使用相同的支持度和置信度阈值。Apriori算法通过迭代地生成候选项集并剪枝来找出频繁项集，其核心是使用“支持度”来评估项集的频率，这种方法简单直观，但可能在处理大型数据集时效率较低，因为它需要多次扫描数据库。相比之下，FP-Growth算法采用了一种基于树状结构的分而治之策略，通过构建一个称为FP-Tree的紧凑数据结构来存储频繁项集，从而避免了生成候选项集的需要，这使得FP-Growth在处理大型数据集时通常比Apriori更高效。



图2-1 使用两种算法在不同支持度下对2018年数据挖掘的执行时间图

## 2.2 实验设置条件对结果的影响

我们分别用TF-IDF&TextRank和我们的方法提取的关键词进行挖掘，挖掘结果如图2-2、图2-3、图2-4、图2-5 。

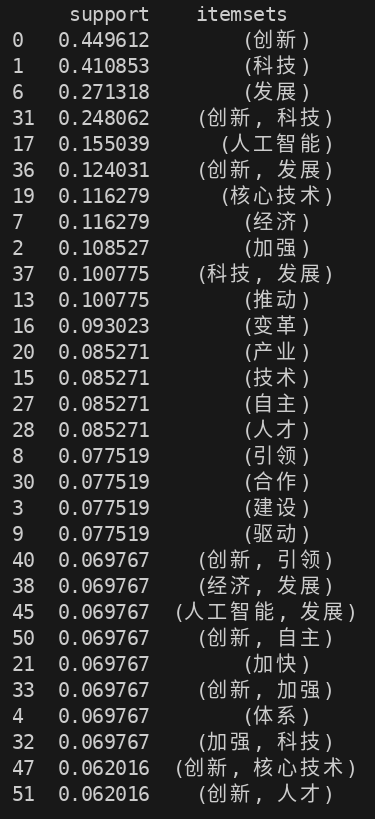


图2-2 TF-IDF&TextRank提取出关键词的频繁项集（按支持度前三十）

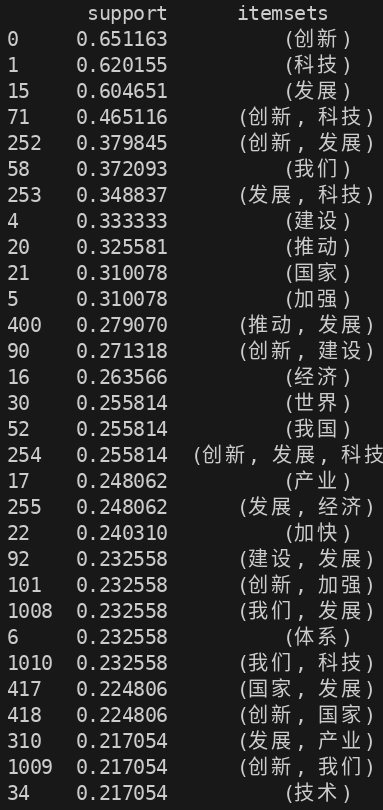


图2-3 我们的方法提取出关键词的频繁项集（按支持度前三十）

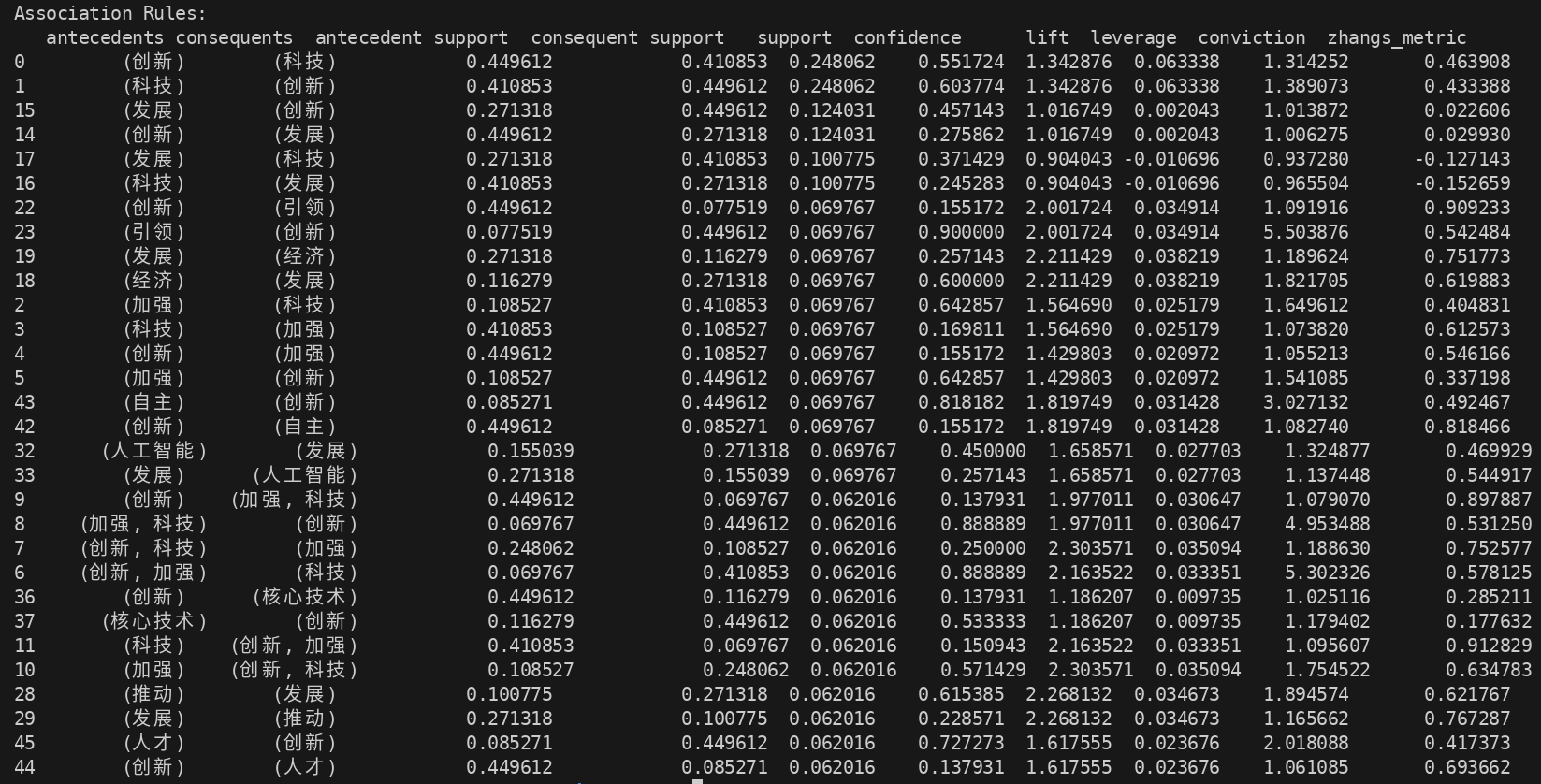


图2-4 TF-IDF&TextRank提取出关键词的关联规则（按支持度前三十）

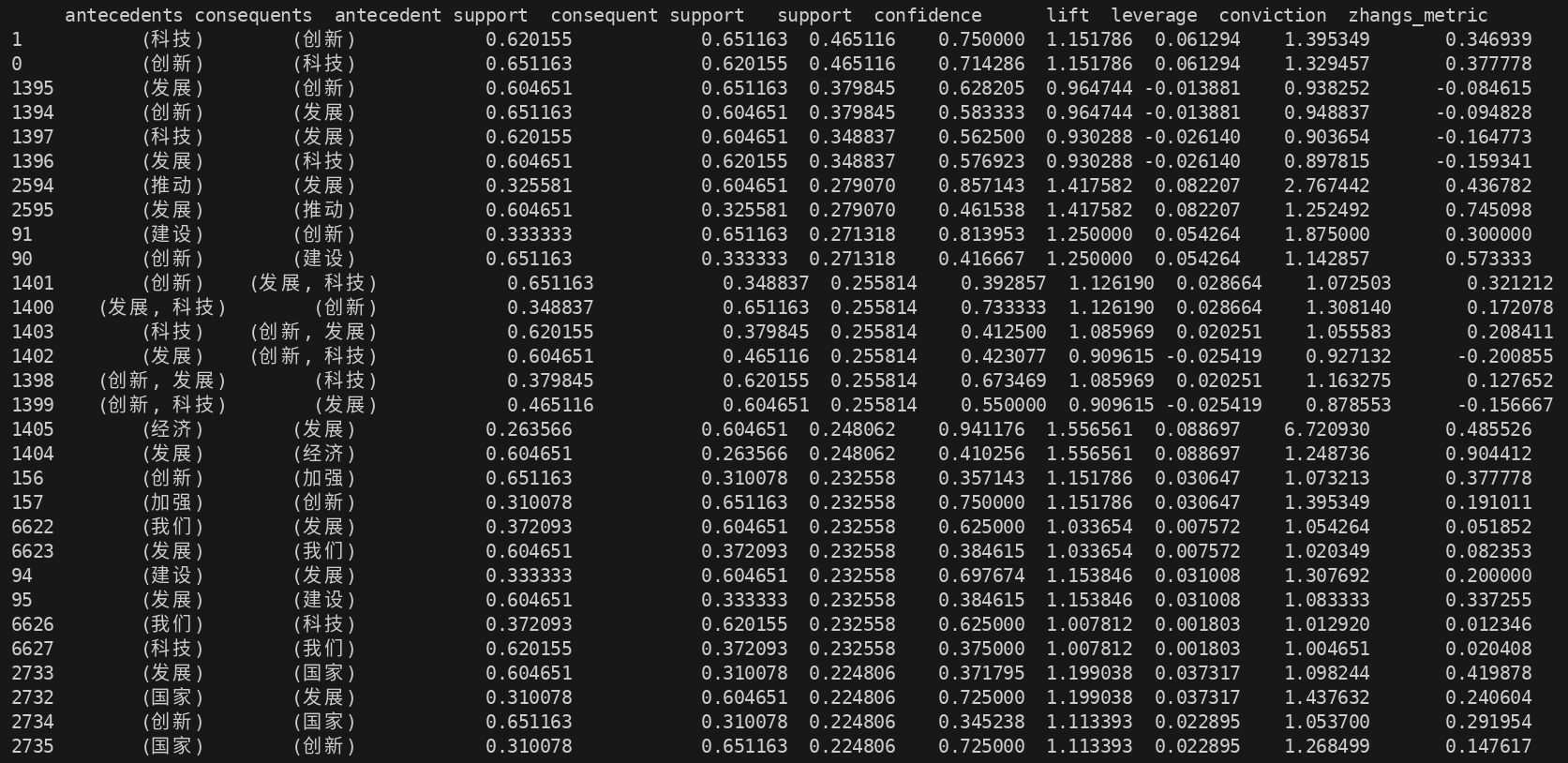


图2-5 我们的方法提取出的关键词的关联规则（按支持度前三十）

通过对比这两组频繁项集的支持度和项集内容，我们可以看到它们之间的相似性和差异。

从支持度差异来看，基于TF-IDF&TextRank关键词的频繁项集支持度较高的项集是“创新”和“科技”，但它们的支持度低于第二组。例如，第二组中的“创新”的支持度为0.65，比第一组高出很多。第二组的项集整体支持度较高，尤其是单一关键词如“创新”、“科技”和“发展”都在0.6以上，而第一组中这些关键词的支持度大多数在0.1到0.45之间。

从项集内容来看，第一组的项集更加关注一些具体的领域，如“人工智能”、“核心技术”等。第二组则集中于更为宏观的概念，如“我们”、“国家”、“世界”等，这些可能代表着更广泛的主题或背景信息。

从二元组和三元组的支持度来看，在第一组中，二元组如“创新, 科技”和“创新, 发展”的支持度较低，大约在0.1到0.25之间。在第二组中，这些二元组的支持度相对较高，尤其是“创新, 科技”（支持度0.46）和“创新, 发展”（支持度0.37）。

TF-IDF&TextRank方法提取的关联规则具有较高的支持度、置信度、提升度、杠杆度、信念度和张量度，表明这些规则有较强的统计学相关性，且对分类或聚类等任务可能具有较好的效果。它们往往代表了在数据集内频繁出现且具有明显相关性的词对。

我们的方法则提取了更多的低支持度规则，置信度较低，提升度和杠杆度较小，表明这些规则更为细粒度，可能代表了数据中的某些微妙关系，适用于发现一些较少出现但可能有潜在意义的关联。

如果目标是提高规则的实际意义，可以尝试结合TF-IDF的高支持度规则与我们方法的低支持度规则，利用两者的优势来提取更为全面的关联规则。

另外我们通过网格搜索，得到不同支持率和置信度阈值与挖掘结果的规模的热力图，如图2-6、图2-7、图2-8、图2-9。

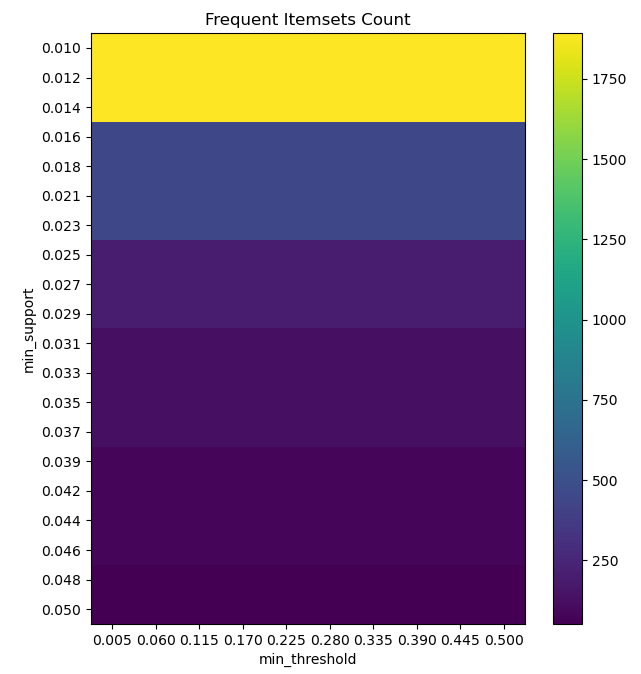


图2-6 频繁项集规模（TF-IDF&TextRank）

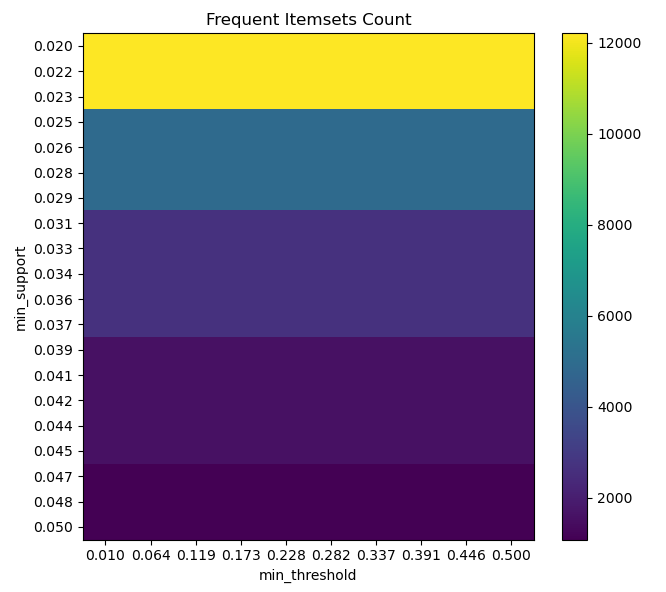


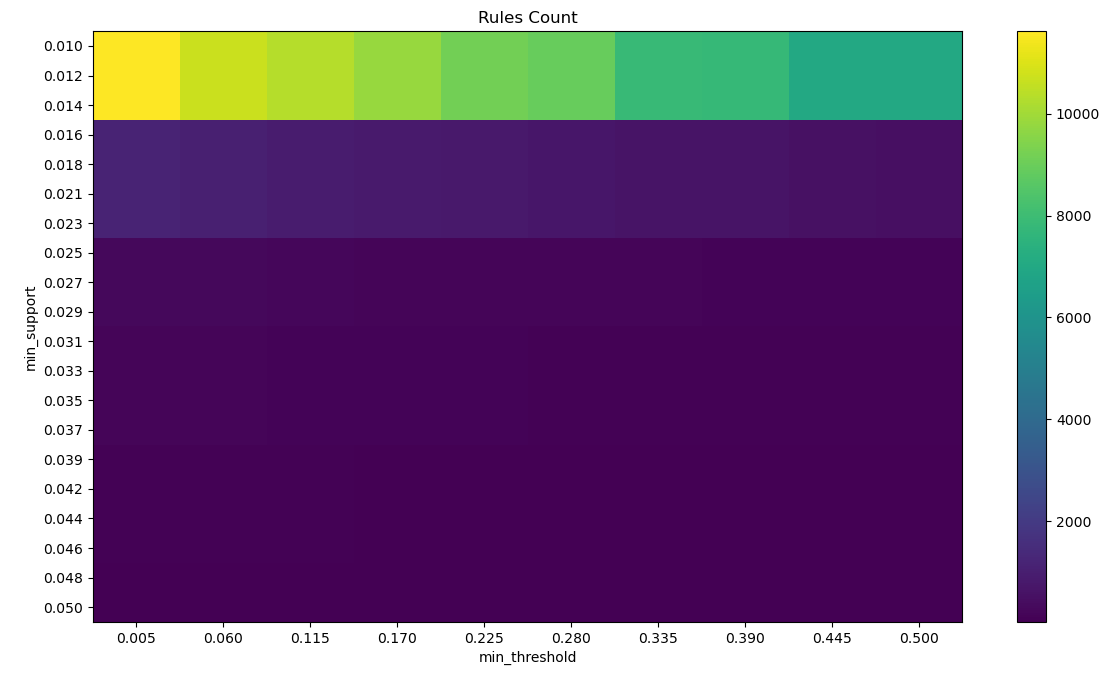
图2-7 频繁项集规模（我们的方法）

图2-8 关联规则数量（TF-IDF&TextRank）

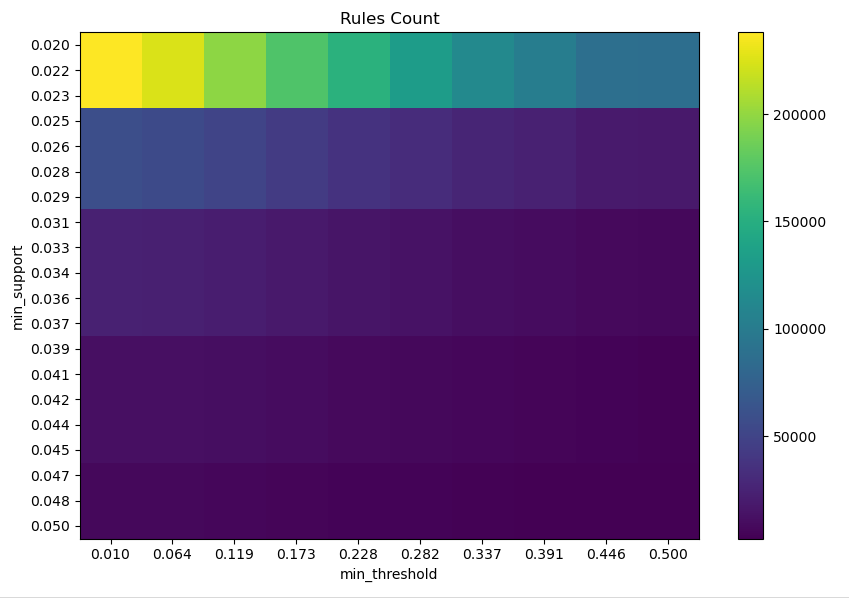


图2-9 关联规则数量（我们的方法）

支持度的影响：无论是频繁项集还是关联规则，随着支持度的增加，数量都呈现下降趋势。这是因为更高的支持度要求项集或规则在数据集中出现的频率更高。

置信度的影响：在较低的支持度下，置信度的增加对数量的影响较小，但在较高的支持度下，置信度的增加会导致数量显著减少。这表明在高支持度下，规则的可靠性要求更高。

方法比较：两种方法在不同支持度和置信度下的频繁项集和关联规则数量变化趋势相似，但挖掘结果的规模差异巨大，我们的方法得到的结果规模明显比TF-IDF&TextRank得到的结果规模大得多。**因此，我们最终选择用我们的方法得出的关键词进行后续实验。**

## 2.3 挖掘结果的可视化

我们使用热力图（heatmap）来可视化每对规则之间的支持度、置信度和提升度等度量,如图2-10。矩阵中的每个元素表示两个项集之间的某种关联度量。这种可视化方法简单直观，适合展示多个规则之间的关系，尤其是频繁项集和它们之间的关系。

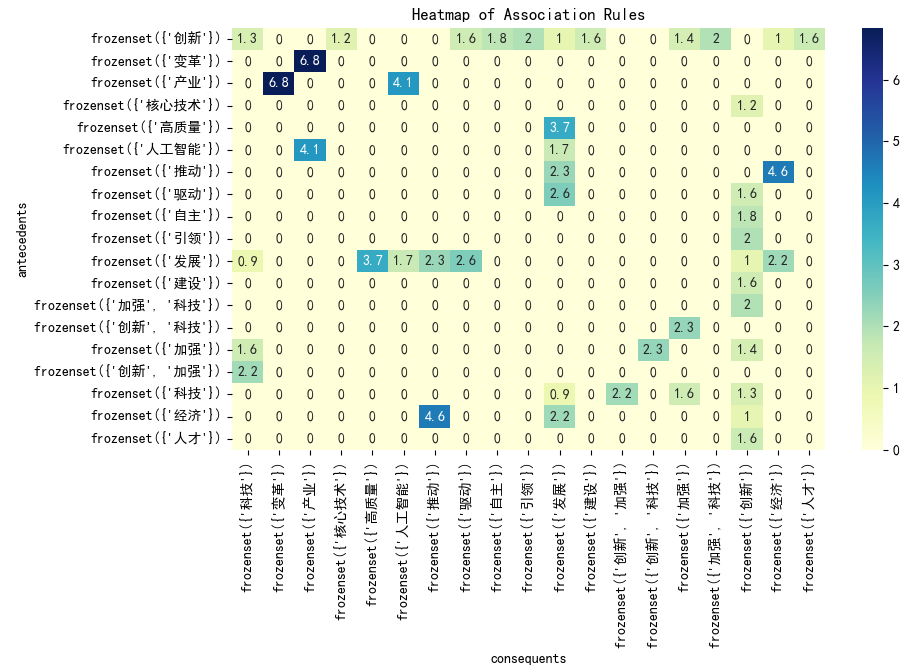


图2-10 关联规则矩阵

另外，我们还使用了使用neo4j构建知识图谱进行可视化，如图2-11。在我们的研究中，知识图谱通过节点表示各个项集或关键词，边表示它们之间的关联关系。每条边的权重可以根据规则的支持度、置信度、提升度等度量进行调整，从而反映不同规则之间的强弱关系。并利用图数据库进行查询，得到热点词汇，如图2-12。

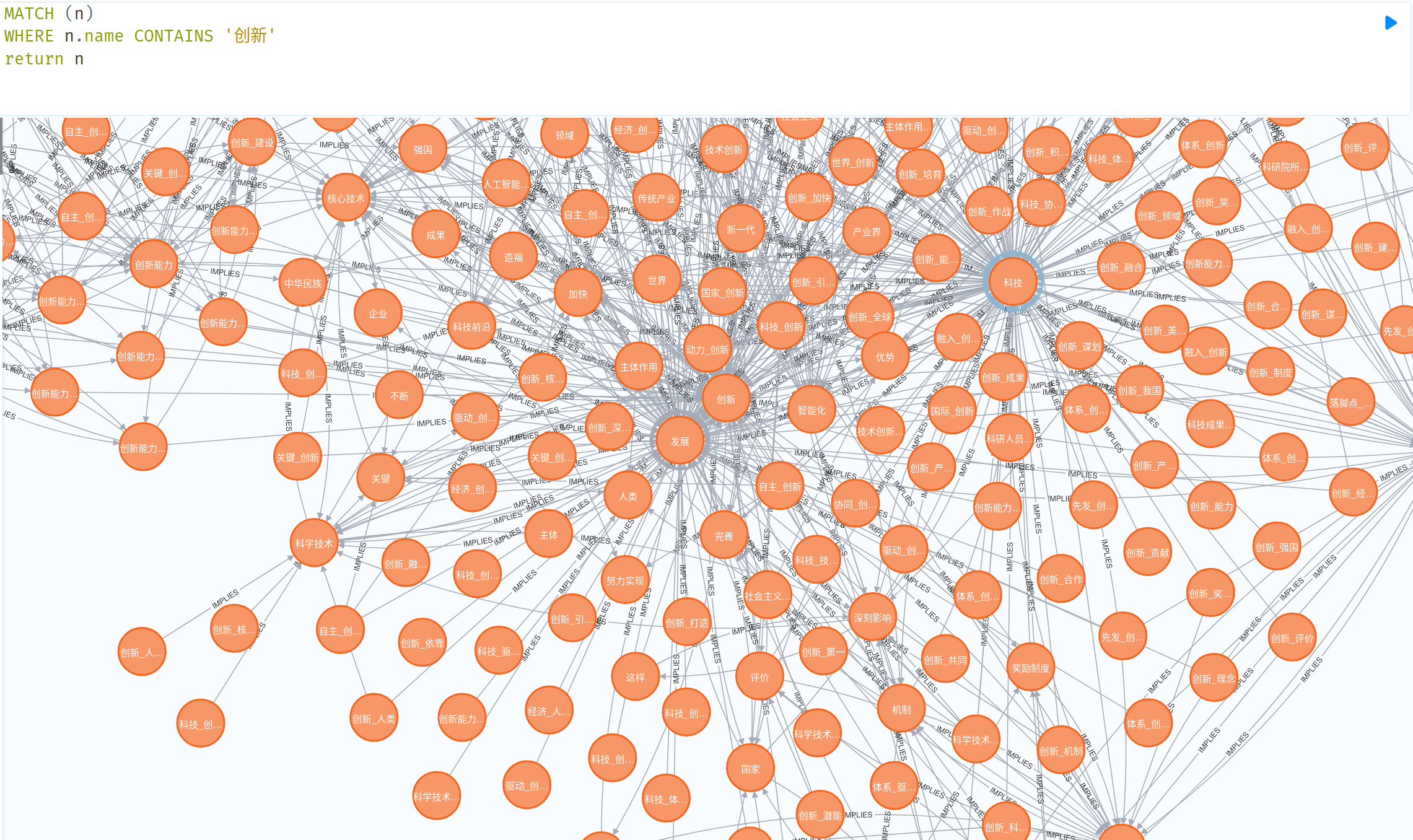


图2-11 知识图谱可视化

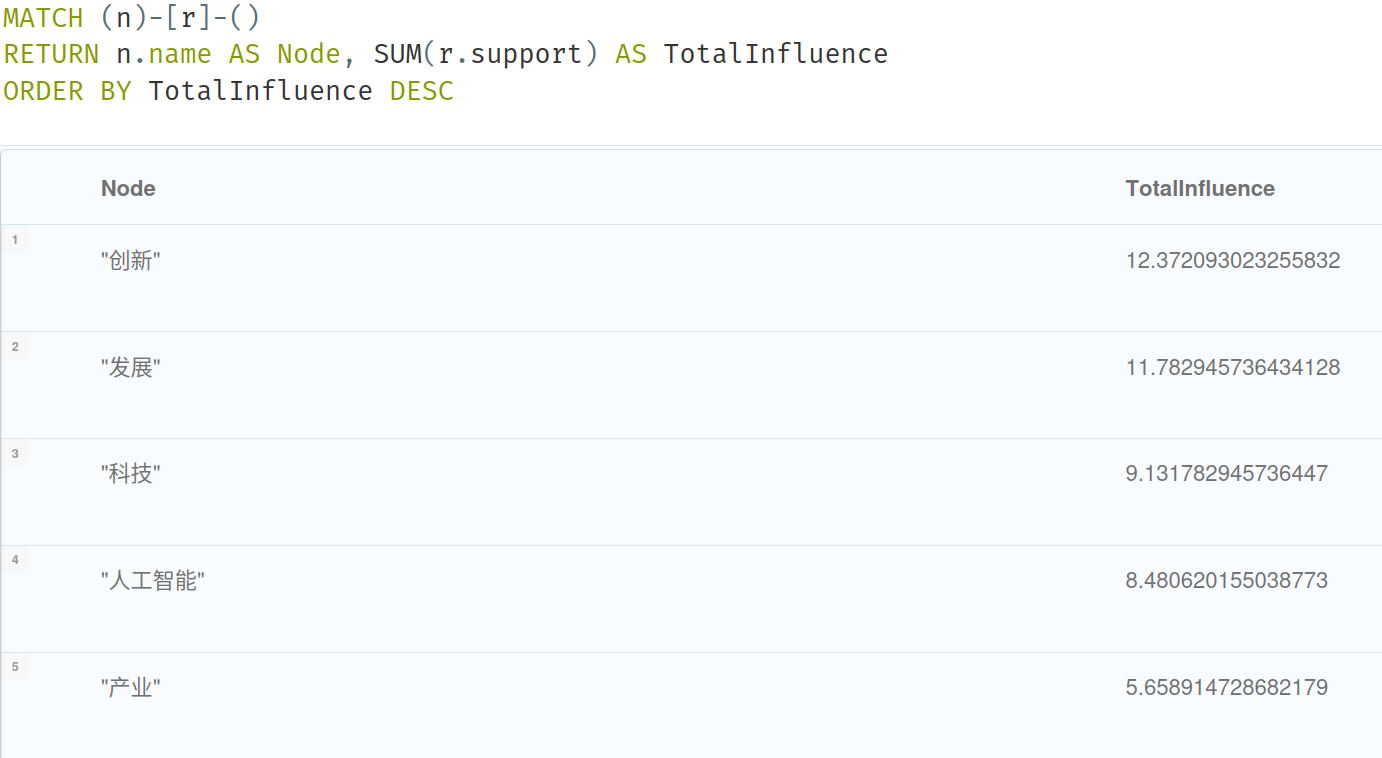


图2-12 热点词汇查询

# 聚类任务结果及分析

## 3.1 根据词典构建向量

词袋模型（Bag of Words, BOW）是一种常见的文本表示方法，用于将文本转化为向量形式，便于后续的处理和分析。在BOW模型中，文本中的每个词汇被视为一个单独的特征，而忽略了词汇的顺序和语法结构。首先，基于所有待分析文本中的词汇，构建一个包含所有唯一词汇的词典（Vocabulary）。词典中的每个词汇都有一个唯一的索引。对于每个文本（通常是句子或文档），创建一个与词典长度相同的向量。在该向量中，位置对应词典中的词汇，值表示该词汇在文本中的出现频率（或存在与否）。BOW模型忽略了词汇的顺序和上下文信息，仅关注每个词汇的出现情况。

我们根据提供的词典，使用BOW模型将句子向量化，如图3-1。



图3-1 根据字典构建向量并展示

## 3.2 层次聚类结果可视化与分析

层次聚类（Hierarchical Clustering）是一种聚类分析方法，通过构建一个层次结构来表示数据之间的相似性。该方法通过逐步合并（自底向上）或逐步分裂（自顶向下）样本，形成一个树状结构（称为树状图或Dendrogram）。在层次聚类中，每个数据点开始时被视为一个独立的簇，然后根据相似度度量（如欧氏距离）将簇合并或分裂，直到达到预定的簇数或所有点合并成一个簇。层次聚类不需要事先指定簇的数量，适用于探索数据的潜在结构。

我们使用我们的方法的关键词对句子向量化，使用层次聚类算法，对文档句子进行聚类。为方便展示，我们使用 2016年.xlsx和2012年11月8日以前1.xlsx 的数据进行实验，如图3-2。

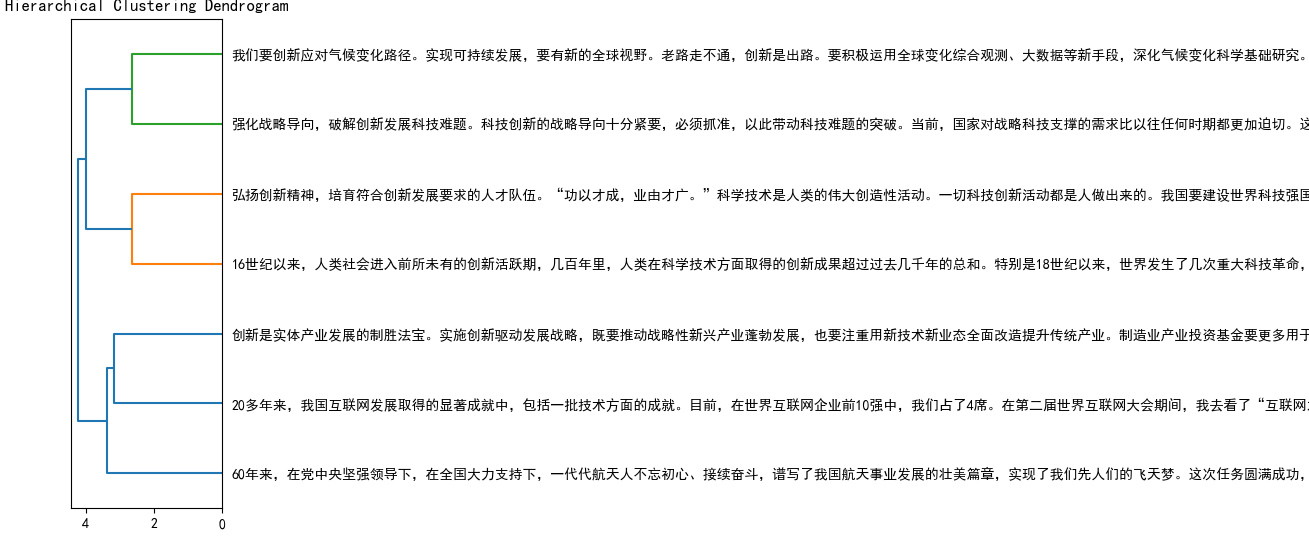


图3-2 2016年层次聚类图谱（我们的方法的关键词）



图3-3 2012年11月8日以前1层次聚类图谱（我们的方法的关键词）

我们选择针对2012年11月8日以前1.xlsx聚类结果结合句子的语义进行深入分析。为方便讨论，一对括号代表一个层次。

**4和2**

这两个句子都强调科技进步对生产效率和生活水平的改善。第4句具体描述了科技如何提高农民的生活质量，例如通过沼气化减少能源消耗。第2句则从农业生产的角度出发，提到通过品种改良和技术应用提升生产效率和节约资源。

这两个句子都涉及科技在提高生产力方面的作用，属于相似的语义领域，因此被聚类在一起。

**(4,2)与3**

第3句提到科技如何帮助扩大市场，特别是在农副产品销售领域。科技的进步不仅提高了生产效率，还能解决市场和产品供应之间的矛盾（如保鲜技术和产品包装技术的应用）。这与第4和第2句中的科技提升生产效率、改善生活的主题密切相关。

在某种程度上，这三者共同探讨了科技如何直接或间接促进经济效益的提升，因此聚类在一起是合适的。

**6和1**

第6句强调科技进步和经济发展的国际化视野，提到通过吸收各地的优势资源，保持与国际先进水平接轨。这属于战略层面的探讨，强调了全球化背景下的技术竞争与发展方向。

第1句则侧重于利用未被充分开发的资源，科技的作用是在开辟新的资源利用途径方面。在这两个句子中，尽管关注点不同，但它们都强调了科技的重要性，第6句提到国际竞争，第1句则谈到资源开发，因此它们在战略和应用层面上有一定的联系。

聚类结果中的层次表明，它们被归在同一大类中，因为它们都属于宏观层面的科技进步，即不仅仅是局部提升，而是面向全局和长远的战略规划。

这两个聚类的结合表明，在科技推动社会进步的整体框架下，具体应用（如农业生产、资源开发）和宏观战略（如全球化视野、竞争力提升）之间存在着深刻的联系。科技不仅在具体领域（如农业、资源开发）提升生产力和市场竞争力，同时也在宏观层面为国家或地区提供提升全球竞争力的路径。

**( (4,2), 3) 与 (6,1)**

第2、3、4三句都强调科技进步如何提升农业生产效率、节约资源、扩展市场。它们主要关注科技在农业、农副产品的生产、管理和销售方面的应用。三者之间的关系是：科技不仅能提升生产力（第4句），还可以节约资源和投入（第2句），并最终通过提升市场竞争力来确保经济效益（第3句）。

第6句和第1句讨论的是宏观层面的科技发展战略，不仅关心如何通过科技提升现有资源的利用效率（第1句），而且关注科技如何推动国家和地区在全球化经济中的竞争力（第6句）。这两句话强调科技的作用不仅仅是局部应用，更是全球竞争力的提升。

**(((4,2),3),(6,1)) 和 5**

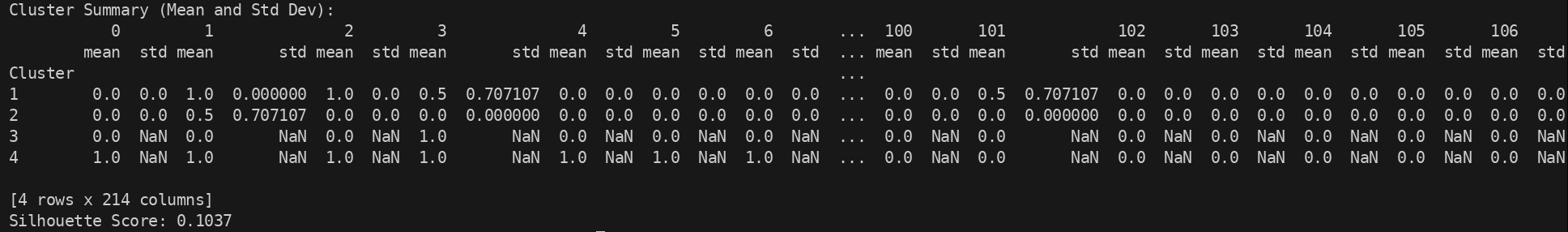
第5句探讨了海洋科技的战略性应用，强调通过科技兴海实现区域经济可持续发展，特别是在海洋产业中的关键技术突破。

这个句子与前面的聚类结果相比，聚焦的是特定区域（海洋经济）的科技发展和应用，其他句子则更广泛地讨论了科技在不同领域（农业、生产等）中的应用。尽管如此，所有这些句子都涉及到科技的战略性支撑作用，因此最终将其与之前的聚类结果结合在一起。

另外，从图3-4来看，Cluster 1 和 Cluster 2 的均值和标准差表现出较强的簇内一致性，说明它们内部的样本在各特征上的表现较为相似。而 Cluster 3 和 Cluster 4 中的 NaN 值可能表明这两个簇中的样本存在缺失数据或高度一致的特征值（如全为零或常值）。

如果某个簇的标准差大，说明这个簇内的样本在某些特征上差异较大，可能是该簇包含了来自不同主题或类别的文本。

Silhouette Score轮廓系数0.1037，轮廓系数的值通常在 [-1, 1] 之间，接近1表示聚类质量较好，接近0表示聚类结果不好，接近-1表示聚类有很大问题。0.1037的轮廓系数相对较低，意味着聚类效果不是很好，可能是簇之间的区分度较低，或者样本在同一簇内的相似度较差。当前的聚类效果（根据轮廓系数）较差，可能需要调整聚类数量、优化特征向量化方法，或考虑使用其他聚类算法。通过进一步分析每个簇的组成和对簇内关键词的分析，可以帮助你更好地理解聚类结果，进而调整聚类过程以提高效果。

图3-4 2012年11月8日以前1层次聚类结果表（我们的方法的关键词）

# 总体任务结果及分析

## 4.1 脉络梳理

我们用饼状图绘制了各时期支持度最高的十条关联规则，如图4-1、图4-2.

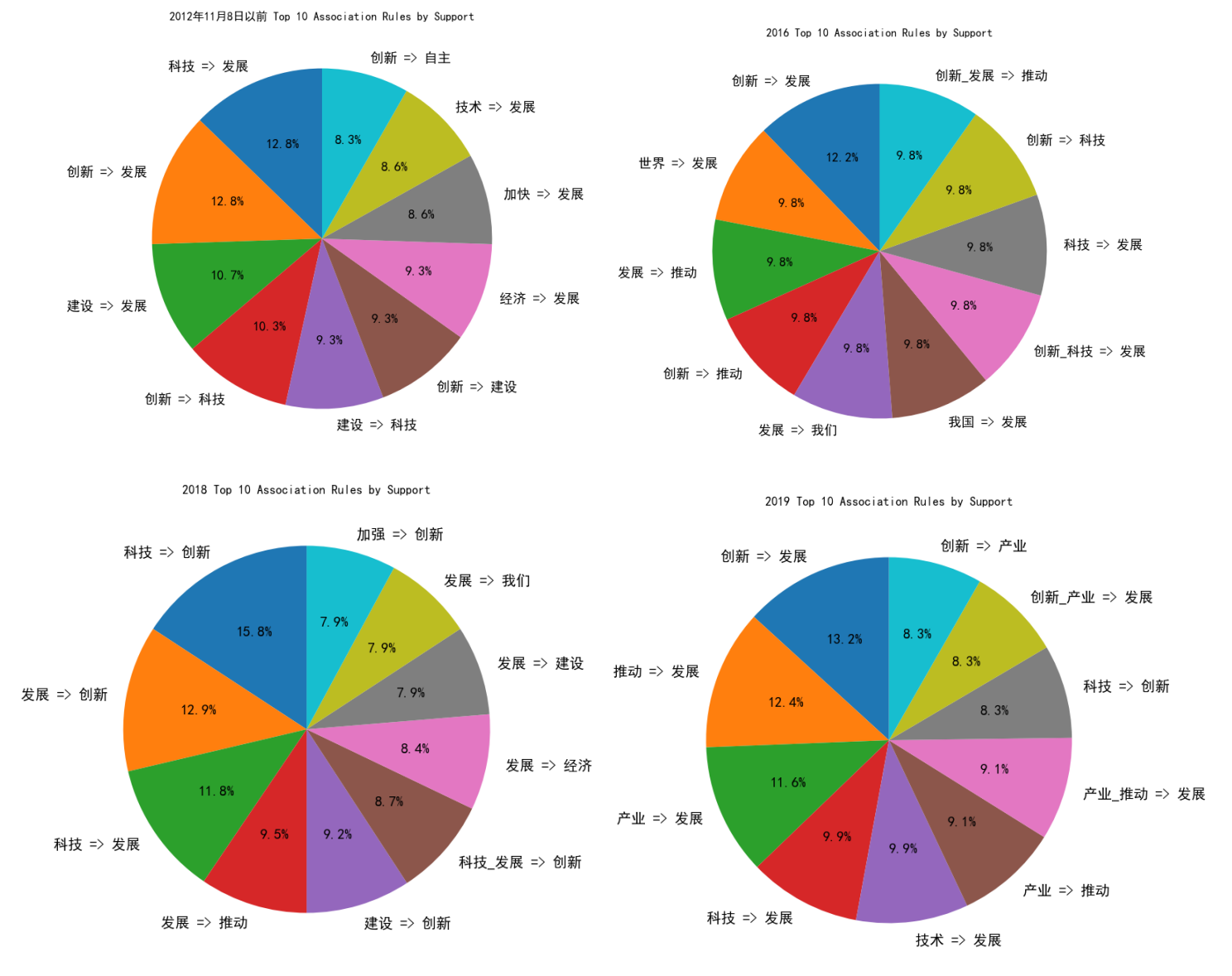


图4-1 2012年11月8日以前~2019年各时期支持度最高的十条关联规则

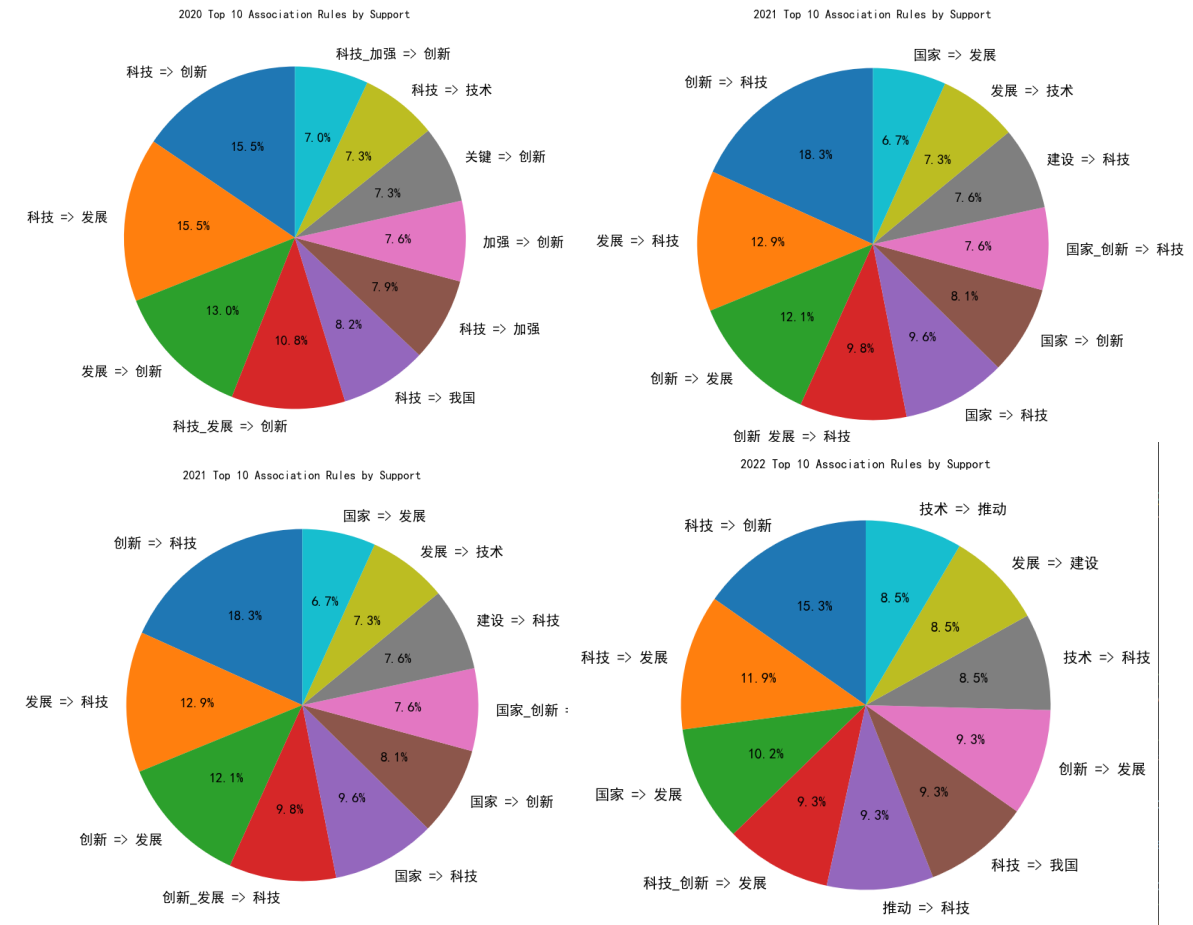


图4-2 2012年11月8日以前~2019年各时期支持度最高的十条关联规则

从2012年到2021年，文档中的关联规则呈现出一些显著的变化。

**"我们要把推动自主创新和实施品牌战略结合起来，以品牌战略带动自主创新，以自主创新支撑品牌战略。"**

最初，在2012年，关联规则主要集中在“科技”与“发展”之间的关系，如“科技”频繁出现在“发展”的前提或结果中。这表明在此时期，推动科技发展被视为一个重要的议题。

**"创新是实体产业发展的制胜法宝。"**

随着年份的推移，关联规则开始逐渐涉及更多元化的关键词，例如“创新”、“建设”以及“经济”。例如，2016年和2018年的规则显示“发展”和“创新”之间的关系逐渐加强，而“推动”也成为了一个常见的后续概念，这可能反映了政策层面上对推动科技创新的强调。

**"创新决胜未来，改革关乎国运。"**

到了2018年，“科技”和“创新”的关联规则进一步深化，显示出这两个概念在各个领域（如“发展”和“建设”）中的紧密联系。而到了2019年，除了科技和创新，产业和推动也开始成为重要的关联元素，表现出产业发展的推动作用。这一趋势在2020年和2021年继续加强，特别是在2020年，科技和创新的组合规则变得更加突出，同时“加强”也出现在与科技和创新相关的规则中，表明在这一阶段，加强科技创新成为了一个更为突出的主题。

总体而言，从2012年到2021年的文档关联规则变化表明，随着时间的推移，技术创新不仅局限于“科技”与“发展”之间的关系，还扩展到多个领域，如“推动”、“产业”、“加强”等，反映了社会发展中的多元化与综合性趋势。

## 4.2 脉络梳理

在进行关联规则挖掘与层次聚类分析的过程中，我们从两个任务中获得了不同类型的可视化结果，这些结果为我们提供了对数据结构和关系的不同视角，并揭示了潜在的内在联系。

首先，关联规则挖掘通过揭示不同特征之间的强关联关系，帮助我们理解数据中各个项目之间的共现模式。例如，在可视化中，我们可能会看到某些关键词或特征在特定条件下频繁一起出现，形成强关联规则。这种信息对于了解数据中的主要模式、发现潜在的依赖关系以及改进决策支持系统具有重要意义。关联规则的结果不仅反映了数据的表面关系，还能挖掘出潜在的因果关系或潜在的趋势，这对于分类、预测等任务提供了有价值的指导。

另一方面，层次聚类通过将相似的数据样本分组，帮助我们理解数据在高维空间中的分布情况。聚类结果可视化展示了不同样本或句子在特征空间中的相似度，分群情况则揭示了数据的内在结构和潜在类别。通过层次聚类，我们能够发现数据中的自然簇，进而为后续的分类、主题分析或个性化推荐提供依据。

从这两个任务的可视化结果中，我们可以发现两者之间的内在联系。首先，聚类分析的结果可以作为关联规则挖掘的基础，帮助我们识别不同簇之间是否存在强烈的依赖关系。例如，某些特定簇中的样本可能表现出特定的特征组合或关键词，这些组合在关联规则中可能形成强规则。因此，聚类结果可以为关联规则挖掘提供潜在的分组信息，使得规则的发现更有针对性，也能揭示每个簇的独特属性或行为模式。

此外，关联规则的挖掘可以为聚类分析提供额外的特征信息。通过识别出强相关的特征，可以将这些特征作为聚类分析中的输入变量，进一步提升聚类的精度和有效性。例如，如果某些特征在关联规则中表现出强烈的相关性，这些特征在聚类过程中可能能更好地帮助分离不同簇，提高聚类质量。

综上所述，关联规则挖掘与聚类分析虽侧重于不同的任务，但两者在分析数据结构和模式时具有重要的互补性。通过结合两者的结果，我们不仅可以在全局层面上识别数据的内在关系，还能够在局部层面上深入挖掘特征的潜在联系。这种跨任务的分析为我们提供了更加全面的理解，帮助我们从多个维度发现数据的复杂性和潜在价值。

# 开放性思考

在neo4j可视化一步中，当出现[a b]->[c d]这类多元关系时，我的处理方式是将关键词用下划线连接起来作为新的结点，而这个新结点不会与其中的关键词关联。如[科技,技术]与[科技,技术,发展]，他们不会有关系。可以考虑在他们之间添加包含关系，让知识图谱更丰富，关键词之间的关系更稠密，也能使知识图谱更易使用。

另外，在挖掘不同的文档时，由于文档数据大小的差异，会出现某些文档挖掘时，因为阈值设置不当，导致程序卡死的情况。在少文档的情景下不会造成太大麻烦，但在海量文档的情境下，预测安全阈值或许值得探讨，可以建立适当的模型进行预测。

# 参考文献