# 基于段落信息增益的政策文本主题识别研究

赵一方　裴雷　康乐乐
（南京大学信息管理学院，南京　210023）

摘要：综合性政策文本通常具有多种政策主张的表述，而现有的基于特征词向量的政策主题识别方法一直无法有效分配特定特征词对相似政策主题的“贡献度”。本文提出一种基于段落信息增益的半监督化政策文本主题识别方法，在不损耗基本词向量信息的同时，显著降低了矩阵计算的复杂度，平衡了不同主题间的贡献差异。基于该方法，本文进一步通过对2018年31个省级行政机构的政府工作报告的政策主题强度差异和政策倾向进行测算，测算结果与人工标注结果具有一定的秩相关性。

关键词：政策文本；主题识别；政策倾向；信息增益
中图分类号：TP391.1
DOI：10.3772/j.issn.1673-2286.2018.11.001

在政策扩散和政策比较研究中，研究者越来越注重从历时或跨区域、跨领域的角度量化研究知识内容、政策模式对关联领域的影响，注重大样本的政策文本分析[1]。因而，在量化政策文本研究领域出现一系列以文本数据（textual data）为对象的政策文本量化研究方法，典型如Laver[2]、Benoit[3]等提出的政策词频统计方法、德国柏林社会科学研究所持续开展的政策语料自动编码分析[4]，以及国内部分学者提出的政策计量（policiometrics）方法[5]，尤其是以内容分析方法与词频分析、文本挖掘相结合的方法[6]，发展为政策文本自动解析和大尺度政策分析的主要方法体系。

在以往政策主题研究中[7]，政策文本自动分析作为一类非介入性研究方法，在海量政策文本分析中已经得到广泛采纳，尤其在选举研究、政策立场与政策倾向分析上体现出一定的方法优势。同时，在具体方法的使用中仍然存在较大争议。一是文本分析领域的若干典型方法在政策文本分析领域均有明显不足。如TF-IDF容易受罕见低频词的干扰、LDA方法的政策主题的可解释性并不理想。而通过词间网络（如共现网络）、引证耦合所形成的潜在语义挖掘中容易缺失权重或强度等重要表意信息。二是政策文本分析的若干特定需求，传统文本分析方法难以满足。如以词频为基础的文本解析方法，无法表征语词在特定政策领域的“语用”意义。而在综合性政策文本分析中，相似的政策主题同时蕴含于政策文本，且邻近政策主题之间极有可能“共享”大量特征词[8]，因此传统基于词向量的政策主题抽取方法往往因提取“最大特征向量”、截频等操作方法而无法有效分配一个具体的特征词对不同政策主题的“贡献度”，存在政策主题的强度统计“偏差”。基于此，本文提出一种基于段落信息增益的半监督化政策文本主题识别方法，在不损耗基本词向量信息的同时，显著降低了矩阵计算的复杂度，平衡了不同主题间的贡献差异。

## 1　相关研究回顾

### 1.1　典型政策文本主题分析方法

文本主题分析广泛应用于主题识别、语义网络建构、情感判断、趋势分析等领域，产生了大量经典的研究方法与工具。一般认为，典型的政策文本主题分析方

法包括统计学、语言学、半结构化文本特征和机器学习等类型。

典型的统计学方法包括以TF-IDF技术为代表的词权重方法[9]和基于词所在文档位置特征方法。这种高度可泛化性的方法易于实现，但是没有考虑到词汇在文档集合类间和类内的分布[10]。将词汇看作单独的个体，无法反映词汇的语义信息，更无法区分不同语境下词汇表征的信息，一般应用于大尺度、粗粒度的文本信息抽取，对于细粒度、个体化的主题跟踪和研究则有所不足。

语言学方法是基于词法分析、句法分析、语义分析等语义角度，通过词义词性词典[11]、词汇链[12]、语义传递关系识别方法[13]对政策的主题进行识别。如马费成等学者[14]建立政策引用类型语义结构提取政策之间的关系，识别政策文本之间的引用主题，基于语义的政策文本方法应用较少，其难点在于词典和规则的批量建立和规则完备性保证。

半结构化文本特征方法通过分析文本所具有的规律化外在特征，根据文本的发布时间、标签、引用关系等结构化特征作为文本主题分析的主要依据，比如王星等[15]基于文献之间的引用关系，以引用行为下的知识关联作为文本主题依据，通过改进的遗传算法实现文献的自动标引，而政策文本的潜在引用行为相对复杂，通过外在特征一般不容易把握文本主题的变迁[16]。

机器学习方法能够基于训练文本数据的文本主题标引先验知识对无主题文本进行标引。如白如江等[17]基于科技文本多元核心特征提取策略通过支持向量机的方法训练分类器模型，对句子级别的科技文献进行标引，但是政策文本数据量较小，主题在不同时间窗口下的表述会产生较大变化，以过去的文本衡量新文本主题可靠性会大幅降低。基于大训练样本的机器学习方法存在样本缺失的问题，而以LDA为代表的非监督主题生成方法对主题的解释性较弱，这些自动标引方法极大减轻了标引任务量，但对政策文本这种半结构化特殊文本类型并不完全适用[18]。

因此，一直以来在政策文本主题识别领域并没有完全适用的单一方法，往往根据具体任务、具体文本特征选择一种或多种方法予以组配使用。

### 1.2 典型政策主题分析工具

随着政策文本的定量处理越来越关注政策语义的分析，一些注重文本价值的定量描述和潜层语义的知识发现工具得以引入，典型的工具如WordScore和WordFish等[7]。WordScore是德国柏林社会科学研究所在政党比较研究项目中开发出来的政策主题词表系统，主张运用政策编码、政策概念词表或政策与语词之间的映射关系进行政策概念的自动处理，进而统计分析政策倾向（preference），目前已经涵盖超过50个国家的4 051个语料集的794 536个政策术语或词条。WordScore方法通过考察词语在先验文本中的概率分布，计算测试文本的政策价值，描述政策文本价值表现力[2]。从算法准确度来看，相较于一般选民的政策主题排序，WordScore方法的政策主题排序更接近于专家主题分析排序[19]；裴雷等[20]将WordScore方法应用于中文政策语言环境，实证表明该方法在比较不同政策来源中相同综合政策主题是有效的。与此相对应，Slapin等[21]指出词与政策倾向的依赖关系并不稳定，而是假定政策立场依赖于文本特征词的分布特征，据此提出无监督的政策倾向计算方法WordFish。

在后续的研究中，学者发现WordScore方法的前提是政策特征词相互独立，忽视了特征词在不同政策语句下的语用表义差异，尤其是段落的词语之间存在语义依存关系，使得词语无法兼顾政策语境的差异性。因此，Proksch和Slapin[22]认为政策文本处理算法的最大缺陷是对政策文本语言特征、文本结构和语境适用性的关注缺失。尽管政策自动分析的有效性得到了学者的证明，但主流政策分析领域普遍认为政策计算分析的结果是非精确性的[3]，加强政策的解释性分析，并融合质性方法的混合方法更具有应用前景[7]。

综上所述，虽然国内外学者提出大量不同的政策文本主题分析方法，但是在政策文本语言特征、文本结构的文本主题分析方法层面的探索仍不充分，尤其在中文政策文本语用信息中，研究方法与应用效果均有待进一步探讨。

## 2 研究思路和研究设计

本文在信息增益算法[23]的基础上，借鉴了信息熵[24]的计算方法，提出引入专家质性知识的基于段落信息增益计算方法的半监督政策文本主题分类模型。具体而言，本文引入了段落作为政策的语义容器，将语词-主题的二元稀疏弱关系转化为语词-段落的非稀疏弱关系、段落-语义的主题关联强关系两阶段的矩阵乘积模式，用以增强语词或段落的语义表达效果。该方法假

定任意语词-主题的二元稀疏弱关系均为可分的，而词语-段落的关系矩阵为易测易得的关系数据，则语词-主题的二元稀疏弱关系表达实质就转化为段落-语义的强关联关系表达及其效果评估。

在现有的段落语义分析方法中，一类是根据先验的语义词典，如WordScore词典，将段落中的特征词汇的语义表达强度进行提取和加总，进而生成段落的语义强度；另一类是根据人工判断段落的语义主题归属。本文提出半监督的方法，即通过人工标注少量段落或段落群，识别段落中的词向量特征，进而得到通过少量标注推及大量段落自动标注的方法（见图1）。

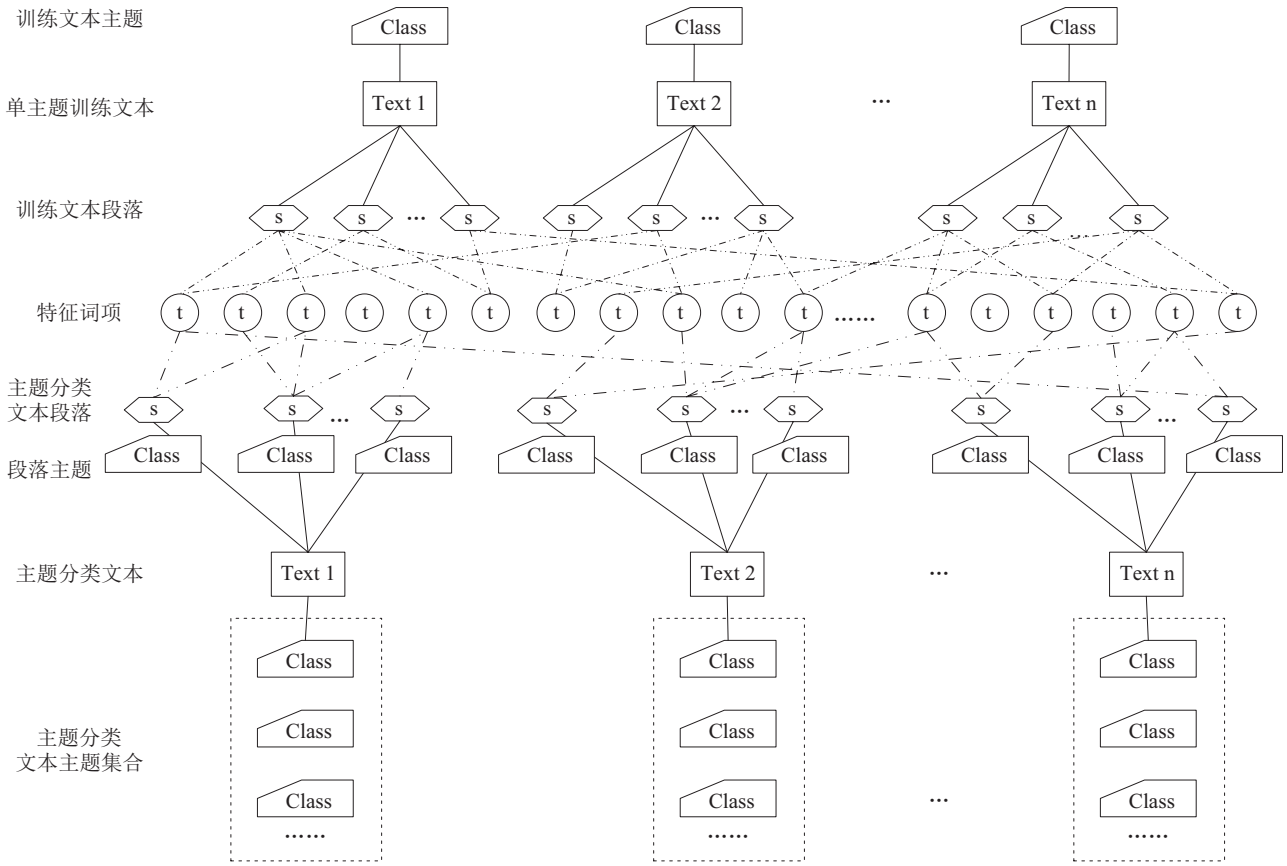


图1 政策文本主题向量生成示意图

## 2.1 段落主题特征与段落信息增益

在政策文本分析中，政策的段落语义体现为政策主题及其表现强度。而基于段落的主题识别主要有两个任务：一是稳健识别段落主题；二是识别段落在政策主题上的表达强度，即政策表现强度。

### 2.1.1 段落主题特征向量与段落频次

为揭示段落对政策主题的表现强度，本文提出两个基本概念。

段落主题特征向量（paragraph term vector，PTV）是由政策文本中名词构成的特征词集及其词频构成的特征词向量。段落主题特征可以描述为全文本中的特征词向量，也可以描述为段落中的特征词向量。但前者因词向量稀疏性会影响相关性测试结果，因此本文选择段落特征词向量作为测试方法。此外，政策文本中的主题和意义往往是由体词表达的，政策文本中的体词包含更多的意义[12]，为了增加算法的普适性，将分词结果中的名词特征项单独提取出来，作为政策文本集的特征向量，将文本转换为向量空间模型，其中模型的特征项集T=$\{t_1, t_2, t_3, \cdots, t_z\}$，$t$表示单个特征词项。

段落频次（paragraph frequency，PF）是特定段落在政策文本中出现的频次统计。从自然语言描述角度

看，政策文本集中可能在语法层面很少出现完全相似的文本，但在语义层面可能存在相似或相容的段落，即具有公共的段落主题特征向量PTV。

本文识别了一种段落相容的情况，并将其定义为段落可见。即段落$s$所包含的PTV均出现在另一段落$s'$中，满足$s\subseteq s'$；那么在文本$d_j$中，若段落$s'\in d_j$，则可声明段落$s$亦包含于文本$d_j$中的某一段落，存在映射$\Phi(s,d_j):s\to d_j$。因而，定义段落相容函数为简单的二值函数见公式（1）。

$$\Phi(s,d_j)=\begin{cases}1, & \text{段落}s\text{在文本}d_j\text{中可见}\\ 0, & \text{段落}s\text{在文本}d_j\text{中不可见}\end{cases} \quad (1)$$

那么，段落$s$在文本$d_j$中出现的概率可描述为公式（2）。

$$Pr(s)=\frac{\left|\{j:\Phi(s,d_j)=1\}\right|}{n} \quad (2)$$

### 2.1.2 段落信息增益

1986年，Quinlan[23]提出采用信息熵和信息增益来衡量特征项对类型归属贡献度，提出了一种基于自信息的分布状态表征类别属性的方法；1997年，Yang等学者[24]将信息增益方法应用于文本分类，并提出了特征项缺失条件下的信息增益改进算法，以互补条件下的条件概率测度特征项缺失时的信息增益，放宽了自信息计算方法中对信息完备性的要求，见公式（3）。

$$Pr(\bar{t}\mid c_i)=1-Pr(t\mid c_i) \quad (3)$$

在Yang信息增益算法的基础上，本文进一步提出以段落作为语义容器来计算段落信息增益（paragraph information gin，PIG），见公式（4）。

$$IG(s)=H(C)-(Pr(s)\times H(C\mid s)+Pr(\bar{s})\times H(C\mid\bar{s})) \quad (4)$$

因段落作为词集的一种表现形式，不同于特征词的彼此独立，段落之间存在相关、相似，甚至相容关系，并不存在完全独立假设。在段落描述尺度下，采用特征词向量作为段落表达形式，形成事实上的两种“效果”：第一，段落之间具有大量相似的特征词或相似的特征词分布，存在相似段落；第二，不同段落对特定主题的信息增益效果不同，造成相同的特征词因处于不同的段落，而对该主题的“表意效果”不同，这点不同于传统的词向量方法，也是本文试图解决的词语的“上下文”对语义的影响。

具体算法如下所示。

算法名称：段落信息增益算法

输入：政策文本训练集（包括单主题文本、政策文本主题）

输出：段落主题及段落信息增益、特征词项知识库

步骤：

Step1：预处理政策文本训练集，对文本进行分词，提取文本词项特征，将特征词项整理输出为特征词项知识库。

Step2：计算主题分类系统的信息熵$H(C)$。

Step3：将政策文本按照段落切分，预处理段落文本，标识段落来源文本及段落所在位置；基于特征词项知识库将段落文本转变为特征词向量，建立段落文本向量空间模型。

Step4：遍历每篇政策文本的段落向量，将政策文本主题赋予段落主题，通过公式（4）计算段落在该主题下的政策表达强度，建立段落的政策主题和政策表达强度与段落向量的链接。

## 2.2 段落政策主题特征及政策主题识别

本文选取同质化程度较高的政策文本作为训练样本，在语言统计层面发现文本中段落长度分布较为均匀，而作为测试政策文本语料的段落长度也相似，同时*Jaccard*系数与信息增益均考察特征词项在文本中出现的可能性[25]，因此采用*Jaccard*系数作为政策段落相似的主要判断依据。

### 2.2.1 段落主题表现强度

段落主题表现强度（paragraph topic representation，PTR）是段落对特定主题（Topic）的表现强度，可由段落对该主题中所有与该段落相似或相容的段落分布的信息增益来表征。信息增益值越大，反映测试段落对该主题的特征描述贡献越大，越能作为判断主题相似的依据。通过PTR的计算，能够识别出段落对主题的表意强度。

具体而言，为计算段落$s$与主题分类知识库S的相似度，选取训练样本集中与段落$s$最相似的段落记作$s^*$，段落$s$的政策主题与$s^*$一致，段落$s$的政策主题表现强度$PTR_s$的计算方法见公式（5）。

$$PTR_s=IG(s^*)\times Jaccard(s,s^*) \quad (5)$$

$PTR_s$的数值越大表示段落$s$的政策主题表现力越强。

### 2.2.2 政策主题识别

在政策文本集中，每个文本由若干段落组成，政

策文本的主题多样性是由段落文本的主题多样性构成的，因此政策文本*police*主题的定义见公式（6）。

$$Topic(police \mid c_i) = \sum_{s \in police} \mathrm{PTR}_s \times col(s, c_i) \quad (6)$$

其中，$Topic(plice \mid c_i)$表示政策在主题类别$c_i$的表达强度；$s$表示政策的每一个段落；$col(s, c_i)$表示段落$s$与主题类别$c_i$之间的关系，如果类别$c_i$是段落$s$的政策主题，那么$col(s, c_i)=1$，否则，$col(s, c_i)=0$。

基于主题类别集合$C=\{c_1, c_2, c_3, \cdots, c_k\}$，政策文本被表示为关于主题类别的向量。对于每一篇政策文本，在每一个主题类别上都具有大于或等于0的主题强度，设定主题强度阈值$\alpha$，$\alpha$表示最低主题强度，用于控制政策文本的主题数量，在主题类别向量中筛选出主题强度大于$\alpha$的主题作为政策文本的主题。

## 3 实验过程与结果

### 3.1 政策文本处理流程

在处理流程上，①将政策文本按照回车符划分为段落的形式，标记每个段落的文本来源与段落序号，对段落文本分词处理；②将标引好训练文本按照段落划分，建立主题分类知识库；③计算当前段落与主题分类知识库中文本特征的相似度，对当前段落赋予主题并计算主题强度；④将政策文本段落主题汇总，将政策文本映射为政策主题向量，通过主题强度累加的方式计算政策文本的主题强度，设置主题最低强度，将高于最低强度的主题赋予政策文本。处理流程如图2所示。

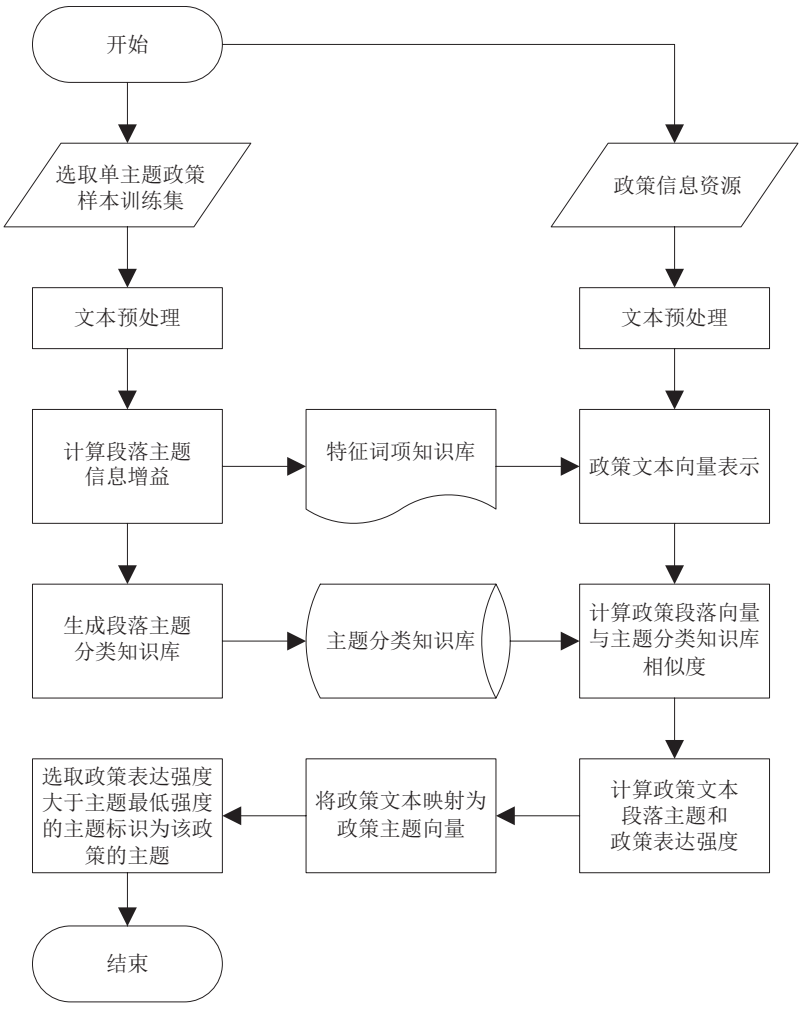


图2 政策文本主题分类模型流程图

## 3.2 训练样本集及其文本特征

假定政策主题主要依托其发布机构的类型属性，即科学技术部发布的训练样本集中，认为其语篇的主题属性可描述为科技类政策。依此假定，本文选取科学技术部、国家民族事务委员会、教育部、文化和旅游部等9部委的政策文件作为主题分类训练样本，训练样本内容包含政策名称、政策主题和政策全文，共包含88篇政策样本，2 905个段落（如表1所示）。

表1 训练样本政策主题分布

| 主题类别 | 政策数量/篇 | 段落数量/个 |
|---|---|---|
| 公共卫生 | 10 | 189 |
| 农业 | 10 | 382 |
| 教育 | 10 | 491 |
| 文化 | 10 | 664 |
| 贸易 | 10 | 242 |
| 民族宗教 | 8 | 54 |
| 法治 | 10 | 417 |
| 生态环境 | 10 | 283 |
| 科技 | 10 | 183 |

## 3.3 段落信息增益方法的主题识别效果

为检验段落信息增益方法的主题识别效果，本文采用Hjorth[19]、Fieller[26]等对自动文本分析效果中提出的spearman相关系数法。Hjorth等[19]分别测算对CMP RILE measure政治演讲语料库的自动分析方法所检测的段落主题编号序列和专家分析、选民分析的主题编号序列的spearman相关系数。结果显示，自动分类与专家分类一致性高于选民与专家分类的一致性，进而证实自动分类法优于一般选民的政策主题识别效果。

笔者以人工标注模拟专家分类，以WordScore算法作为本研究的参照算法，以主题强度排序作为spearman相关排序的依据。随机抽取5篇政府工作报告作为测试集，采用人工编码的方法对5省政府工作报告的共414段政策进行主题标引，统计测试文本中9个政策主题的表现强度；同样测算PIG算法和WordScore算法对政策主题的强度排序（见表2）。其中，T1至T9分别为公共卫生、农业、教育、文化、民族宗教、法治、生态环境、科技、贸易9个政策主题（下同）。

综合而言，虽然PIG算法整体与人工标引的秩相关系数约0.66，并非特别突出；但相对于WordScore算法

表3 秩排序及相关度计算结果

| | 编码者 | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | $R_h$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TESTPOLICY1 | 人工标引 | 6 | 3 | 2 | 8 | 5 | 9 | 4 | 7 | 1 | - |
| | PIG | 6 | 1 | 3 | 9 | 8 | 5 | 2 | 7 | 4 | 0.63 |
| | WordScore | 7 | 1 | 5 | 4 | 8 | 9 | 2 | 6 | 3 | 0.60 |
| TESTPOLICY2 | 人工标引 | 4 | 5 | 1 | 6 | 9 | 8 | 7 | 3 | 2 | - |
| | PIG | 8 | 2 | 1 | 4 | 9 | 5 | 6 | 7 | 3 | 0.53 |
| | WordScore | 7 | 1 | 4 | 2 | 8 | 9 | 3 | 6 | 5 | 0.28 |
| TESTPOLICY3 | 人工标引 | 7 | 2 | 1 | 6 | 9 | 8 | 4 | 5 | 3 | - |
| | PIG | 7 | 3 | 2 | 6 | 9 | 4 | 1 | 8 | 5 | 0.67 |
| | WordScore | 7 | 1 | 4 | 2 | 9 | 8 | 3 | 6 | 5 | 0.73 |
| TESTPOLICY4 | 人工标引 | 6 | 2 | 3 | 7 | 9 | 8 | 4 | 5 | 1 | - |
| | PIG | 7 | 2 | 3 | 5 | 9 | 8 | 1 | 6 | 4 | 0.83 |
| | WordScore | 7 | 1 | 5 | 6 | 8 | 9 | 3 | 4 | 2 | 0.90 |
| TESTPOLICY5 | 人工标引 | 7 | 4 | 2 | 5 | 9 | 8 | 6 | 3 | 1 | - |
| | PIG | 5 | 2 | 1 | 6 | 9 | 7 | 4 | 8 | 1 | 0.67 |
| | WordScore | 7 | 1 | 5 | 2 | 9 | 8 | 4 | 6 | 3 | 0.63 |

的平均秩相关系数0.60，PIG算法在政策主题识别领域仍表现出微弱的优势。尤其是从测试样本中，发现基于段落信息增益的PIG主题识别方法稳定性要显著优于WordScore算法，其识别精度（以相关性考量）的标准差为0.08，远小于WordScore方法的0.2。

## 3.4 2018年省级政府工作报告的政策主题结构区域差异

在综合性政策主题分类中，以往多以频次截断方式获取政策主题的近似表达，而且LDA生成的主题表达强度的可解释性并不理想。PIG算法中，以段落多测试主题的信息增益贡献为标准，可以近似为后验关联概率，具有实际价值，因而不仅能实现政策主题的归属判定，而且能通过加总、加权实现对政策主题结构的量化解析。本文选取除香港特别行政区、澳门特别行政区、台湾地区外的2018年政府工作报告省级政府工作报告文件作为测试样本，分别测算了不同区域的政府工作报告在不同政策主题的表达强度差异与整体结构差异（见表3）。

从政策强度可观察各省份的政策主题结构的区域差异与倾向。以$\alpha$=0.15为政策强度阈值，TOP3作为各省级行政区域2018年政府工作报告的政策文本主题表达，可识别各省份的政策主题倾向如表4所示。

更进一步，根据政策主题及其表达强度的分布特

表3 政策主题表达强度的区域差异

| | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 |
|---|---|---|---|---|---|---|---|---|---|
| 山西 | 0.0494 | 0.1420 | 0.1840 | 0.0493 | 0.0056 | 0.0454 | 0.1918 | 0.0729 | 0.1649 |
| 广东 | 0.0156 | 0.2165 | 0.2818 | 0.0491 | 0.0098 | 0.0284 | 0.0740 | 0.0541 | 0.1468 |
| 青海 | 0.1169 | 0.1355 | 0.1715 | 0.0177 | 0.0000 | 0.0285 | 0.0957 | 0.0360 | 0.0557 |
| 重庆 | 0.0214 | 0.0805 | 0.0815 | 0.0099 | 0.0050 | 0.0473 | 0.0924 | 0.0473 | 0.0804 |
| 北京 | 0.0810 | 0.0993 | 0.1525 | 0.0205 | 0.0111 | 0.0404 | 0.1843 | 0.0458 | 0.0811 |
| …… | …… | …… | …… | …… | …… | …… | …… | …… | …… |
| 海南 | 0.0201 | 0.1004 | 0.0821 | 0.0222 | 0.0137 | 0.0580 | 0.0986 | 0.0182 | 0.1104 |

表4 2018年省级政府工作报告的政策主题倾向识别

| 省/市/自治区 | 政策主题倾向 |
|---|---|
| 山西 | 生态环境、教育建设、贸易往来 |
| 广东 | 教育建设、农业建设、贸易往来 |
| 海南 | 贸易往来、农业建设、生态环境 |
| 青海 | 教育建设、农业建设、公共卫生 |
| 陕西 | 教育建设、农业建设、公共卫生 |
| 重庆 | 生态环境、教育建设、农业建设 |
| 北京 | 生态环境、教育建设、农业建设 |
| 福建 | 农业建设、贸易往来、教育建设 |
| 江苏 | 教育建设、生态环境、法制建设 |
| 辽宁 | 教育建设、农业建设、生态环境 |
| 上海 | 教育建设、农业建设、贸易往来 |
| 湖北 | 生态环境、教育建设、贸易往来 |
| 河南 | 公共卫生、教育建设、贸易往来 |
| 湖南 | 生态环境、贸易往来、科技发展 |
| 四川 | 农业建设、贸易往来、生态环境 |
| 河北 | 生态环境、教育建设、贸易往来 |
| 云南 | 生态环境、教育建设、农业建设 |
| 安徽 | 教育建设、贸易往来、生态环境 |
| 甘肃 | 农业建设、生态环境、贸易往来 |
| 黑龙江 | 生态环境、农业建设、公共卫生 |
| 山东 | 生态环境、农业建设、贸易往来 |
| 贵州 | 公共卫生、教育建设、贸易往来 |
| 吉林 | 农业建设、生态环境、教育建设 |
| 江西 | 农业建设、贸易往来、教育建设 |
| 天津 | 农业建设、教育建设、生态环境 |
| 广西 | 生态环境、农业建设、教育建设 |
| 西藏 | 农业建设、生态环境、教育建设 |
| 宁夏 | 生态环境、教育建设、农业建设 |
| 内蒙古 | 教育建设、农业建设、贸易往来 |
| 新疆 | 农业建设、生态环境、教育建设 |
| 浙江 | 教育建设、农业建设、贸易往来 |

征，可通过层次聚类的降维方法（见图3），将2018年省份政府工作报告体现出5种区域类型。

（1）突出公共建设服务优先的政策导向。以陕西、青海、贵州、河南为代表，在公共卫生环境、公共文化服务、公共基础设施等公共建设服务领域着墨更多，突出强调了放管服与政府服务建设在年度工作报告中的施政力度。

（2）着重强调生态环境建设力度。以新疆、甘肃、山东、山西、河北为代表，生态友好型政策导向并不突出，这类区域的政策主题发展比较均衡和全面，对生态环境的相对关注占比略显突出。在区域分布特征上，也发现以自然环境保护为中心的新疆、西藏等生态大省，以人居环境保护为中心的山东、四川等人口大省，以及具有典型能源钢铁等重污染产业布局的山西、河北等生态弱省，是这类政策输出省/自治区的主要特征。

（3）注重外向型经济发展。以北京、云南、上海为代表的政府工作报告强调商贸与对外出口，以北京、上海为代表的经济发达地区，以黑龙江、云南为代表的边境贸易地区，以山西、河北为代表的产业链上游地区，相较于其他省份，这3类地区更强调外向经济与对外贸易对本区域经济发展的战略地位。

（4）突出农业建设等区域特色的政策类型。以辽宁、吉林、江西为代表，突出地方产业特色；辽宁、吉林均是中国的农业大省，集中了国家大量的商品粮基地，农业政策的稳定是该区域的施政基础；而江西省在产业结构上，农业占比也较高，基本反映了其主要政策着眼点与地方特色。

（5）主张创新驱动型政策。以江苏、湖南为代表，二者均重视科技与教育，并不同于全国其他地区的政策主题结构的分布，比较强调结合自身优势和特点推行相关政策。

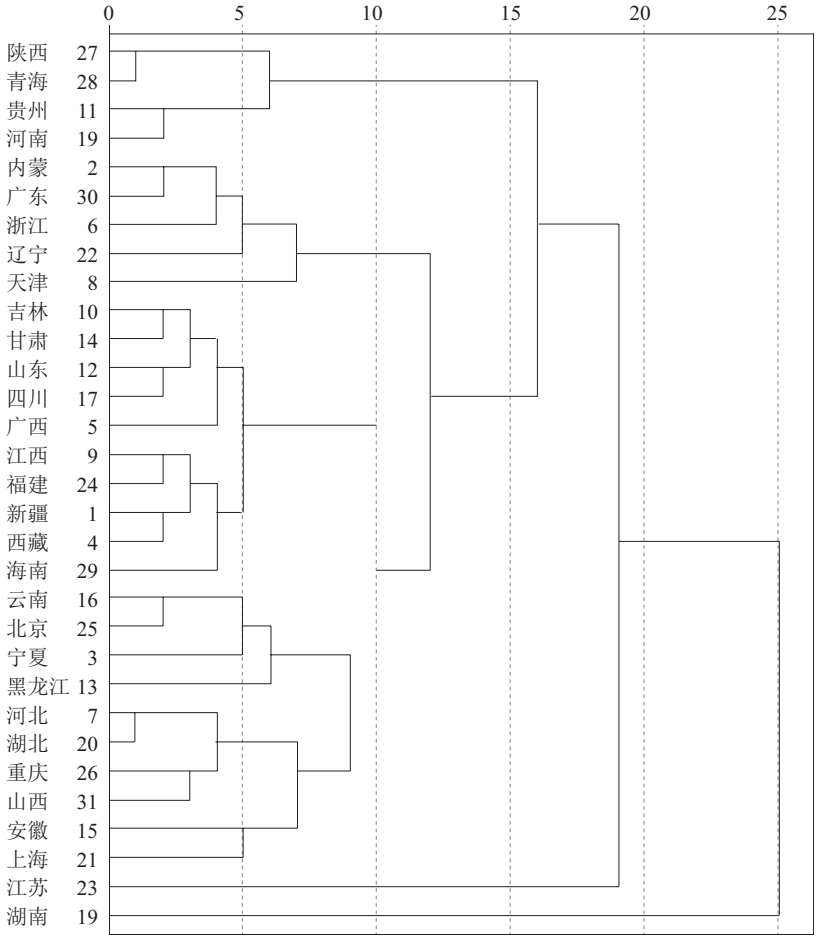


图3 2018年省级政府工作报告区域类型聚类分析

## 4 结论与不足

本文基于政策文本段落间的语义相似特征，提出了一种可用的、具有一定容错性的基于段落信息增益（PIG）的政策主题识别方法。相对于以往的主题识别方法，该方法体现出两个典型特征：一是计算量不大，并能有效解决特征词向量主题识别中的稀疏矩阵问题，充分考虑了低频词在政策主题中的表现强度；二是体现出主题判定的稳健性，相对于WordScore算法更接近于人工判定结果。此外，在功能上也提供了一种面向综合政策主题的结构解析方法，通过政策主题表达强度、政策主题表达结构的分析，提供了更高维度的政策解析方法。同时，在研究和设计过程中也存在一些不足。如段落之间的关系可以更加精细化表达，而非二值化处理；政策主题的设定依赖于先验政策文本集的规模和质量，对于新兴政策话题可能并不一定具有很好的识别效果。

总之，本文提出并验证了一种可能用于大规模综合性政策文本主题比较的解析方法和框架，未来将进一步用于典型的非综合性政策文本解析，用以判定专题性政策中相近政策主题的识别效果。

### 参考文献

[1] 施茜，裴雷，李向举，等. 信息化政策理论与实践的交互扩散研究——以江浙信息化政策样本为例[J]. 情报学报，2016(10)：1081-1089.

[2] LAVER M，GARRY J. Estimating policy positions from political texts [J]. American Journal of Political Science，2000，44(3)：619-634.

[3] BENOIT K，LAVER M，MIKHAYLOV S. Treating words as data with error：Uncertainty in text statements of policy positions [J]. American Journal of Political Science，2009，53(2)：495-513.

[4] VOLKENS A，LEHMANN P，MATTHIE T，et al. The Manifesto Data Collection. Manifesto Project（MRG/CMP/MARPOR）[EB/OL]. [2018-10-01]. https://visuals.manifesto-project.wzb.eu/mpdb-shiny/cmp_dashboard_dataset/.

[5] 黄萃，任弢，张剑. 政策文献量化研究：公共政策研究的新方向 [J]. 公共管理学报，2015（2）：129-137.

[6] GRIMMER J，STEWART B M. Text as data：the promise and pitfalls of automatic content analysis methods for political texts [J]. Political Analysis，2013，21（3）：267-297.

[7] 裴雷，孙建军，周兆韬. 政策文本计算：一种新的政策文本解读方式 [J]. 图书与情报，2016（6）：47-55.

[8] NAMENWIRTH J Z. Some long and short term trends in one American political value：A computer analysis of concern with wealth in 62 party platforms [J]. Historical Methods A Journal of Quantitative & Interdisciplinary History，1967，1（1）：6.

[9] SALTON G，BUCKLEY C. Term-weighting approaches in automatic text retrieval [J]. Information Processing & Management，1987，24（5）：513-523.

[10] 赵京胜，朱巧明，周国栋，等. 自动关键词抽取研究综述 [J]. 软件学报，2017，28（9）：2431-2449.

[11] 刘宏哲，须德. 基于本体的语义相似度和相关度计算研究综述 [J]. 计算机科学，2012，39（2）：8-13.

[12] 索红光，刘玉树，曹淑英. 一种基于词汇链的关键词抽取方法 [J]. 中文信息学报，2006（6）：25-30.

[13] HORROCKS I，GOUGH G. Description logics with transitive roles [J]. Universite Paris-Sud，1997，38（4）：25-28.

[14] 马费成，李小宇，张斌. 中国互联网内容监管体制结构、功能与演化分析 [J]. 情报学报，2013，32（11）：1124-1137.

[15] 王星，刘伟. 基于引文的中文学术文献自动标引方法研究 [J]. 图书情报工作，2014，58（3）：106-110，105.

[16] 李江，刘源浩，黄萃，等. 用文献计量研究重塑政策文本数据分析——政策文献计量的起源、迁移与方法创新 [J]. 公共管理学报，2015（2）：138-144，159.

[17] 白如江，王晓笛，王效岳. 基于支持向量机和核心特征词的科技文献自动标引研究 [J]. 情报理论与实践，2014，37（7）：129-134.

[18] HOFMANN T. Unsupervised learning by probabilistic latent semantic analysis [J]. Machine Learning，2001，42（1/2）：177-196.

[19] HJORTH F，KLEMMENSEN R，HOBOLT S，et al. Computers，coders，and voters：Comparing automated methods for estimating party positions [J]. Research & Politics，2015，2（2）：1-9.

[20] 裴雷，孙建军. 基于WordScore原理的信息政策价值评价模型与方法 [J]. 数字图书馆论坛，2011（8）：61-70.

[21] SLAPIN J B，PROKSCH S O. A scaling model for estimating time-series party positions from texts [J]. American Journal of Political Science，2008，52（3）：705-722.

[22] PROKSCH S O，SLAPIN J B. How to avoid pitfalls in statistical analysis of political texts：the case of Germany [J]. German Politics，2009，18（3）：323-344.

[23] QUINLAN J R. Induction of decision trees [J]. Machine Learning，1986，1（1）：81-106.

[24] YANG Y，PEDERSEN J O. A Comparative Study on Feature Selection in Text Categorization [C]//Proc. International Conference on Machine Learning. 1997：412-420.

[25] JACCARD P. The distribution of the flora in the alpine zone [J]. New Phytologist，2010，11（2）：37-50.

[26] FIELLER E C，HARTLEY H O，PEARSON E S. Tests for rank correlation coefficients. I [J]. Biometrika，1957，44（3）：470-481.

## 作者简介

赵一方，女，1994年生，硕士研究生，研究方向：自然语言处理与政策文本分析，E-mail：mf1714067@smail.nju.edu.cn。
裴雷，男，1981年生，博士，教授，研究方向：信息资源管理与政策分析。
康乐乐，男，1987年生，博士，副教授，研究方向：协同创新与政策研究。

## A New Method of Topic Detection in Hybrid Policy Documents Based on PIG

ZHAO YiFang PEI Lei KANG LeLe
( School of Information Management, Nanjing University, Nanjing 210023, China )

Abstract: In hybrid policy documents, a number of policy topics being mixed in context may not be completely extracted or computed by the former algorithms based on featured terms. Thus the paper tried to propose a semi-supervised subject classification method and a subject intensity calculation method based on paragraph information gain. In methodology test, 31 provincial government reports issued in 2018 were chosen as test samples, and a significant relevance was observed between automatic topic identification and expert tagging.

Keywords: Policy Text; Topic Detection; Policy Preference; Information Gain

（收稿日期：2018-10-25）