科技成果转化政策文本中的领域关键词汇提取研究

张 越 刘琦岩 张玄玄 望俊成
（中国科学技术信息研究所，北京 100038）

摘要：为了解决政策文本研究领域尚未建立其关键词表的问题，尝试利用语法分析特征，挖掘核心词汇构成模式，构建政策文本核心词汇的抽取模型，采用基于混合指标的政策领域关键词汇抽取和结果筛选方法对关键词进行识别，最后在科技成果转化政策上对该方法进行实证研究。结果表明，该方法可以从政策文本中发掘出潜在的信息，为政策内容分析和决策支持提供数据基础与效率支撑。

关键词：科技成果转化；政策文本；混合指标；领域关键词汇；信息抽取
中图分类号：F204 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2018.03.010

Methods of Domain Keywords Extraction in the Policy of Transformation of Scientific and Technological Achievements Based on Hybrid Index

ZHANG Yue, LIU Qiyan, ZHANG Xuanxuan, WANG Juncheng
(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: The keywords table in the policy domain is not yet established. We try to solve this problem by analyzing the characteristics of grammar, mining structure pattern of keywords, constructing the extraction model of the core vocabulary of the policy text, using the methods of keywords extraction and results screening in the policy domain based on hybrid index. Finally, we make an empirical study in the policy of transformation of scientific and technological achievements. The results show that the method can discover potential information from the policy text and provide data basis and efficiency brace for policy content analysis and decision support.

Keywords: transformation of scientific and technological achievements, policy texts, hybrid indicators, domain keywords, information extraction

0 引言

提取文本中关键词的价值主要在于可以简洁地表示文档的基本内容、改进文件的显示（例如，突出显示重要的短语）、帮助分类文件和为自然语言处理提供外部词典支持等，其任务是自动识别最能描述文档主题的术语。虽然术语不同，但功能是一样的，都有在文档中讨论话题的特征。目前，学界对关键词抽取算法可根据是否需要人工标注的语料进行训练分为有监督的提取

作者简介：张越（1994—），男，中国科学技术信息研究所情报学专业研究生，研究方向：科技政策与科技战略；**刘琦岩**（1964—），男，中国科学技术信息研究所研究员，研究方向：科技发展战略与政策研究；**张玄玄**（1995—），女，中国科学技术信息研究所情报学专业研究生，研究方向：科技政策与科技战略；**望俊成**（1984—），男，中国科学技术信息研究所副研究员，研究方向：科技政策与科技管理，数据可视化、情报学基础理论（通讯作者）。

基金项目：中国科学技术信息研究所创新基金“政策亮点挖掘模型的构建——以科技成果转化政策文本为例”（QN2017-01）。
收稿时间：2017 年 12 月 15 日。

和无监督的提取。有监督的提取算法需要人工标注的语料进行训练，人工预处理的代价较高。而无监督的抽取算法直接利用需要提取关键词的文本即可进行关键词的提取，其适用性较强。

无监督的方法主要采用统计量的方法完成关键词提取任务，解决了文本挖掘和自然语言处理中的重要问题。在这之前，通常需要对文本进行解析提高后续步骤的准确度。文本解析主要基于语法与规则，对一类文本的组织结构特征进行归纳，其缺陷在于由于语言的表达方式不尽相同，需要定义出一套通用的语言模型，如定义词串的词性序列[1]。

有监督的方法是对文本中的一部分数据进行训练以得到统计参数，建立相关模型后对词汇进行抽取。常用的方法有利用分类算法中的支持向量机（SVM）[2]、最大熵模型[3]以及将短语抽取问题视为序列标注问题并利用条件随机场（CRF）[4]进行抽取。基于统计量度的方法是从分类语料中分析领域词汇之间的用词规律，从而发现并获取领域关键词汇，其中有一些相对简单的统计量度包括首次出现位置[5]、词频和词的共现信息[6]等。一个词汇如果越符合领域关键词汇的特征，则越有可能成为候选关键词。这些特征即是抽取候选关键词的属性。特征选取的好坏将直接影响关键词抽取质量的高低。对用数学方法进行特征选择的算法，决定文本特征提取效果的主要因素是评估函数的质量。研究学者在过去的研究中常用评估函数包括：期望交叉熵[7]（Expected Cross Entropy）、TF−IDF[8]、信息增益方法[9]（Information Gain）。

为了避免单纯使用其中一种方法抽取关键词汇的局限性，有关学者使用无监督的方法提出并建立了规则和统计方法相结合的中文关键词汇自动抽取模型。杜波[10]设计了一个将统计方法与规则方法相结合的专业领域内关键词汇抽取算法。董洋溢[11]提出了一种基于文本特征和复合统计量的中文领域核心词汇抽取方法，在对中文文档中的词语进行粗粒度筛选后，再综合考虑候选关键词汇的词性、长度、边界词语等文本特征，构造出信息熵和TFIDF等统计量，计算其综合权值。

本文针对科技成果转化政策文本中的关键词汇抽取的相关技术进行研究，主要包括分析研究该领域政策关键词汇的组成特点，制定相应的标注规则。根据语言学特征和统计学特征两方面制定候选关键词汇组合，采用基于混合指标的政策领域关键词汇抽取和结果筛选方法，得到候选领域词汇构建自定义词表等。

1　领域关键词汇的定义与语言特征

关键词汇是在特定上下文中具有特定含义的词或短语。本文讨论的领域词汇是一种在特定领域具有特殊意义的技术关键词汇，即特定领域或行业中限定专业概念的典型词或短语，并被相应领域或行业的人熟悉和使用[12]。词作为中文语言中最小的语意单位，在自然语言处理（NLP）领域中进行文本挖掘和语义分析时难以发挥良好的效果，使用领域词汇组合能够满足文本研究和分析的实际需求。

要实现对领域词汇的抽取，就必须对其特征进行全面的认识，并针对其特征设计特定的抽取方法。从语言学结构角度看，领域词汇有如下特征：一是长度特征。领域词汇的长度主要为3～6个字，根据关键词汇的长度可以限定长度窗口对目标进行词长的筛选。二是词性特征。词组结构中落脚的中心词多为名词，少量兼有动词。三是词法构成特征。如名词（n.）+名词（n.）+名词（n.），形容词（a.）+名词（n.）+名词（n.），动词（v.）+动词（v.）等形式。根据关键词汇的构成模式特征，可以构造词法模板来过滤候选关键词汇。同时，领域词汇具有一定的统计特征：一是领域词汇是仅在一个学科领域中使用、表示该学科领域内概念、特征或关系的词语，且不局限于本领域，在后续分析时假设仅在该领域内出现；二是领域词汇可以只在一个学科领域中存在，也可并存于多个学科领域中。但是领域词汇在本领域的出现频率很高，且远远高于在其他领域出现频率。

领域关键词汇作为领域知识的一种描述形

式，可以独立承载一定的专业信息。首先必须具有一个成熟的自然语言表达模式，满足一定的语法结构[13]。引入词性过滤模板可以筛选出目标词组。将传统分词中的词长进行扩展，用粒度较大的词法模板进行设计是有必要的。大量的关键词汇构成都以名词词汇组合的短语为主要形式如“技术市场”，因此在大量的文献研究中，研究人员都采用名词短语构造候选集合。但是存在的问题：一是根据文本所处领域限制，其关键词汇应包括少部分存在的动词短语如“协议定价”，极少量的形容词、副词短语；二是中文词汇具有多义性，而且在表示不同词义时词形几乎没有变化，最常见的就是名词和动词的混淆，如“成果转化”标注为“成果/n 转化/v”，显然这里的“转化”表示“转化过程或行为”的意思，应该标注为名词。因此，在抽选关键词汇候选集时，并不是人为地设定某一个或多个固定的语法结构或组成模式。

本文分析了一个政策领域的文本库子集，其中包含 2001 条词长为 1 ～ 2 个中文字符的分词结果。使用中国科学院计算技术研究所的词法分析工具NLPIR对这些关键词汇进行分词及词性标注，人工划分出词长为 2 ～ 8 个中文字符的短语组合条目有 323 条。因此，将候选关键词汇的长度限制在 2 ～ 8 个中文字符。由于包含 2 ～ 4 个词的关键词汇结构紧凑而稳定，出现的词法模式种类少而且集中。从这些关键词汇条目中自动学习获得 18 种词法模式，这些模式对语料库中词长 4 ～ 6 个中文字符的关键词汇覆盖率达到 90% 以上。表 1 中列举了出现最多的 10 种词法模式。

词长介于 4 ～ 6 个中文字符的关键词汇结构相对比较松散，不能采用少量的词法构成模式来覆盖大多数词条。例如，在 4 词关键词汇中，出现频率最高的 5 种词法模式只能覆盖 83% 的关键词汇条目。可见，随着关键词汇长度的增加，词长大于等于 5 中文字符的关键词汇已经不适合采用词法模式作为识别规则。语料库中共包含 7 条词长大于等于 7 个中文字符的关键词汇，通过分析这些关键词汇首部、中部和尾部词的词类特征，总结出以下 4 条规则。

表 1　长度 2~3 中文字符的关键词汇中 top 5 的词法构成模式

词组长度∈(2,4]	词组长度∈(4,8]
n.+n.	n.+n.+n.+n.
v.+n.	n.+n.+n.
v.+v.	v.+n.+n.
n.+v.	n.+a.+n.
a.+n.	a.+n.+n.

规则一：领域关键词汇构成为“定语+中心词”形式，其中至少有一个词或中心词属于名词、动词、量词、形容词、缩略语或后接成分。

规则二：领域关键词汇候选项中不得包含如下性质的词语：代词、状态词、非语素词、处所词、拟声词、叹词、语气词等虚词、常用语。

规则三：领域关键词汇不得以连词、助词和后接成分作为词首，不得以连词、方位词和前接成分性质的词语结尾[14]。

规则四：领域关键词汇中由连词连接的复合动宾短语，其形式为“v.1+conj+v.2+n.”，如“开发和服务能力”，以及由连词连接的复合主谓短语，其形式为“n.+v.1+conj+v.2”，如“科技成果转让、许可和作价投资”。这几个v.表达的意思可能相似或递进，这时需要进行短语扩展，除划分为v.2+n. 或n. +v.1 仍需补充标注v.1 +n.或n.+v.2 后继步骤中外，还要在词语的粒度控制下使用这些词法模式作为语言规则对候选关键词汇进行抽取。

2　领域关键词词的提取方法

对于关键词提取，词频统计通常认为是最直接和常规的方法，但是词的重要性和词频没有完全对应关系。统计学指标的选择需要考虑词汇的多维特征，文中所指的混合指标考虑了多种关键词提取算法的特征和优势。在调研中发现，TF-IDF主要用以评价字/词在一个文档集合或语料库中单个文档的分布重要度，字/词的重要度会随着其在文档中的出现次数成正比增加，也会随着其在语料库出现的频率成反比下降；由规范化Google距离演变而产生的NPD主要用来衡量一个

词在语料库中与其他词的相似重要度；TextRank在迭代计算中考虑了所构建词图中，词在节点之间的关系重要度。

由于3个统计学指标从3个方面描述了重要度特征的属性，不分主次，因此在计算时将权重平均分配。指标混合的结果就是稀有的或者重要的词被给予了更高的分值，而对于常用却被认为比较不重要的词则在考虑权重时有较小的影响。

2.1　利用TF-IDF提取

要使用统计的方法判断关键词汇所包含的信息量，就要研究关键词汇在语料中的统计分布特征。TF−IDF（Term Frequency−Inverse Document Frequency）算法是一种基于统计特征的非常经典的算法，通过计算一个词TF值和IDF值的乘积作为该值的得分，然后根据得分从大到小进行排序，选择分数高的词语作为关键词。对于在某一特定文件里的词语t_i来说，其重要性可表示为：

$$TF_{i,j}=\frac{n_{i,j}}{\sum_k n_{i,j}}$$

在上式中，$n_{i,j}$是该词t_i在文档d_j中出现的次数，而分母则是在某个文档d_j中分词并过滤停止词后词语的数量。

逆向文件频率（inverse document frequency，IDF）则指词语在整个语料库中出现的频率大小。首先要指出的是TF−IDF算法是针对一个语料库（也就是多篇文档）进行关键词提取的算法。

$$IDF_i=\log\frac{|D|}{\left|\left\{j:t_i\in d_j\right\}\right|}$$

其中：$|D|$表示语料库中的文档总数，文中使用的北大法宝政策法规数据库中中央法规司法解释273698篇，地方法规规章1213430篇，共计1487128篇文件；$\left|\left\{j:t_i\in d_j\right\}\right|$表示出现词语$t_i$的文件数目（即$n_{i,j}\neq 0$的文件数目）。

然后，$TF-IDF_{i,j}=tf_{i,j}\times idf_i$

根据这个值对候选词从大到小排序，选择前n个作为候选关键词即可。

2.2　利用规范化政策距离提取

规范化Google距离[15]是对于给定一组关键词从Google搜索引擎返回的命中数中测量语义相似度的一个指标。在自然语言意义上具有相同或相似含义的关键词往往以“规范化Google距离”为单位“接近”，而具有不同含义的词往往距离更远。

从上文中获得衡量一个政策关键词汇集合中构成词组的语素x和y之间距离的灵感，一个关键词汇的形成过程中，语素之间的距离在逐渐减小，相似度增大，搜索引擎所包含的语料库也弥补了单一领域缺乏相对性指标验证关键词汇抽取科学性的不足。

定义规范化政策距离（NPD）为：

$$NPD(x,y)=\frac{\max\left\{\log f(y),\log f(y)\right\}-\log f(xy)}{\log N-\min\left\{\log f(y),\log f(y)\right\}}$$

其中：N表示北大法宝政策文本库搜索到的网页总数，共计1565375条记录，包括中央法规司法解释274310条，地方法律规章1215177条，法律动态70337条，立法背景资料5551条；$f(x)$，$f(y)$表示搜索词x和y的命中数量；$f(xy)$表示关键词汇xy的网页命中数量。

利用NPD计算领域词汇构成语素之间在北大法宝政策文本库的规范化政策距离。当NPD介于0～1时，能够在一定程度上说明构词的合理性，在搜索引擎上共同出现的概率较高。当$NPD\geqslant 1$时，说明规范化政策距离较远，不具备构成领域词汇的条件。

2.3　利用TextRank算法提取

TextRank算法可用于提取文本中的关键词，但需要将文本分为单个词的集合。而每个词类似于PageRank算法中的一个节点。以PageRank为例，假如一个大型网站有一个超链接指向某个小网站，那么小网站的重要性上升，而上升量则依据指向它的大网站的重要性而定。同理，假如两个词的距离小于预设的距离，那么就认为这两个词间存在语义关系，否则不存在。这个预设的距离在TextRank算法中被称为同现窗口（co−occurance window）。这便可构建一个词的图模型。

假设一个句子由下列词依次组成ω_1，ω_2，…，ω_n，存在一系列长度为k的窗口ω_1，ω_2，…，ω_k、ω_2，ω_3，…，ω_{k+1}、ω_3，ω_4，…，ω_{k+2}等。每个窗

口中任意两个词的对应节点间存在一个无项无权的边，以k个词的共现关系建立起节点之间的链接。基于这种构成原理，可以计算出每个节点的重要性，节点对应的重要的词可以作为关键词。如果原文中存在若干关键词相邻的情况，这些关键词即可组成一个关键词汇。这里给出TextRank的公式：

$$WS(V_i)=(1-d)+d\times\sum_{V_j\in In(V_i)}\frac{\omega_{ji}}{\sum_{V_k\in Out(V_j)}\omega_{jk}}WS$$

在式中，V_i表示某个网页；V_j表示链接到V_i的网页（即V_i的入链）；$WS(V_i)$表示网页V_i的PR值；$In(V_i)$表示指向网页V_i的所有入链节点的集合；$Out(V_j)$表示V_j指向的的所有出链节点的集合；d表示阻尼系数，常取0.85；ω_{ji}是权重项。

3　领域关键词汇的识别与候选库的建立

按照基于规则的语言规则，对文本数据集中符合句法模板的多元词组进行识别。为了保证数据的质量，早期采用专家人工标注的方式。采用3位专家交叉验证的方式，即将数据集合平均分成3个容量大致相同但不相交的子集，3位专家分别标注后取标注完成集合的公共集。

领域关键词汇识别的详细步骤是：（1）词汇定位。查找词汇在文本中的开始位置，位置信息加入关键词汇列表中，并标注其类型。（2）文本截取。以“词汇开始位置+词汇长度”为起点。截取剩余文本作为匹配文本。（3）词条筛选。遍历关键词汇列表，如果有词汇的起始位置处于更长的一个词汇中，则去除较短词汇在关键词汇列表中的信息，保留较长词汇的信息。

目标词汇提取结果要满足科学性。科学性指标的建立主要依靠德尔菲法。由专家对关键词汇候选集中关键词汇进行筛选，专家设定关键词汇科学性评价指标，根据每一位专家给出的权重意见，计算指标权重结果，最后对多位专家给出的权重结果取均值，尽可能消除个人主观因素对指标权重的影响。

首先，通过大量的文献阅读和调研进行领域词汇指标阈值的范围限定，后期放入语料库中进行自动或半自动地测试。通过计算根据是否在阈值范围内来确定词是否有意义。设定各指标按“十分制”评价，筛选前50%作为结果，即相关的标准为共现词频≥9来判断所选词的科学合理性。其次，相同的发文机构或相同类型的文章，其用词、句法、词法等都会有一些共性，因此同样效力的文献在选取时贯穿国家、省部、地级市3个层面，内容贯穿基于指导意见的宏观层面、基于行动方案的中观层面以及基于实施细则的微观层面，增加样本的广泛性与人工标注关键词汇的效率，同时通过Pkulaw API把政策文本搜索引擎作为词汇的样本来提高识别准确率。

4　实证研究

4.1　数据获取

中文科技成果转化政策领域词汇候选词抽取构建的主要思路为：网页数据抓取，对目标网页信息进行爬取，构建基于大量大样本的政策文献的完善的语料库并建立中文政策领域自定义候选词库。由于中文政策目前尚无主题词表作为基准词使用，即需要对特定领域特定政策文本的特定表达习惯进行语言偏好分析，并利用已有知识和专家意见率先进行人工切词和标注，并构建自定义词表的考量因素或指标依据，需要对初始词组进行筛选，将符合要求和标准的词加入到领域自定义候选词库中。

科技成果转化政策语料库选择《促进科技成果转化行动有关重点政策摘编》中时间范围为2015年10月至2016年12月的科技成果转化政策文献，共计77137字。利用构建的词法模板，在国家层面选取国务院文件，识别到177个候选词组；在省级层面选取上海市文件，识别到111个候选词组；在市级别层面识别到55个候选词组。经合并再去重后初步得到324个词组。候选词组见表2。

4.2　数据处理

按照第2节中所述方法，分别计算初始词汇的3个指标和对应值，见表3。

为了证明各指标描述数据的独立性，研究每个领域政策关键词汇的指标之间的线性相关关

表 2　部分抽取到的初始词汇

词汇组合
作价投资、转移机构、转化收益、转化工作、收益分配、服务机构、技术市场、科技型企业、科技成果转让、作价入股、创新创业、转化情况、技术交易、离岗创业、股权激励、产学研合作、产业化、技术转让、管理制度、中介服务、单位预算、完成单位、股权奖励、技术开发、拟交易价格、科技计划、知识产权运营、人事关系、约定奖励、报告制度、高新技术企业、考核评价、技术职务、资产处置、人才培养、知识产权信息、技术推广、转化应用、转型升级、科技型中小企业、专项资金、知识产权管理、服务能力、开放共享

表 3　统计指标计算结果（部分）

共现词频	候选词	TF−IDF 值	规范化政策距离值	TextRank 平均值
55	服务机构	0.1560	0.30	0.005540
40	作价投资	0.2978	0.55	0.001463
39	转移机构	0.2757	0.63	0.003904
36	技术成果	0.1549	0.40	0.008265
29	技术服务	0.0905	0.30	0.008179
26	转化收益	0.1869	0.45	0.006151
26	转化工作	0.1595	0.63	0.007139
26	科技创新	0.1391	0.46	0.006229
25	收益分配	0.1265	0.29	0.000847
22	服务机构	0.0629	0.3	0.005540
21	转化项目	0.1198	0.32	0.007852
20	技术市场	0.1077	0.58	0.006351
20	科技型企业	0.1036	0.4	0.008818
19	科技成果转让	0.1395	0.48	0.009028
19	作价入股	0.1091	0.17	0.000000
18	转化服务	0.1153	0.59	0.008279
17	创新创业	0.0742	0.3	0.004598
17	转化情况	0.1276	0.74	0.006508
17	科技服务	0.1114	0.71	0.006415
16	技术交易	0.091	0.55	0.006524
16	离岗创业	0.1129	0.38	0.001808
15	股权激励	0.0909	0.35	0.002688
15	产学研合作	0.0737	0.3	0.001548
15	研发机构	0.1063	0.69	0.004697
15	科技成果转化项目	0.0962	0.35	0.010539
14	产业化	0.0414	0	0.001388
14	技术转让	0.0686	0.44	0.006275
14	管理制度	0.0284	0.25	0.003315
14	中介服务	0.0509	0.31	0.002976
14	完成单位	0.0776	0.61	0.005676
14	科技项目	0.0623	0.30	0.005988
14	技术研发	0.1143	0.78	0.007336
13	股权奖励	0.0712	0.67	0.003277
13	技术开发	0.0925	0.53	0.006035
13	转化管理	0.1180	0.95	0.007339
12	拟交易价格	0.0459	0.36	0.000000
12	高校科技成果	0.1134	0.68	0.010648
11	科技计划	0.1014	0.48	0.004365
11	知识产权运营	0.049	0.41	0.002082
11	人事关系	0.0784	0.48	0.000000
10	约定奖励	0.0552	0.42	0.002721
10	报告制度	0.0965	0.82	0.001279
10	科技成果转化服务	0.0688	0.58	0.010824
9	高新技术企业	0.0299	0.32	0.000000

系。首先绘制矩阵散点图，然后通过Pearson简单相关系数分析变量之间线性相关性的强弱。绘制的矩阵散点图如图1所示。图1表明各指标之间不具备较强的相关关系。

规范化政策距离和TF-IDF的相关系数检验的概率p值近似为0。因此，当显著性水平为0.05时，应该拒绝相关系数检验的零假设，认为两者总体存在线性关系。但由于二者Pearson相关系数仅为0.267，不具备强相关性。规范化政策距离和TextRank值以及TF-IDF和TextRank值的概率p值都很大，因此，当显著性水平为0.05时，应该接受相关系数检验的零假设，认为两者总体不存在线性关系。具体的皮尔森相关系数分析结果见表4。

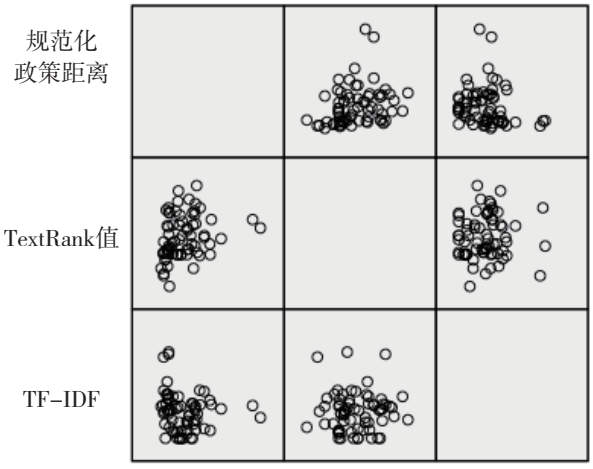

图1　指标相关关系散点矩阵

按照每一个指标的值大小进行打分，并对分数进行归一化。由专家设置最终得分≥总分一半的词作为候选领域词汇，见表5。

从表5可知，科技成果转化政策领域词汇组合中“收益分配”“作价入股”“转化收益”“作价投资”“股权激励”等词与成果转化收益方式和激励机制密切相关；“科技计划”“技术交易”“技术开发”等词涉及创新价值链中的基本活动；“科技型企业”“转移机构”“服务机构”“中介服务”“产学研合作”等主体名词词组揭示了技术市场中的创新中介与制造企业的合作关系。这些相对重要的领域词汇的发现可以发掘政策本身语言形态的构成以及政策话语本身隐含的系统性特征，如科技成果转化工作重心的转移等，从而进一步完善政策词表的构建和文本的语义解析任务。

表4　Pearson相关性

		规范化政策距离	TF-IDF	TextRank值
规范化政策距离	Pearson 相关性	1	0.267*	0.034
	显著性（双侧）	—	0.028	0.781
	N	68	68	68
TF-IDF	Pearson 相关性	0.267*	1	−0.169
	显著性（双侧）	0.028	—	0.169
	N	68	68	68
TextRank值	Pearson 相关性	0.034	−0.169	1
	显著性（双侧）	0.781	0.169	—
	N	68	68	68

注：*为在0.05水平（双侧）上显著相关。

表5　筛选所得候选领域词汇

科技成果转化政策领域词汇组合
收益分配、作价入股、转化收益、作价投资、创新创业、离岗创业、科技成果转让、科技型企业、转移机构、服务机构、产学研合作、转化工作、股权激励、技术市场、转化情况、中介服务、科技计划、技术交易、技术开发

5　结语

（1）本文针对科技转化政策文本的中文关键词汇及关键词汇抽取的相关方法进行了研究。结果表明，语义识别度高、符合用户文本分析需求的领域关键词汇，为创建自定义词表提供了依据，为政策文本进一步在自然语言处理中的使用和文本分析建立了基础，可以节约人工成本来提供决策判定支持。

（2）本文设计的一种快速识别并抽取领域关键词汇的方法，通过语言规则和多重指标混合的方法进行过滤和筛选得到一批具备参考意义的科技成果转化政策领域关键词汇，各指标之间相对独立，从多角度衡量词汇的重要性从而将政策文本中粒度适中的语义单位进行识别并抽取，所抽取的词基本涵盖了科技成果转化政策中所反复提及的政策落脚点。

（3）本研究的不足之处在于使用何种限定的规则、何种统计量以及二者结合方法将直接影响到抽取的结果。在混合指标权重的分配中，尝试了将 3 种指标权重平均分配的处理，希望未来会有更加科学的自动分配方法和对于抽取结果更加科学的评价方法。

参考文献

[1] 邹纲, 刘洋, 刘群, 等. 面向Internet的中文新词语检测[J]. 中文信息学报, 2004, 18(6): 1−9.DOI: 10.3969/j.issn.1003−0077.2004.06.001.

[2] ZHANG K，XU H，TANG J，et al. Keyword extraction using support vector machine[C]// Yu J X, Kitsuregawa M, Leong H V. Proceedings of the Seventh International Conference on Web−Age Information Management. Hong Kong: European Languages Resources Association, 2006:85−96. DOI:10.1007/11775300_8.

[3] 李素建, 王厚峰, 俞士汶, 等. 关键词自动标引的最大熵模型应用研究[J]. 计算机学报, 2004, 27(9): 1192−1197.DOI: 10.3321/j.issn: 0254−4164.2004.09.007.

[4] ZHANG C. Automatic keyword extraction from documents using conditional random fields[J]. Journal of Computational Information Systems, 2008, 4(3): 1169−1180.

[5] 张永刚, 梁颖红, 颜振祥, 等. 基于统计的中文关键短语自动抽取[J]. 江南大学学报(自然科学版), 2010, 9(1): 26−29.DOI: 10.3969/j.issn.1671−7147.2010.01.006.

[6] MATSUO Y, ISHIZUKA M. Keyword extraction from a single document using word co−occurrence statistical information[J]. International Journal on Artificial Intelligence Tools, 2004, 13(1): 157−169.DOI: 10.1142/S0218213004001466.

[7] 刘桃, 刘秉权, 徐志明, 等. 领域术语自动抽取及其在文本分类中的应用[J]. 电子学报, 2007, 35(2): 328−332. DOI: 10.3321/j.issn: 0372−2112.2007.02.031.

[8] SALTON G, BUCKLEY C. Term−weighting approaches in automatic text retrieval [J]. Information Processing & Management, 1988, 24(5): 513−523.DOI: 10.1016/0306−4573(88)90021−0.

[9] 任永功, 杨荣杰, 尹明飞, 等. 基于信息增益的文本特征选择方法[J]. 计算机学, 2012, 39(11): 127−130. DOI: 10.3969/j.issn.1002−137X.2012.11.029.

[10] 杜波, 田怀凤, 王立, 等. 基于多策略的专业领域术语抽取器的设计[J]. 计算机工程, 2005, 31(14): 159−160.DOI: 10.3969/j.issn.1000−3428.2005.14.057.

[11] 董洋溢, 李伟华, 于会. 文本特征和复合统计量的领域术语抽取方法[J]. 西北工业大学学报, 2017, 35(4): 729−735.DOI: 10.3969/j.issn.1000−2758.2017.04.026.

[12] PETERLICEAN A. Challenges and perspectives in teaching specialised languages[J]. The Journal of Linguistic & Intercultural Education,2015(8):149−162.

[13] BOURIGAULT D. Surface grammatical analysis for the extraction of terminological noun phrases [C]// Proceedings of COLING’92, 1992: 977−981.DOI: 10.3115/992383.992415.

[14] 周浪, 史树敏, 冯冲, 等. 基于多策略融合的中文关键词汇抽取方法[J]. 情报学报, 2010, 29(3): 460−467. DOI: 10.3772/j.issn.1000−0135.2010.03.011.

[15] CILIBRASI R L, VITANYI P M B. The google similarity distance[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(3): 370−383..DOI: 10.1109/tkde.2007.48.