# 政府公开信息自动标引的设计与实现

江华丽[1] 曹祺[2] 陈刚[1]
(1. 武汉大学国家网络安全学院，武汉 430072；2. 灰视集团有限公司，北京 100080)

摘要：目前政府信息公开主要依据《中华人民共和国政府信息公开条例》，但站在用户视角，需要根据不同的使用场景进行适配，因此对公文进行自动化标引具有重要意义。本文基于自然语言处理技术，通过词频、词性和词义的实验和分析，提炼公文标题中的范式，对国务院1969—2018年的4 388条公文进行自动化标引。其中以地域关键词和行业关键词为例进行标引，标引后提炼相关关键词可以供相关渠道进行搜索和二次加工。本文主要处理标题的标引，尚未对全文进行标引。

关键词：政府信息公开；文本挖掘；自动化标引
中图分类号：G350
DOI：10.3772/j.issn.1673-2286.2019.01.006

## 1 研究背景

政府信息公开制度，是确保关联方及时获悉和学习国家方针政策，了解政府工作动态，进而有效执行的前提和基础。鉴于该制度的重要性，2007年4月5日，国务院通过《中华人民共和国政府信息公开条例》（国令492号）[1]，并于2008年5月1日起正式实施。李盛[2]指出，“如果政府和群众之间的信息沟通渠道不畅通，就可能引发社会恐慌，甚至造成严重的社会危机”。国务院网站公布的文件，是按照政府信息公开条例目前的执行标准《政务信息资源目录体系》（GB/T 21063），该标准的主要作用在于尽可能相对完整和科学地保存政府公文的内容信息。而广大普通用户通常是通过大众的媒体渠道，如各种搜索引擎和新媒体等方式获取公开海量的政府信息，这也需要对相关公文进行自动化标引。同时，具体通过标引来添加相关标签，以便更好地被用户阅读访问和二次加工。

## 2 研究现状

目前，对于政府公开信息的自动化标引研究主要分为两类研究方法：一类是基于计算机相关技术，先从技术上来分析词频、词义，然后进行相关的标引，最后由行业专家进行修正；另一类是基于行业知识，先由行业专家进行分析和加工，进而进行人工或者计算机类辅助分析。本文主要是第一类研究方法。

对于第一类研究方法，贾君枝等[3]对比分析了各种自动标引方法的优缺点，将标引法分为词典标引、统计标引、单汉字标引、语义标引、神经网络标引和专家系统标引。本文的研究在词典标引的过程中，核心是元数据的分析和解析。在公文元数据分类领域，张新民等[4]对比了中国、英国、美国的政府信息公开的元数据定义指南，本研究参考了他的元数据定义规范。

第一类研究方法的核心是生成词库和标引流程，吴洁明等[5]设计流程图和标引流程，主要分为词库维护、自动化标引模块和人工修正模块。基于不同政府信息的词库可以挖掘出不同的结论，如邓雪琳[6]通过对政府工作报告中的关键词、高频词、关键段落字数开展计量，提出了相同同意度和信度两种测量模式，回溯性地测量了改革开放以来中国政府职能转变的特点，并预测了中国政府职能未来转变的趋势。在基于计算机技术建立词库的过程中，除根据政府公开的文件建立主题词库外，还可以基于相关的期刊进行词库建立。利用这些基于期刊的关联词库来对政府公文的词库进行修正。如朱晓峰等[7]选取了2006—2015年WOS和CNKI数据库，利用citespace分析了相关政府公开信息的主题

词，研究相关趋势并生成相关主题词库。

对于第二类研究方法，王志刚[8]基于文本挖掘提出理念词频率概念，即在每万字的政府工作报告中某政府理念出现的次数，他通过政府公文计算其理念词频率，进而分析理念词频率和实际经济增速的变化。然而，理念词的内涵首先需要由相关行业的专家来定义，为此，这种分析的难度就在于行业专家定义词库本身。潘松[9]分析了公文中的汉语成语，基于汉语成语的语义来判断分析公文的褒贬色彩，但这种分析需要行业专家对公文中的成语进行预处理和定义。程大荣[10]通过政府公开刊载的意见统计来分析意见的使用频率、发文机关、发文形式、文种名称前缀修饰词、正文内容层次等5个方面的情况，但这种分析同样需要行业专家对发文机关的业务有一定的认识和理解。

## 3 实验与分析

### 3.1 数据分类

本文的政府公文实验数据来自国务院官方网站的“政府信息公开”专栏。国务院公文数据分为公文元数据和公文全文数据。其中，公文元数据包含索引号、主题分类、发文机关、标题、发文字号、发布日期、成文日期、主题词8种类别。本文研究发现，公文的索引号并不是唯一标识符，如国函〔2016〕64号和国办函〔1992〕4号这两份公文的索引号均为000014349/2016-00057，但它们的发文字号是唯一标识符。

发文字号主要分为国令、国发、国函、国发明电、国办发、国办函、国办发明电和其他8种类别。发文机构分为国务院和国务院办公厅，其中，“国发”代表国务院发文，“国办发”代表国务院办公厅发文，国务院发文的权威性高于国务院办公厅发文。从发文类别上看，国务院的发文分为国令、国发、国函和国发明电4类。

发文分为自上而下制定和自下而上制定两类。自上而下制定是指基于现有法律制定的条例，如针对全国人大通过的《中华人民共和国个人所得税法》制定《中华人民共和国个人所得税法实施条例》；又如本文研究的《中华人民共和国政府信息公开条例》（国令492号）就暂无对应的法律，属于先实施条例，待条例成熟后再决定是否转化为法律。

国发和国函的区别在于，国发一般是针对全国，而国函是针对行业或者地域，不具备全国性。如《国务院关于支持自由贸易试验区深化改革创新若干措施的通知》（国发〔2018〕38号），而同样是自贸区公文文件，针对地方的公文文件《国务院关于同意设立中国（海南）自由贸易试验区的批复》，则是属于国函〔2018〕119号。

一份国令往往对应多份国务院文件或者国务院办公厅文件，如《中华人民共和国政府信息公开条例》（国令492号），其相关的国办文件有《国务院办公厅关于推进社会公益事业建设领域政府信息公开的意见》（国办发〔2018〕10号）、《国务院办公厅关于推进公共资源配置领域政府信息公开的意见》（国办发〔2017〕97号）和《国务院办公厅关于推进重大建设项目批准和实施领域政府信息公开的意见》（国办发〔2017〕94号），它们分别是政府信息公开条例在社会公益、公共资源配置和重大建设项目领域的细则。

明电属于一般不具备保密属性的政府公文，如《国务院办公厅关于2019年部分节假日安排的通知》（国办发明电〔2018〕15号）。对于国务院办公厅的文件的规则也类似于此。

### 3.2 数据采集

本文研究发现，政府公文的唯一标识符是政府文号，考虑到历史公文数据的正确性，本文对于实验数据的准备主要是基于政府信息公开目录纸质文档。这些文档可以从政府信息公开的专栏中下载获取，本文具体的数据清洗流程如下。

（1）下载政府信息公开目录纸质文档，采用Apache PDFBox抽取政府文号，并校对政府文号。文号校对主要是针对历史公文，如国发〔1969〕50号文件，其纸质目录的文号为（69）国发文50号。本文利用Java模板技术进行字符串替换，共计得到1969—2018年的4 833条公文元数据。

（2）根据公文号下载对应的公文，采用Selenium引擎Chrome版进行下载。本文研究发现，并非所有的公文都保存为文本或者网页形式，还有一些公文是以图片数据形式保存。作者对下载的公文采用Apache PDFBox抽取全文数据，发现全图片公文数据只有190条，占比仅为3.9%。如《国务院办公厅转发国家计委、交通部关于加强港口建设宏观管理意见的通知》（国办发〔1995〕56号）就是全图片公文。

（3）本文将采集的数据存入MySQL数据库，方便下一步进行标记实验。MySQL中元数据主要采用直接

文本存储，而对于公文的全文则通过HtmlParser的Java库去掉HTML标签，存储到MySQL数据库。

## 3.3 数据标引流程原理

本文标引流程设计的核心是先分类后标引，如对行业、地域等进行分类。标引的原理是根据词性建立范式，并基于固定表达（如“军民融合”）进行范式的过滤。因此，该方法标引的同时，程序会生成中间临时文件。这些中间临时文件就是根据范式的规律和政府规划用语的固定表达而生成的词库。本文的范式是基于词性创建的，固定表达来自规划文件中高频词（如“一带一路”“军民融合”）的统计，且不对这些固定表达进行修改。本文与TF-IDF等其他方法的一个明显区别，就在于它在标引过程中产生了新的分类体系，以及基于文件本身生成的词库文件。这也是本文数据标引流程设计的创新之处。

## 3.4 数据标引定义过程

本文研究是通过文本标引来进行自动分类，自动标引生成标签（Tag）。

本文进行自动标引的主要原理是建立标题、公文号和自动标引关键词集的三元关系，见公式（1）。

$$Result(pcode) = F(title, keywords, state, method) \quad (1)$$

其中，Result代表是否成功标记；pcode代表公文号；title代表标题；keywords代表自动标引生成关键词集合，每个关键词之间用“/”符号分割，keywords初始值为NULL；state代表自动标引后状态，未参与标引初始state值为0，已经参与自动标引但是标引失败state值为1，成功标引state值为2；method代表生成标引的分类器方法，初始方法method值为NULL。

对于《国务院关于苏州市城市总体规划的批复》（国函〔1986〕81号）标记为：Result（国函〔1986〕81号）= F（国务院关于苏州市城市总体规划的批复，NULL，0，NULL）。

对于《国务院办公厅关于印发“十二五”全国城镇生活垃圾无害化处理设施建设规划的通知》（国办发〔2012〕23号）则标记为：Result（国办发〔2012〕23号）= F（国务院办公厅关于印发“十二五”全国城镇生活垃圾无害化处理设施建设规划的通知，NULL，0，NULL）。

## 3.5 基于地域关键词的标引

本研究主要基于自然语言处理技术（natural language processing，NLP），采用的NLP引擎为复旦大学开发的FudanNLP。本文实验主要分为7个步骤。

（1）加载数据源。对全集按照公式（1）进行标记，共有4 833条，计数方法为Count，见公式（2）。

$$Count_{全集}(Result_{全集}(pcode)) = 4\ 833 \quad (2)$$

（2）初始化数据。对全集中每个标题进行筛选，并且进行初始化标记。如对于《国务院关于苏州市城市总体规划的批复》（国函〔2016〕134号），按照公式（1）标记记录为公式（3）。

$$Result(国函〔2016〕134号) = F(国务院关于苏州市城市总体规划的批复，NULL，0，NULL) \quad (3)$$

（3）遍历数据分词。对公式（3）的字符串采用FudanNLP进行分词，分词的字符串为“国务院关于苏州市城市总体规划的批复”，分词的结果为：

国务院/名词 关于/介词 苏州市/地名 城市/名词 总体/名词 规划/名词 的/结构助词 批复/量词

从结果中可以看到，如果关系的是地域，则需要提取苏州市。由此公式（3）就变换为公式（4）。

$$Result(国函〔2016〕134号) = F(国务院关于苏州市城市总体规划的批复，苏州市，1，地域) \quad (4)$$

（4）数据标引。对于分词的数据需要进行标引，标引的目标是把词变成统一的词组。则可以对不同的词组从多个不同维度添加标签。对于公式（4）中，苏州市是地域名词，地域名词之后的名词可能是对该地域的说明，如“城市”。而名词之后出现的词语如果是“总体”，则属于修辞名词。“总体”之后“规划”属于类型名词关键词。因此需要进行检查地域名词之后的名词，使其满足范式。本文通过词频统计，发现满足的范式为：［地名］［地名补充关键词］［修辞名词］［类型名词关键词］。

类型名词代表该公文的主要作用，如“规划”，则公式（4）的状态改为公式（5）。

$$Result(国函〔2016〕134号) = F(国务院关于苏州市城市总体规划的批复，苏州市/城市/总体/规划，1，地域) \quad (5)$$

根据词频统计结果，地名补充关键词结合LOCATION集合主要包含以下名词，见公式（6）。

$$集合_{LOCATION}=\{土地，城市，区，乡，经济，群\} \quad (6)$$

（5）标引压缩。考虑到关键词搜索时要进行歧义消除，需要对公式（5）进行压缩，即删除“［修辞名词］”，同时对名词的并列语进行处理。如对《国务院关于辽河、松花江流域综合规划的批复》（国函〔1994〕82）进行提取，关键词集合见公式（7）。

$$集合_{KEYWORD}=\{辽河，松花江，流域，规划\} \quad (7)$$

（6）标引存储。对于公式（5），进行标引压缩后，其最终结果如公式（8）。

$$Result（国函〔2016〕134号）=F（国务院关于苏州市城市总体规划的批复，苏州市/城市/规划，1，地域） \quad (8)$$

Result（国函〔2016〕134号）由于标引成功，其结果Result（国函〔2016〕134号）值为真。

（7）更新计数器。完成“国函〔2016〕134号”后，更新公式（2）中数量，计数器减1。基于地域方法的遍历，本文研究发现共有316条符合地域关键词范式，即对于符合地域关键词的范式标记为：Count地域关键词（Result地域关键词（pcode））。

实验结果在MySQL数据中如图1所示。

| document_report_id | document_meta_id | id | title | report_type | report_keyword |
|---|---|---|---|---|---|
| 244 | 244 | 000014349/1985-00019 | 国务院关于规划和开发辽宁省兴城旅游区的批复 | 1 | 辽宁省/兴城/旅游区/规划 |
| 270 | 270 | 000014349/1985-00021 | 国务院关于长春市城市总体规划的批复 | 1 | 长春市/城市/规划 |
| 271 | 271 | 000014349/1985-00022 | 国务院关于大连市城市总体规划的批复 | 1 | 大连市/城市/规划 |
| 272 | 272 | 000014349/1985-00011 | 国务院关于南昌市城市总体规划的批复 | 1 | 南昌市/城市/规划 |
| 273 | 273 | 000014349/1985-00025 | 国务院关于乌鲁木齐市城市总体规划的批复 | 1 | 乌鲁木齐市/城市/规划 |
| 274 | 274 | 000014349/1985-00026 | 国务院关于桂林市城市总体规划的批复 | 1 | 桂林市/城市/规划 |
| 439 | 439 | 000014349/1986-00175 | 国务院关于苏州市城市总体规划的批复 | 1 | 苏州市/城市/规划 |
| 442 | 442 | 000014349/1986-00030 | 国务院关于天津市城市总体规划方案的批复 | 1 | 天津市/城市/规划 |
| 446 | 446 | 000014349/1986-00035 | 国务院关于上海市城市总体规划方案的批复 | 1 | 上海市/城市/规划 |
| 447 | 447 | 000014349/1986-00019 | 国务院关于宁波市城市总体规划的批复 | 1 | 宁波市/城市/规划 |
| 448 | 448 | 000014349/1986-00018 | 国务院关于贵阳市城市总体规划的批复 | 1 | 贵阳市/城市/规划 |
| 452 | 452 | 000014349/1986-00020 | 国务院关于哈尔滨市城市总体规划的批复 | 1 | 哈尔滨市/城市/规划 |
| 747 | 747 | 000014349/1988-00073 | 国务院关于唐山市城市总体规划的批复 | 1 | 唐山市/城市/规划 |
| 931 | 931 | 000014349/1990-00025 | 国务院关于乌江干流沿岸地区国土规划综合报告的批复 | 1 | 乌江/干流/规划 |
| 980 | 980 | 000014349/1990-00038 | 国务院关于海口市城市总体规划的批复 | 1 | 海口市/城市/规划 |
| 1504 | 1504 | 000014349/1994-00160 | 国务院关于辽河、松花江流域综合规划的批复 | 1 | 辽河/松花江/流域/规划 |
| 1623 | 1623 | 000014349/1995-00028 | 国务院关于南京市城市总体规划的批复 | 1 | 南京市/城市/规划 |
| 1630 | 1630 | 000014349/1995-00042 | 国务院关于指定深圳等市的城市总体规划由国务院审批的通知 | 1 | 深圳/城市/规划 |
| 1959 | 1959 | 000014349/1999-00015 | 国务院办公厅关于批准新乡市城市总体规划的通知 | 1 | 新乡市/城市/规划 |
| 1960 | 1960 | 000014349/1999-00016 | 国务院办公厅关于批准开封市城市总体规划的通知 | 1 | 开封市/城市/规划 |
| 1962 | 1962 | 000014349/1999-00020 | 国务院办公厅关于批准淮南市城市总体规划的通知 | 1 | 淮南市/城市/规划 |
| 1963 | 1963 | 000014349/1999-00034 | 国务院关于西安市城市总体规划的批复 | 1 | 西安市/城市/规划 |

图1 基于地域关键词的数据标引结果

## 3.6 基于行业关键词的标引

除地域关键词外，还有一类属于行业，如《国务院关于印发新一代人工智能发展规划的通知》（国发〔2017〕35号）。这类规划中不包含地域名词，与上述处理方法的差别主要在于数据标引和标引压缩两步。

从词频的统计结果来看，这类规划多为五年规划，如《国务院办公厅关于“七五”期间以煤代油计划安排问题的通知》（国办发〔1985〕76号）。范式如：［规划关键词］…［行业名词或者物品名词］。

其中，规划关键词包含“五年规划”中的关键词“五”，采用类似的程序处理流程，文本发现满足规律的共计114条，记为公式（9）。

$$Count_{行业关键词}（Result_{行业关键词}（pcode））=114 \quad (9)$$

实验结果在MySQL数据中如图2所示。

基于以上研究发现，不论是基于地域关键词还是基于行业关键词，对于标引后的标题数据在搜索时都可以更加精确；但考虑到政策文件的特殊性，不适合直接基于NLP的方法标引，而需要根据标题结合词频统计的原理分析得出不同的范式，并生成不同的模板进行处理。加之，国家的五年计划都具备连续性，如果用同一类数据生成同样的数据标引，则能更高效地搜索该规划从“十一五”“十二五”“十三五”发展的历史沿革。具体来说，以基于地域关键词的数据进行查询，可以看到长春市、大连市、乌鲁木齐市的城市规划在国务院文件中至少提过3次，进一步以关键词“长春市/城市/规划”为例查询数据，结果如图3所示。

“长春市/城市/规划”在国函〔1985〕63号、国函〔2011〕166号和国函〔2017〕87号分别被提到，从这3份文件中

| document_report_id | document_meta_id | id | cate | title | puborg |
|---|---|---|---|---|---|
| 185 | 185 | 000014349/1985-00014 | 国土资源、能源煤炭 | 国务院办公厅关于“七五”期间以煤代油计划安排问题的通知 | 国务院办公厅 |
| 275 | 275 | 000014349/1985-00001 | 科技、教育科技 | 国务院批转国家计委关于“七五”行业技术政策和技术改造问题报告的通知 | 国务院 |
| 379 | 379 | 000014349/1986-00069 | 农业、林业、水利水利 | 国务院批转水利电力部关于“七五”期间治准问题报告的通知 | 国务院 |
| 473 | 473 | 000014349/1986-00050 | 公安、安全、司法司法 | 国务院办公厅转发国务院法制局《关于〈一九八六年立法计划〉和拟订“七五”期间立法计划问题的请示》的通知 | 国务院办公厅 |
| 499 | 499 | 000014349/1987-00021 | 综合政务其他 | 国务院办公厅关于印发国务院“七五”期间立法规划的通知 | 国务院办公厅 |
| 1065 | 1065 | 000014349/1991-00048 | 工业、交通机械制造与重工业 | 国务院批转国务院机电产品出口办公室关于“八五”期间进一步扩大机电产品出口意见的通知 | 国务院 |
| 1587 | 1587 | 000014349/1995-00083 | 国民经济管理、国有资产监管经济体制改革 | 国务院关于“九五”期间上海浦东新区开发开放有关政策的通知 | 国务院 |
| 1607 | 1607 | 000014349/1995-00048 | 农业、林业、水利农业、畜牧业、渔业 | 国务院关于各省、自治区、直辖市“九五”期间年森林采伐限额问题的批复 | 国务院 |
| 1701 | 1701 | 000014349/1996-00047 | 城乡建设、环境保护环境监测、保护与治理 | 国务院关于国家环境保护“九五”计划和2010年远景目标的批复 | 国务院 |
| 1722 | 1722 | 000014349/1996-00032 | 国防国防建设 | 国务院办公厅、中央军委办公厅关于“九五”期间军队三线企业战略转移工作有关问题的通知 | 国务院办公厅、中央军委办公厅 |
| 1756 | 1756 | 000014349/1997-00060 | 商贸、海关、旅游对外经贸合作 | 国务院批转国家经贸委等部门关于“九五”期间进一步扩大机电产品出口意见的通知 | 国务院 |
| 1776 | 1776 | 000014349/1997-00003 | 卫生、体育卫生 | 国务院关于印发全国综合治理血吸虫病“九五”计划的通知 | 国务院 |
| 1779 | 1779 | 000014349/1997-00002 | 人口与计划生育、妇女儿童工作人口与计划生育 | 国务院关于下达“九五”计划分地区人口指标的通知 | 国务院 |
| 1788 | 1788 | 000014349/1997-00021 | 民族、宗教民族事务 | 国务院关于“九五”期间民族贸易和民族用品生产有关问题的批复 | 国务院 |
| 1898 | 1898 | 000014349/1998-00076 | 城乡建设、环境保护环境监测、保护与治理 | 国务院关于太湖水污染防治“九五”计划及2010年规划的批复 | 国务院 |
| 1957 | 1957 | 000014349/1999-00014 | 城乡建设、环境保护环境监测、保护与治理 | 国务院办公厅关于批准辽河流域水污染防治“九五”计划及2010年规划的通知 | 国务院办公厅 |
| 2061 | 2061 | 000014349/2001-00008 | 农业、林业、水利林业 | 国务院批转国家林业局关于各省、自治区、直辖市“十五”期间年森林采伐限额审核意见报告的通知 | 国务院 |
| 2086 | 2086 | 000014349/2001-00092 | 商贸、海关、旅游对外经贸合作 | 国务院办公厅转发外经贸部等部门关于“十五”期间进一步促进机电产品出口意见的通知 | 国务院办公厅 |
| 2089 | 2089 | 000014349/2001-00102 | 商贸、海关、旅游其他 | 国务院办公厅转发国家计委关于“十五”期间加快发展服务业若干政策措施意见的通知 | 国务院办公厅 |
| 2109 | 2109 | 000014349/2001-00098 | 科技、教育教育 | 国务院办公厅转发教育部等部门关于“十五”期间进一步推进特殊教育改革和发展意见的通知 | 国务院办公厅 |
| 2115 | 2115 | 000014349/2001-00087 | 卫生、体育卫生 | 国务院办公厅关于转发卫生部国家计委全国综合治理血吸虫病“十五”计划的通知 | 国务院办公厅 |
| 2129 | 2129 | 000014349/2001-00042 | 公安、安全、司法公安 | 国务院批转公安部关于“十五”期间消防工作发展指导意见的通知 | 国务院 |

图2 基于行业关键词的数据标引结果

| document_meta_id | id | title | report_type | report_keyword |
|---|---|---|---|---|
| 270 | 000014349/1985-00021 | 国务院关于长春市城市总体规划的批复 | 1 | 长春市/城市/规划 |
| 3457 | 000014349/2011-00130 | 国务院关于长春市城市总体规划的批复 | 1 | 长春市/城市/规划 |
| 4577 | 000014349/2017-00128 | 国务院关于长春市城市总体规划的批复 | 1 | 长春市/城市/规划 |

图3 长春市城市规划次数

也可以看出长春市规划的历史沿革和改革的延续性。

# 4 对比实验与分析

## 4.1 对比实验的设计

本文的研究思路是基于词频统计结果，通过不断地分类和提取范式，是一种半自动化半人工、交互式的分析方法。如果利用已经分析完成的词库和范式库，可以对其他相同领域的公文数据进行标引。

除了本文的标引方法，常见的其他标引方法还有TF-IDF（Term Frequency-Inverse Document Frequency）模型和主题模型。这几类模型的发展顺序是首先是TF-IDF模型，在TF-IDF模型的基础上发展了LSA（Latent Semantic Analysis）模型，在LSA的技术上发展了pLSA（probabilitistic Latent Semantic Analysis）模型，最后在pLSA模型的基础上发展了LDA模型。

TF-IDF模型通过对语料库进行分词，进而建立关键词库。其中词频（TF）指的是关键词在语料文件词中的频率，而逆文档频率（IDF）指关键词在文件总数的频率，程序处理时需要先对逆文本频率取对数，然后计算TF和IDF的乘积（即为TF-IDF值），最后根据TF-IDF进行提取的关键词用来对该文档进行标引。TF-IDF模型是对一份文档的所有词频进行标记，导致词频和文档的构成矩阵维度巨大。相关学者在此基础上，用文档部分关键词代替全文档，这就是LSA模型的核心思想。

LSA模型是通过建立关键词和文档的矩阵，对矩阵进行奇异值分解（singular value decomposition，SVD）降维，用部分关键词代替整个文档的属性，而LSA技术的主要作用就降维。LSA的局限性在于，对于矩阵进行SVD分解需要保证各个主题关键词是互相垂直的向量，如果一个关键词存在多个含义，则会导致无法区分而造成错误，这在中文的政府公文分析中尤为严重，因为中文的词义比英文复杂。

在LSA的基础上，相关学者提出了pLSA模型。与LSA的区别就是pLSA增加了文档在语料库中的概率。

在pLSA模型基础上发展的是LDA模型，LDA模型在pLSA的基础上增加了主题的Dirichlet先验分布后得到的贝叶斯模型，pLSA只能对语料库中文本进行语义识别，而无法识别不在语料库的文本语义，但是LDA模型可以。

本文测试TF-IDF模型和LDA模型采用的语料库为上文中基于地域关键词的数据的公文标题数据，测试数据为316条，本文对比研究的程序基于Python的NLP框架gensim进行编码，具体测试代码分为模型训练代码和模型测试代码。

## 4.2 对比实验的结果

如图4~图7所示，对比实验结果（均取前10）。

通过图4和图5对比发现，默认gensim并没有将长春作为一个单词给分割，因此TF-IDF模型的实验效果和LDA模型的实现效果都不好，都不如本文研究方法。

通过对比图6和图7发现，默认gensim虽没有将长

春作为一个单词给分割，但是将规划作为一个词语，并且LDA模型会优先考虑包含“规划”的词组，如“总体规划”，而TF-IDF则更考虑完全匹配的词。

| 序号 | TF-IDF权重分值 | 公文标题 |
|---|---|---|
| 1 | 0 | 长春市/城市/总体规划/ |
| 2 | 0 | 大连市/城市/总体规划/ |
| 3 | 0 | 南昌市/城市/总体规划/ |
| 4 | 0 | 乌鲁木齐市/城市/总体规划/ |
| 5 | 0 | 桂林市/城市/总体规划/ |
| 6 | 0 | 苏州市/城市/总体规划/ |
| 7 | 0 | 天津市/城市/总体规划/方案/ |
| 8 | 0 | 上海市/城市/总体规划/方案/ |
| 9 | 0 | 宁波市/城市/总体规划/ |
| 10 | 0 | 贵阳市/城市/总体规划/ |

图4 关键词为“长春”（TF-IDF模型结果）

| 序号 | LDA权重分值 | 公文标题 |
|---|---|---|
| 225 | 0.682325 | 石家庄市/城市/总体规划/ |
| 160 | 0.680125 | 太原市/土地利用/总体规划/ |
| 50 | 0.549091 | 同意/调整/江苏/苏州高新区/出口/加工区/规划/范围/复函/ |
| 313 | 0.532995 | 兰州/—/西宁/城市群/发展/规划/ |
| 38 | 0.531895 | 北京/城市/总体规划/ |
| 243 | 0.529361 | 呼和浩特市/城市/总体规划/ |
| 251 | 0.529352 | 淄博市/城市/总体规划/ |
| 260 | 0.529351 | 济南市/城市/总体规划/ |
| 105 | 0.52935 | 武汉市/城市/总体规划/ |
| 257 | 0.52935 | 合肥市/城市/总体规划/ |

图5 关键词为“长春”（LDA模型结果）

| 序号 | TF-IDF权重分值 | 公文标题 |
|---|---|---|
| 57 | 0.376669 | 松花江流域/防洪/规划/ |
| 55 | 0.349665 | 辽河流域/防洪/规划/ |
| 58 | 0.349665 | 太湖流域/防洪/规划/ |
| 59 | 0.349665 | 长江流域/防洪/规划/ |
| 60 | 0.349665 | 黄河流域/防洪/规划/ |
| 79 | 0.349665 | 淮河流域/防洪/规划/ |
| 246 | 0.313801 | 长江三角洲/城市群/发展/规划/ |
| 248 | 0.313801 | 中原/城市群/发展/规划/ |
| 270 | 0.313801 | 北部湾/城市群/发展/规划/ |
| 56 | 0.275538 | 海河/流域/防洪/规划/ |

图6 关键词为“长春/规划”（TF-IDF模型结果）

| 序号 | LDA权重分值 | 公文标题 |
|---|---|---|
| 243 | 0.999544 | 呼和浩特市/城市/总体规划/ |
| 290 | 0.999544 | 包头市/城市/总体规划/ |
| 66 | 0.999527 | 贵州省/土地利用/总体规划/ |
| 78 | 0.999527 | 四川省/土地利用/总体规划/ |
| 21 | 0.999478 | 西安市/城市/总体规划/ |
| 29 | 0.999478 | 汕头市/城市/总体规划/ |
| 61 | 0.999478 | 西安市/城市/总体规划/ |
| 280 | 0.999478 | 汕头市/城市/总体规划/ |
| 94 | 0.999421 | 武汉市/土地利用/总体规划/ |
| 128 | 0.999421 | 呼和浩特市/土地利用/总体规划/ |

图7 关键词为“长春/规划”（LDA模型结果）

## 4.3 现象的原因分析

根据4.1的实验和4.2的数据显示，LDA的效果不如TF-IDF，因此，本文主要分析TF-IDF方法和本文方法存在的区别及原因。

在实验中，TF-IDF方法对搜索“长春”的结果错误，原因是在进行gensim分词时，采用Python的jieba分词引擎。该分词引擎是基于中文通用词库，在分析本文时，分词的词语是“长春市”，而不包含“长春”，由此导致结果错误。

本文的数据标引是基于NLP的模板范式，先考虑词性，即基于词性自建词库，然后建立基于词性的句式模板，并用该模板来匹配关键词，而不采用通用词库。

尽管TF-IDF的优势在于自动化处理，但是TF-IDF加载通用分词引擎，需要增加停用词（stopwords），而停用词只能删除，不能合并。在本文研究中，jieba分词引擎将“军民融合”分词为“军民/融合”。对于“军民融合”属于规划中的固定表达，TF-IDF采用通用词库分词会导致结果错误，而本文的优势就在于基于词法、句法的规律提取词性范式，并且是先分类后标引。如“建立人工智能特色小镇”属于建设类而不属于技术类规划，“人工智能芯片”则属于技术类规划。因此，通过本文的研究方法，不会出现词语搜索错误的情况，即不会出现图4和图6的情况。这也是本文相较于TF-IDF的主要优势。

本文提取词时会根据FudanNLP引擎考虑词义，如根据地理位置和行业进行区分。另外，在未来的研究中，本文会进一步考虑词语的上位词关系、同义词关系，这是TF-IDF无法处理的情况。

然而，本文方法的局限在于，需要半人工的提取范式和行业词库。

## 5 研究结论

本文提出了一种综合词典标引和语义标引的方法来分析政府公开信息数据，使用复旦大学的FudanNLP开源包对公文标题进行分词、词性标注等处理，再通过句法和语义分析方法提炼范式，最后寻找满足范式的关键词作为标引。

同时，将本文的标引方法与TF-IDF模型和LDA模型进行了对比研究。本文标引的准确度比TF-IDF模型和LDA模型高，主要在于该标引方法不会出现搜索不到的情况，但是本文标引的自动化程度不如TF-IFD模型和LDA模型。本文需要结合行业知识构建词库和范式。

由于本文的范式是基于词性和行业特点，本文的词库会根据词性抽取具体的固定表达；同时，本文先分类后标引，标引基于范式提炼词库，因此，效果会比常见TF-IDF好，不会出现明显错误和搜索不到的问题。

另外，本文的主要工作为将自然语言处理方法应用于对政府公开信息数据的分析，而对相关公文进行标引便于民众在新媒体和搜索引擎中搜索政府公开信息，

因此对公文进行自动化标引具有重要意义。

同时通过本方法进行自动化标引的关键词库，可以作为其他相关研究者底层技术的词库。

## 参考文献


[1] 中华人民共和国政府信息公开条例 [J]. 中华人民共和国国务院公报，2007（15）：15-18.

[2] 李盛. 《中华人民共和国政府信息公开条例》的制定背景、主要内容及目录编制 [J]. 电子政务，2008（5）：21-26.

[3] 贾君枝，闫晓美，武晓宇. 政府信息公开的自动标引的设计与实现 [J]. 情报理论与实践，2012，35（2）：109-113.

[4] 张新民，罗卫东. 我国政府信息公开工作中的技术问题探析 [J]. 图书情报工作，2008（8）：58-61.

[5] 吴洁明，赵文丽. 新闻出版行业标准碎片化标引的研究与实现 [J]. 计算机工程与设计，2017，38（8）：2281-2286.

[6] 邓雪琳. 改革开放以来中国政府职能转变的测量——基于国务院政府工作报告（1978—2015）的文本分析 [J]. 中国行政管理，2015（8）：30-36.

[7] 朱晓峰，崔露方，陆敬筠. 国内外政府信息公开研究的脉络、流派与趋势——基于WOS与CNKI期刊论文的计量与可视化 [J]. 现代情报，2016，36（10）：141-148.

[8] 王志刚. 政府理念和经济增长：基于文本挖掘 [J]. 经济社会体制比较，2016（6）：5-6.

[9] 潘松. 国务院公报中成语的运用 [J]. 宿州学院学报，2011，26（1）：54-57.

[10] 程大荣. 从《国务院公报》看“意见”处理的规范化 [J]. 档案学通讯，2015（1）：35-38.


## 作者简介


江华丽，女，1980年生，博士，研究方向：企业项目管理，E-mail：daisy_jiang@whu.edu.cn。

曹祺，男，1988年生，博士，研究方向：信息计量学。

陈刚，男，1970年生，教授，博士生导师，研究方向：Web搜索与挖掘。


### Design and Implementation of Automatic Indexing of Government Public Information


JIANG HuaLi[1] CAO Qi[2] CHEN Gang[1]

( 1. School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China; 2. Greysh Group Co., Ltd., Beijing 100080, China )



Abstract: At present, government information disclosure is mainly based on the “Regulations on the Openness of Government Information of the People’s Republic of China”, but from the perspective of users, it needs to be adapted according to different usage scenarios. Therefore, it is of great significance to automate indexing of official documents. Based on natural language processing technology, this paper refines the paradigm in the official document title through the experiment and analysis of word frequency, part of speech and word meaning, and automatically indexes 4 388 official documents of the State Council from 1969 to 2018. In the case of regional keywords and industry keywords as an example, the relevant keywords can be searched and secondary processed after indexing. This article mainly deals with the indexing of the title, and the full text has not been indexed.



Keywords: Government Information Disclosure; Text Mining; Automated Indexing

（收稿日期：2018-12-26）