科技政策术语自动识别技术初探

曾 文 李智杰 王小玉 董 诚
（中国科学技术信息研究所，北京 100038）

摘要：在对科技政策领域术语的特点分析基础上，提出一种适用于科技政策领域的术语识别方法，即结合科技政策术语的语言特点，采用统计计算的方法进行两次术语过滤过程，实现科技政策术语的自动识别。实验结果表明，本文提出的基于科技政策术语语言特点和统计计算相结合的科技政策术语自动识别的方法具有一定的可行性，将用于科技政策词典的构建和科技政策文本内容的深层次语义分析。

关键词：科技政策；科技政策术语；术语特点；统计计算；自动识别
中图分类号：TP391 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2017.03.004

Research on Automatic Recognition Technology of Science and Technology Policy Term

ZENG Wen, LI Zhijie, WANG Xiaoyu, DONG Cheng
(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: The paper proposed an automatic recognition method based on characteristics and statistical computing of term. The method fully combined language characteristics and statistical information of terms. And the method of statistical calculation is adopted to carry out the two filtering process of terms. Experimental results showed that the proposed method had certain feasibility. It will have certain application value. In the next step, the method will be used for constructing the dictionary and deep semantic analysis of science and technology policy.

Keywords: science and technology policy, science and technology policy term, term characteristic, statistical calculations, automatic recognition

1 引言

随着网络的发展和应用普及，国家和各级政府的科技政策通过网络进行实时发布，例如：科学技术部、中国科学院和各省市科技厅（委）均设有科技政策相关网站，并建有科技政策数据库，如全国科技创新政策数据库（http://www.kjcxzc.com/contentlist.asp?parentid=4）、万方数据的政策法规知识服务平台（http://s.wanfangdata.com.cn/Claw.aspx?f=claw.Cateogory&q=effectlevel%3a司法解释），可按时间排序提供科技政策信息浏览和全文下载功能。科技政策的数据量日益增

作者简介：曾文（1973—），女，中国科学技术信息研究所副研究员，研究方向：智能信息处理、情报分析和知识组织等（通讯作者）；**李智杰**（1992—），男，中国科学技术信息研究所硕士研究生，研究方向：知识组织；**王小玉**（1992—），女，中国科学技术信息研究所硕士研究生，研究方向：科技资源管理；**董诚**（1970—），男，中国科学技术信息研究所研究员，研究方向：科技管理与科技创新。

基金项目：国家社会科学基金项目“基于事实型科技大数据的情报分析方法及集成分析平台研究”（14BTQ038）。
收稿日期：2017 年 1 月 16 日。

长，科技政策涉及的内容广泛而复杂，如何准确快速地挖掘科技政策中的核心信息，急需对科技政策进行深层的内容分析，而其基础工作是科技政策术语的识别。所谓科技政策术语是指科技政策文本中的词语。本文拟对科技政策术语的自动识别进行初步探讨。

术语自动识别方法主要分为3类：基于规则的方法、基于统计的方法以及统计与规则相结合的方法。除此以外，还有一些新颖但应用相对较少的方法[4]。其中，基于规则的方法主要利用术语词典和规则模板进行术语抽取，即把一些常用的术语收入词典作为基础，对于词典中没有的术语，则通过构建规则模板的方法来识别[5]。该方法对特定领域和特定类型的术语识别具有良好的效果，但该方法需要掌握术语的构词规则，其适应性及可移植性较低。基于统计的方法是以统计理论为基础，利用术语已经在语料库中的分布统计属性来识别术语，即从概率意义上衡量多字单元是否为术语[6]。比较经典的方法是词频统计方法、互信息和信息熵方法、基于统计机器学习的方法等。基于统计的方法相对于基于语言学的方法来说不需要特定专业知识或资源，因此可移植性较好，但是其统计计算需要依靠大规模的语料库。单独地使用基于规则的方法或基于统计的方法进行术语自动识别，或多或少会存在较大的误差，而将两者结合起来使用，则会提高术语自动识别的准确度。因此，使用两者结合的方法进行术语自动识别是目前的主要方法，其具有代表性的方法是*C-value*方法及其改进的方法。此外，其他新方法还有扩展法。其主要思想是通过种子术语[7]、中心词串、术语部件等术语核心部分进行扩展，以抽取术语。如，文献[8]设计了一种串扩展算法，对一个中心串集的每一个中心串，需在领域语料中找出包含这个中心串的句子集合，并对其中每一个句子进行单句串扩展操作。这种算法用于密码学领域的术语识别，取得了不错的效果。文献[9]提出一种术语部件扩展算法来自动识别术语。术语部件是特定领域中构成术语能力较强的单词或语言片段，通过对领域文本分词后，判断每一分词串是否包含术语部件，若包含则对其向左向右进行扩展，扩展时结合词性及词语长度的规则进行判断，向左扩展时若到了符合首词规则的词时则终止，向右扩展时若到了符合尾词规则的词时则终止，两者都终止即得到了候选术语。该方法的不足之处是随着术语的不断更新变化，无法保证构建出一个完整的术语部件库，同时很多术语并不包含所谓的术语部件。

因此，本文首先从科技政策的术语特点和语言规则入手，分析适用的术语识别方法。

2　科技政策术语的语言规则与统计分析

科技政策术语主要有以下5方面的语言特点。

（1）中心词普遍存在。中心词指科技政策文本中频繁出现的基本术语，多数非单个词的术语是由中心词组成的名词性结构或谓词结构等。

（2）连接结构。科技政策术语中存在一些专有术语，它们的词素和词素之间通过符号连接，如协议标准GB/T19596−2004。

（3）数据存在稀疏现象。科技政策中有些术语只出现一次或少数几次。

（4）术语存在嵌套。多个词组成的术语由单个术语组合而成，使这些术语存在嵌套关系。

（5）停用词表内容不同。科技政策文本用词较为严谨，政策领域的停用词和通用停用词表相比，没有“哦”“哈”等语气词，没有拟声词，没有相对白话的转折词，没有人物代词，没有相对特殊的符号，但是有部分公文领域常用词。

在科技政策文本中，作为反复使用且形式较为固定的、表达某一特定概念的词语，术语的组成结构一般具有词性特点。能够构成术语的词一般为名词、动词、形容词等，有些词性的词是不能作为术语出现的，如连词、介词、副词、语气词等。考虑到科技政策中术语的特点，保留部分动词性成分、形容词性成分和前后缀。相关词性如表1所示。

此外，在科技政策文本中，领域专业术语用词较为严谨，同时对比已有的公文术语词典，发现其构词长度大部分是单个词、2个词和3个词，

所以本文选择识别长度在 1 ～ 3 个词的术语。可以发现：多数科技政策术语的语言规则为二元术语及三元术语，如汽车节油率、资源节约型等，也含有少量单词术语和四元及其以上的术语。针对这些科技政策术语的词性构成，构造科技政策术语常用的语言规则模板如表 2、表 3 和表 4 所示。其中，一元指一个词性标记代表一个术语；二元指两个词性标记代表一个术语；三元指三个词性标记代表一个术语等。

经语言规则过滤处理可得到初次的候选术语集，其中还会包含非术语的普通词语搭配、无意义的词语搭配。为了进一步得到正确的术语，本文采用统计计算进行二次过滤候选术语的策略。由于科技政策术语存在嵌套现象，因此本文基于*C-value*的统计方法进一步过滤候选术语。*C-value*方法是一种实现多词语自动术语识别，且与

表 1　词性释义

词性标记	含义	词性标记	含义
n	名词	*vn*	名动词
nl	名词性惯用语	*vi*	不及物动词
ns	地名	*vg*	动词性语素
ng	名词性语素	*a*	形容词
nsf	音译地名	*an*	名形词
nt	机构团体名	*ag*	形容词性语素
nz	其他专名	*b*	区别词
v	动词	*h*	前缀
vd	副动词	*k*	后缀

表 2　科技政策术语一元语言规则

规则含义	词性标记说明	含词个数
4个以上名词	大于等于4个字的n（如：市场占有率，高新技术）	一元
地名	*ns*（如：北京市）	一元
机构团体名	*nt*（如：财政部）	一元
其他专名	*nz*（如：高新，比亚迪）	一元

表 3　科技政策术语二元语言规则

规则含义	词性标记说明	含词个数
名词+名词	*n+n*（如：汽车产业）	二元
动词+动词	*v+v*（如：贯彻落实）	二元
动词 + 名词	*v+n*（如：汽车研发）	二元
名词 + 名动词	*n+vn*（如：工业信息化）	二元
名动词 + 名词	*vn+n*（如：混合动力）	二元
区别词 + 名词	*b+n*（如：电动汽车，公交车）	二元
形容词+名动词	*a+vn*（如：轻量化）	二元
区别词+名动词	*b+vn*（如：公共服务，高效节能）	二元
名词+名词性语素	*n+ng*（如：城市群，正负极）	二元
名动词+名词性语素	*vn+ng*（如：循环链）	二元
名词 + 名形词	*n+an*（如：交通安全）	二元
名动词+名动词	*vn+vn*（如：示范推广）	二元
动词 + 名动词	*v+vn*（如：维护保养）	二元
动词 + 名词性语素	*v+ng*（如：污染物）	二元
不及物动词+名词	*vi+n*（如：充电设施，转向系统）	二元

表 4　科技政策术语三元搭配规则

规则含义	词性标记说明	含词个数
动词+名词+后缀	*v*+*n*+*k*（如：插电式）	三元
名动词+名词+后缀	*vn*+*n*+*k*（如：节能环保型）	三元
名词+名动词+后缀	*n*+*vn*+*k*（如：汽车节油率）	三元
名词+动词+后缀	*n*+*v*+*k*（如：资源节约型）	三元
名词+形容词+后缀	*n*+*a*+*k*（如：环境友好型）	三元
形容词+名词+后缀	*a*+*n*+*k*（如：低碳化）	三元
形容词+名动词+后缀	*a*+*vn*+*k*（如：单一充电式）	三元
形容词+名词+名词	*a*+*n*+*n*（如：新能源汽车）	三元
形容词+区别词+名词	*a*+*b*+*n*（如：纯电动汽车）	三元
区别词+名词+名词	*b*+*n*+*n*（如：电动汽车产业）	三元
区别词+名动词+名词	*b*+*vn*+*n*（如：远程监控装置）	三元
区别词+名词+名动词	*b*+*n*+*vn*（如：一次性财政补贴）	三元

领域无关的方法，其综合运用了统计学和语言学的信息，目的是改进嵌套术语（nested terms）的识别。由于*C-value*方法充分考虑了术语长度的问题和嵌套术语的问题，可以在一定程度上改进嵌套术语的抽取准确率。

*C-value*算法的基本思想是：（1）如果一个字符串作为子串出现在长的多词术语中的频率很高，而它作为单独术语出现的频率很低，那么尽管这个字符串的整体词频很高但其很有可能不是术语；（2）如果一个字符串经常作为子串出现在多个不同的多词术语中，那么这个字符串是术语的概率更大；（3）如果两个长短不同的候选术语具有相同的词频，那么长字符串是术语的可能性更大。一个候选术语的*C-value*值的大小与其在语料中的词频和长度成正比，如果候选术语是其他词语的子串即候选术语被嵌套，其*C-value*值会相应降低。其计算公式如下：

$$C-value(s)=\begin{cases}\log_2|s|f(s) & \text{如果}s\text{未被嵌套}\\ \log_2|s|\left(f(s)-\dfrac{1}{p(T_s)}\sum_{w\in T(s)}f(w)\right) & \text{其他}\end{cases} \quad (1)$$

其中：s表示一个候选术语；$|s|$表示候选术语s的长度；$f(s)$表示候选术语s的词频。

如果s被嵌套，T_s指以s为子串的候选术语，指以s为子串的候选术语总个数，w指T_s中任意的以s为子串的候选术语，为w在候选s的上下文中出现的次数。通过以上公式计算每个候选术语的*C-value*值。*C-value*值越高，该候选术语成为术语的可能性就越大。把*C-value*小于某个阈值的术语去掉，则可得到二次过滤后的术语结果。

3　科技政策术语的停用词表与自动识别的算法设计

与其他领域术语的停用词相似，科技政策术语停用词表也包含符号、数字和无实际意义的某些词。为了找到停用词，需要依据一定的标准计算得到。最基本的计算标准是利用词频的大小判断。词频评估函数的理论假设是，通常高频词与高噪声值具有相关性，即当一个词的词频非常高时，很有可能是噪声词。本文利用中国科学院NLPIR−ICTCLAS 2014分词系统对所搜集的科技政策进行分词，统计经过分词及词性标注后的政策文本中所有词的词频，可以发现，一些没有实际意义的词，如“的”“是”“和”等虚词、连词（即停用词）出现次数非常多，这些词不能出现在术语中。同时，一些频繁出现的常用词，如“服务”“推广”“加快”“我们”等，虽然有实际意义，但不包含领域专业信息，同样不能出现在术语中。所以，对于停用词，直接将它们存入停用词表中；对于常用词，对照相应公文领域及科

技领域主题词表，以词频及主题词表判断作为依据，选择不是术语的常用词，存入停用词表文件中。如表 5 所示。

为提高科技政策术语自动识别的准确性，本文将科技政策术语的自动识别过程分为 3 部分：一是利用停用词表过滤科技政策文本数据；二是利用科技政策术语语言规则识别候选科技政策术语；三是利用统计计算实现科技政策术语的再识别和过滤，以形成整个科技政策术语的自动识别过程。具体算法流程见图 1。

表 5　科技政策领域停用词表（部分）

通用停用词	常用词
对	发展
为	提高
与	建立
而	加快
按照	加强
各	形成
由	达到
之间	实现
一般	主要
只	开展
就	相关

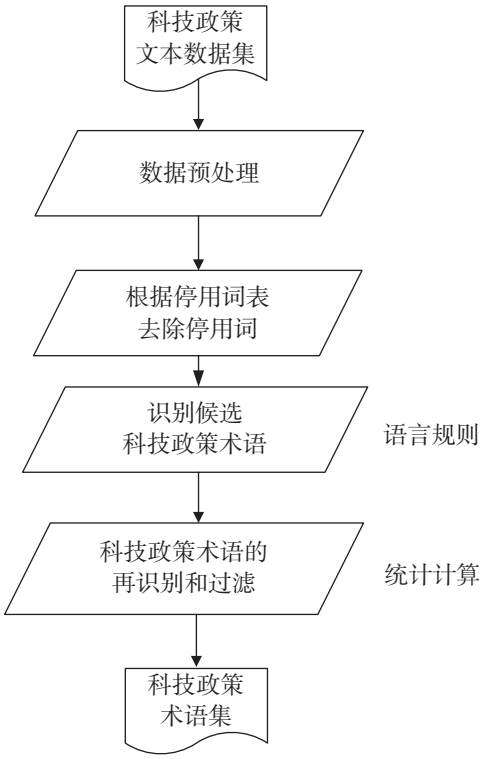

图 1　科技政策术语自动识别算法流程图

图 1 中的数据预处理主要指实现分词和词性标注，以实现后续的组词过程。经过停用词表去除科技政策中的停用词，使用科技政策术语的语言规则进行术语的第一次过滤，将不满足条件的词语删除。之后，采用统计计算进行候选术语的第二次过滤，最终得到科技政策术语集。

4　实验分析

本文选取 1426 条科技政策作为实验数据，通过开发 JAVA 编程语言程序实现上述科技政策术语的自动识别算法，并进行术语识别效果的检验。利用科技政策术语的语言规则进行术语的过滤。第一次过滤后得到科技政策术语 69720 个，第二次统计计算过滤后得到科技政策术语 83745 个，并对识别后的术语进行排序。通过对比，可以发现第二次统计计算解决了术语的嵌套问题，增加了识别后术语的数量。由于目前国内外尚无有关中文科技政策术语的抽取算法及相关科技政策主题词表或词表用于术语结果的对比，因此对文中经过算法识别的术语结果只能通过人工方法来判断是否正确。由于数据集较大，同时为了保证人工判断的相对客观性，本文分别在 83745 个术语集中随机选取前端 1000 个术语（$T1000$，统计计算值较高），中间 1000 个术语（$M1000$，统计计算值中等），后部 1000 个术语（$B1000$，统计计算值较低），并对这 3000 个经过算法识别的术语进行人工判断，分别计算 $T1000$、$M1000$、$B1000$ 术语的准确率，最后通过取平均值得到本文术语识别方法的准确率。准确率计算方法如下：

$$\begin{aligned} &Accurcy\_terms(T,M,B) \\ &=(\frac{Number\_accurcy\_terms(T)}{Number\_terms}\times 100\%+ \\ &\frac{Number\_accurcy\_terms(M)}{Number\_terms}\times 100\%+ \\ &\frac{Number\_accurcy\_terms(B)}{Number\_terms}\times 100\%)/3 \end{aligned} \quad (2)$$

准确率的计算结果如表 6 所示。

表 6　术语识别准确率

术语选取	准确率/%
*T*1000	87.80
*M*1000	84.50
*B*1000	79.60
总术语	83.97

根据实验结果可知，本文设计的科技政策术语识别方法具有一定的可行性，

5　结论

科技政策术语既是构建科技政策领域词表的词汇基础，也是对科技政策进行深层次数据挖掘的基础。本文提出的基于科技政策术语语言特点和统计计算相结合的术语自动识别方法，可以应用于科技政策词表的构建过程和科技政策语义分析过程。实验结果表明，该方法具有一定的术语抽取效果，但将受到数据集选择规模的大小或数据集内容质量的高低的影响，术语识别的准确度达不到人工识别的精确度和智能性。此外，实验结果有效性的对比问题仍有待进一步的研究。因此，在科技政策术语自动识别的具体算法设计和实现有待进一步的深入研究 [10]。

参考文献

[1] BERNIER−COLBORNE G,DROUIN P.Creating a test corpus for term extractors through term annotation[J]. Terminology,2014,20(1):50−73.
[2] 袁劲松,张小明,李舟军.术语自动抽取方法研究综述[J].计算机科学,2015(8):7−12.
[3] 张二艳.术语自动抽取技术研究[D].哈尔滨:哈尔滨工业大学,2009:20−50.
[4] 杨雅娜,刘胜奇.基于TValue融合领域度的术语抽取法[J].情报工程,2015(5):25−31.
[5] 闫琪琪,张海军.中文领域术语自动抽取方法进展研究[J].电脑知识与技术,2014(28):6716−6718.
[6] 季培培,鄢小燕,岑咏华.面向领域中文文本信息处理的术语识别与抽取研究综述[J].图书情报工作, 2010(16):124−129.
[7] MEIJER K,FRASINCAR F,HOGENBOOM F.A semantic approach for extracting domain taxonomies from text[J].Decision Support Systems,2014,62:78−93.
[8] 陈士超,郁滨.面向科技领域的术语自动抽取模型[J].系统工程理论与实践,2013(1):230−235.
[9] 闫琪琪,张海军.一种混合策略的领域术语自动抽取方法[J].电子制作,2015(8):50−51.
[10] 曾文,李颖,韩红旗,等.海量数据的组织与管理方法研究[J].情报工程,2016,2(1):109−113.

（上接第 12 页）

项目获资助后，注意在经费上给予相应的倾斜，以保证课题顺利执行。同时，研究单位管理部门应该根据自身的发展，积极规划、建设与科研水平相适应的动物设施、做好实验动物人才贮备，并投入相匹配的经费以促进实验动物科学的发展，使实验动物科学发挥其对生物医学的支撑作用。

参考文献

[1] 李继平, 金剑, 秦川.实验动物在医学创新研究与发展中的作用[J].中国医药导报, 2014, 11(31): 152−155.
[2] 张连峰.我国常用实验动物资源的现状及对未来发展的思考[J].中国比较医学杂志, 2011, 21(10): 39−44.
[3] 卫茂玲.医学动物替代研究发展现状研究[J].中国医学伦理学, 2016, 29(2): 304−307.
[4] 陈振文.首都医科大学实验动物学科建设10年回顾[J].实验动物科学, 2015, 32(5): 40−42.
[5] 李大鹏, 刘文清, 王永清, 等.实验动物工作在军队医院科研中的价值及实现要素[J].实用医药杂志, 2008, 25(9): 1142−1143.
[6] 陈领, 胡景杰, 陈越, 等.重视和加强中国实验动物的研究[J].动物学杂志, 2013, 48(2): 314−318.
[7] 徐艳霞.医院科研课题实验动物成本核算研究[J].中华医学科研管理杂志, 2014, 27(2): 233−235.
[8] 中国医学科学院.中国医学科学院北京协和医学院年鉴[M].北京: 中国协和医科大学出版社, 2014.
[9] 国家自然科学基金委员会科学基金网络信息系统[EB/OL][2017−01−08].https: //isisn.nsfc.gov.cn/egrantweb/.
[10] 梁力均, 石晶, 张纯, 等.实验动物在综合医院科研工作中的使用情况[J].实验动物科学, 2009, 26(6): 47−49.