Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

基于汉语科技词系统的专利文献标引及应用研究*

□ 张兆锋 / 中国科学技术信息研究所 北京 100038
/ 南京大学信息管理学院 南京 210093
桂婕 张运良 / 中国科学技术信息研究所 北京 100038
刘喜文 / 南京大学信息管理学院 南京 210093

摘要：文章介绍了利用汉语科技词系统的词表及词间关系对中文专利文献进行主题标引的研究进展，根据专利文献的特点设计了相应的标引策略和流程，并进行实验及结果分析，证明了本标引方案的有效性，最后对标引结果在专利检索中的应用特色进行了介绍。

关键词：标引，专利标引，主题标引，科技词系统
DOI：10.3772/j.issn.1673—2286.2013.11.003

1 引言

专利作为一种科技文献，与期刊论文相比，它提供更全面、更直接的技术信息，其内容具有广泛性、可靠性、创造性、实用性的特点，是掌握最新技术的重要信息源之一。据研究，全世界的发明成果70%~90%出现在专利文献中。如果充分利用专利文献，可以缩短60%的科研周期，节约40%的科研经费[1]。专利文献是科技创新的体现，同时又是创新的基础。随着十八大“实施创新驱动发展战略”，作为创新载体的专利文献资源的挖掘与利用必将受到越来越多的重视。

专利文献与科技论文相比，无关键词字段。因此它不能像论文一样可通过关键词准确地揭示论文的主要内容，提高检索的准确性和效率，并基于关键词进行内容方面的深度挖掘，如文献自动分类和相似性计算等。为了更有效地利用专利文献资源，服务于技术创新、科研和支持管理决策，本文尝试基于汉语科技词系统对中文专利文献进行主题词标引，进而给每篇专利赋予主题词，以便更好地揭示资源，充分利用专利文献，实现专利信息的精准检索及与主题相关的分析挖掘服务。

本文在接下来的章节会首先简要介绍现有的文本标引方法，基于此提出本文采用的标引方法，并介绍相应的词表及标引策略设计。然后进行实际的标引程序开发实验，并对实验结果进行分析，总结此种标引方法的优点和不足，最后对标引的结果的应用场景进行设想。

2 专利标引概述与汉语科技词系统

2.1 专利标引概述

专利标引指用一个或多个词来表现专利内容特征及相关技术、算法、组件的过程[2]。根据自动化程度可分为手工标引、机助标引和自动标引；根据标引的词语的来源不同，可分为抽词标引和赋词标引[3]。专利标引的主要对象是专利主题、核心技术、重要算法、关键部件等，便于建立专利内容层面的知识关联，实现对隐含信息的挖掘。李宏芳等人对三个较权威的中文专利数据库的标引质量进行了测评[4]，发现中文专利数据库的主题标引深度不够，对非题名关键词的标引不足，不便于从内容层面对相似题名的专利进行区分检索。同时，由于目前还是以手工标注为主，标引效率较低，标引结果也受标引人员主观影响较大。要改善此类问题，需要借助于

* 本文获得中国科学技术信息研究所预研基金项目“基于汉语科技词系统的专利文本标注模型构建与应用”（编号：YY201225）的资助。

大规模的权威词表和计算机的自动化技术[5]。

2.2 汉语科技词系统

汉语科技词系统（简称词系统：http://www.vocgrid.org/）是中国科学技术信息研究所提出并开发的面向中文为主的科技信息资源管理及深层次知识服务的知识组织系统[6]。该系统为中文科技信息资源的内容处理提供词汇层面的语义支撑，并建立了一定规模的领域科技词表，使对海量文献资源的智能、全面、准确的赋值标引提供了可能。本文探索如何利用大规模词表结合计算机智能技术对海量专利文献进行标引的方法。

3 标引策略设计

对专利文献进行标引，要首先了解专利文本的特点，根据特点设计标引的策略和流程。

3.1 专利文献的特点

专利文献作为一种科技成果载体，详细客观地描述了发明创造的对象名称、原理、组成、流程等内容。一般包括专利号、发明名称、摘要、权利要求书、国际分类号、发明人、申请人、申请日期等题录信息。专利文献不同于文学作品，专利描述的语言风格客观、朴实，不使用比喻、拟人等修辞手法。专利描述用词具体说有如下特点：

（1）文中的词语都是如实反映所描述的物体、组件、元素等对象本身的概念，为主题词赋词标引提供了很好的前提条件。

（2）专利主题词会多次出现。作为专利描述的主要对象，能够代表或接近专利主题概念的词汇在专利全文中往往会重复出现，多次被提及，因此是专利的主题词概率更高。

（3）由于专利发明多是对某一小部件或某一新类型的发明创造，因此词系统中的上下位词会在专利文本中有较多体现。而下位词往往是专利发明的具体对象，上位词是该发明的所属类别描述，因此标引时应使用下位词优先标引策略。

（4）专利发明标题与专利文摘和权利要求项相比，标题更能体现专利主题所在，在标引策略设计时应给予更高的权重。

（5）专利文本行文比较规范，很少出现口语化的词汇和缩略语、简称等。因此，在文中出现的能和主题词表中词汇匹配的词语都可作为主题候选词。

3.2 标引策略

以前的相关研究多为对新闻材料[7]、科技文献[2]、学位论文[8]等进行标引，对专利文献的标引研究较少。各种文献由于文体特点、内容、结构不同，需要制定不同的标引策略。本文详细分析了专利文献的特点，制订了如下的标引策略：

（1）标引源。专利发明名称一般能比较明确地指出发明的对象，但有些专利直接以一个较上位类的概念词作为发明名称，如“汽车”、“电池”等。仅以此为标引词，检索时还不能提高查准率。而专利摘要和权利要求书可以对发明名称有很好的补充。摘要是对发明的具体原理、结构、功能的概要介绍，权利要求书是对专利所声明保护权利的具体描述，专利所要保护的核心技术和对象会在权利要求书中有所体现。因此，本文选择发明名称、摘要和权利要求书作为标引源。

（2）标引权重。自动标引策略设计中对标引源权重的设计很重要，设置不当可能会遗漏主题词，或者引入干扰词，需要根据各个部分对主题的表达能力不同给予适当的权重。根据侯汉清、章成志、郑红等人对Web语料标引源加权方案的研究知道，“题名具有很强的表达能力”[9]，同样在专利中，专利发明名称应该具有最高的权重，同时根据专利标引源的特点，摘要和权利要求书描述中同样的主题或部件名称会重复出现，因此需要提高标题中主题词的权重，保证标题中出现的主题词被标引的优先权。基于此，对专利标引源权重设计如表1所示。

（3）选词。词系统中有一些单字主题词，如“碲、锆、镉、铑”等。单字主题词多为某元素名或很上位的概念，标引专利意义不大，且对确定正确的标引词有较大干扰，本文中的标引词选择词系统中词长大于或等于4个字节的主题词。

（4）标引算法。在专利文本中，下位词比上位词更具体，为了提高检索的查准率，优先标注下位词，一般来说下位词比上位词长度更长，因此标引时根据词长顺序进行文本匹配标引，并采用正向最大匹配算法。

表1 标引源权重分配表

字段	权重
发明名称	3
摘要	1
权利要求书	1

（5）确定标引词。根据文本中出现的主题词词频加权求和（简称权和），结果从高到低排序，取前5个主题词作为本篇专利的标引主题词。若与第5个主题词权和相同的还有其他主题词，则都列为标引主题词，权和为1的主题词舍弃，即使不够5个。

4 标引实验

4.1 实验环境

本实验采用的软硬件环境如下：

硬件环境：服务器内存2GB及以上，服务器CPU3.0GHz及以上，服务器硬盘空闲空间100G及以上。

软件环境：操作系统Windows XP SP2/SP3、Windows Vista、Windows7，客户显示器分辨率1024×768及以上，数据库SQL Server 2008及以上版本，浏览器采用IE 7.0以上，IIS7.0、.Net 3.5及以上。

4.2 数据库设计

本实验选择的标引源为12041条专利新能源汽车领域的中文专利，主要字段为专利号、发明名称、摘要和权利要求书。用于匹配的词系统主题词为54750（包括核心词，不含单字主题词）。

数据库表存储标引源和标引结果，本实验用到的主要表格如表2、表3、表4所示。表2用来存储标引源数据，表3存储领域主题词，表4为词间关系表。

4.3 标引流程图

标引的流程图如图1所示。先取一条专利，读取该专利的发明名称，然后调用词系统中的相应领域词表进行正向最大匹配。如果某主题词在标题中有匹配，则计算该主题词权和为词频数乘3，并记录在标引库中。接下来依次对摘要和权利要求书进行标引统计，权和计算为词频乘1，存入标引库中。在该专利三部分标引完成后根据标引库中的记录计算各主题词的总权和，根据权和的大小从高到低排列，取权和最大的5个词为标引主题词，然后处理下一条记录，直到所有待标引专利处理完成。

表2 标引专利表

字段	属性	说明
PN	Varchar(20)	专利号（主键）
TI	Varchar(50)	发明名称
AB	Varchar(max)	摘要
CL	Varchar(max)	权利要求书
Label	Varchar(max)	标引记录
Words	Varchar(200)	标引词

表3 领域词表

字段	属性	说明
ID	Int	序号（主键）
Word	Varchar(50)	主题词
Type	Varchar(10)	基础词或核心词
Color	Varchar(10)	用于展示标引结果
Times	Int	记录被用来标引总次数
Length	Int	词长

表4 词间关系表

字段	属性	说明
Word1	Varchar(50)	主题词1
Relation	Varchar(20)	关系描述
Word2	Varchar(50)	主题词2

5 实验结果讨论

5.1 标引结果展示与分析

为了便于分析标引的结果，把标引的结果以网页的形式展示出来，并通过不同的颜色来区分标引词是基础词还是核心词。如图2所示，左侧为被标引的专利列表，右

侧为标引结果显示，能与主题词表匹配的词都以颜色标注出来，显示红色的为核心词，蓝色的为基础词。词频统计部分为在该篇专利中涉及的主题词及数据统计结果展示。主题词后边括号内“/”前后有两个数字，前者为该词在本篇专利中出现的词权和，后者为该词在所有标引源专利中出现的词权和。

在专利技术检索时，检索者最重要的检索途径是专利产品名称、产品部件、核心技术、核心算法等。因此，在对标引结果进行评估时主要是看能指引到这条专利的这些核心部分是否标出。由于专利标引即使是手工标引，不同的人标引结果差别也比较大，而对标引结果的评估主观性也比较强，因此，笔者采用多人打分取平均值的方法进行结果评价。具体做法是，随机取500条标引结果，分为5组，由5人对结果进行打分，打分方案如表5所示，根据标引词对专利内容主题的覆盖度进行打分。通过对打分结果的统计计算，标引结果的平均得分为81.5分，最多的标引词为8个，最少的标引词为5个，平均单篇的标引词数为6.3个。

根据统计结果可知，标引词对专利文本内容有较好的覆盖，但也有不足的地方，在已选为标引词的主题词中也有一些是没有标引意义的，如“产品(4/509)”，说明“产品”一词在某专利中出现权和为“4”，总权和为“509”，“产品”一词为普通概念，没有专指性，不适合做标引词，同样的情况还有“运行(3/1962)”、“系统(3/3042)”等。通过分析可知，词系统中收集的领域词汇是该领域尽量全的词汇，包括一般性概念词汇，而专利中的检索大多以名词为主，专指性强，而标引

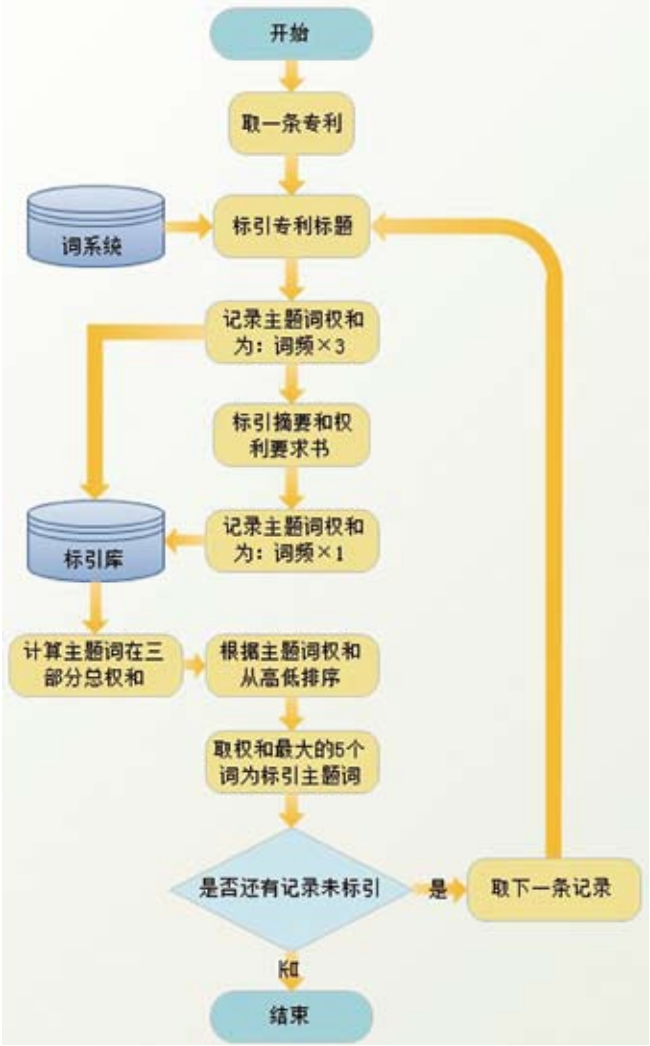

图1 专利标引流程图

图2 专利主题词标引结果

的正确性还有赖于词表的有效性。因此，应在词系统中建立专门用于专利标引的词表，同时评价时考虑词语之间的语义关系、部件名词之间的组合关系，可以有效提高标引结果的有效性和满意度。

5.2 标引结果应用

利用词系统的领域主题词对专利文献的主题标引，可以充分利用词系统的特色功能，对专利资源进行合理存储、深度揭示和精确检索，并利用主题词建立与其他科技资源的关联。具体的特色应用如下：

（1）通过词间关系进行扩缩减，有效提高检索的查全率和查准率。用来标引的主题词都是词系统中收录的主题词，由于主题词之间建立了各种关系，可以充分利用词间关系进行检索。如图3所示，当在检索框中输入关键词“ABS”时，在输入框下自动列出与“ABS”有相关关系的主题词，包含“材料-成品”、“拆解为”、“借助”、“类属”、“全称-缩略同义”、“异名同义”、“子类”等7种关系，同时在右侧列出相应关系对应的主题词，通过勾选主题词前的复选框可以扩缩检索范围或者提醒用户具有相关关系的主题词，提高用户检索的针对性。

（2）提高检索效率。由于专利申请量逐年激增，据统计，今年到目前为止（9月）的专利申请量已超过去年全年。标引后的专利可以根据标引词检索，避免对上千万条专利全文匹配检索的存在歧义、效率低下的缺点，实现专利技术精确快速检索定位。

（3）实现与其他科技资源的关联。若用类似的方法把科技文献、科技论文或科技新闻等资源也用词系统的主题词进行标引，可以实现以主题词为纽带的资源关联，更有效地把各类资源整合起来，实现为科研技术人员的一站式资源提供服务。

（4）新词发现。由于专利文献是发明创造的描述，经常会有新的词汇创造出来，而在词系统中本来是没有的。通过对标引结果的分析可看出，有些标引词在文中是连在一起的，而且本身可以作为一个主题词，而词系统中却没有收录。比如，有篇专利名称为“折叠式电动踏板车前置儿童座椅”，在本系统标引结果页面显示“儿童座椅”四字皆为蓝色、但统计结果是：儿童(3/134)、“座椅(3/387)”，说明系统中只收录了“儿童”、“座椅”两个主题词，而“儿童座椅”没有被收录，它可以作为“座椅”的下位词添加进词系统。因此，标引结果可以

表5 标引结果评价打分方案

分值	属性
100	主题词、重要部件全部标出，标引词能覆盖专利主要内容主题
80	遗漏1个主题词，标引词能覆盖80%的内容主题
60	遗漏2个主题词，标引词能覆盖60%的内容主题
40	只标出2-3个主题词，标引词能覆盖40%的内容主题
20	仅标出1个相关主题词，标引词只能覆盖20%的内容主题
0	所有标引词都与专利内容主题无关

图3 基于词系统的专利检索

用来进行新词发现，通过设置一定的推荐机制，根据标引的结果向词系统推荐新词，经过专家审核后正式成为主题词。

此外，还可以根据标引结果数据的统计反过来优化词系统的构建。比如，在主题词表中检索词词长大于16个字节且被用来标引次数为0的主题词中，会发现有些不是主题词的记录，如“变速器输入轴与输出轴以各自的速度旋转”、“能自动对各车轮的制动和发动机动力进行控制”等。通过这种方法可以快速地对加工后的词表质量进行评价，发现并删除词表建设中所收录的错误词条，提升词系统建设的质量。

6 结语

本文利用汉语科技词系统新能源汽车领域词表的建设成果，对该领域的中文专利进行主题标注模型设计，并进行实证分析。实验结果表明，基于词系统的权威性、语义性、全面性，标引结果能达到令人满意的结果，通过建立针对专利标引的专用词表，更能有效提高标引质量。此外，通过对标引后的专利与词系统的结合，提供专利的语义检索，提高了检索的查全查准率，同时降低了用户的检索难度，提高了专利检索系统的易用性。同时，通过标引系统与词系统的接口设计，保持了标引系统用词与词系统主题词建设同步更新。

本文主要探讨利用主题词表及关系对专利标引的方法，未来可以把语法、语义的因素结合进来，实现综合的智能标引，进一步提高标引的准确性和完备性，更有效地实现专利资源的揭示和挖掘，为企业创新和决策支持服务。

参考文献

[1] 魏衍亮.企业专利情报战略初探[J].中国科技产业,2004(7):45-49.
[2] 苏新宁,邹晓明.文献信息自动标引研究[J].现代图书情报技术,2000(1):23-26.
[3] 章成志,苏新宁.基于条件随机场的自动标引模型研究[J].中国图书馆学,2008(5):89-94,99.
[4] 李宏芳,邹小筑.中国专利数据库标引质量测评[J].现代情报,2010(12):58-61.
[5] 章洪流,徐伟,吴倩,等.关键词标引常见问题探讨[J].中国发明与专利,2008(8):65-67.
[6] 乔晓东,张运良,朱礼军.汉语科技词系统建设与应用进展[J].情报学报,2010,29(6):978-986.
[7] 查贵庭,侯汉清.基于多词表的自动标引技术研究:新华社新闻稿自动标引的实验[J].情报学报,2002(3):273-277.
[8] 全根先.学位论文的主题标引及其规范[J].学理论,2011(30):89-91,97.
[9] 侯汉清,章成志,郑红.Web概念挖掘中标引源加权方案初探[J].情报学报,2005,24(1):87-92.

作者简介

张兆锋，男，1979年生，在读博士，助理研究员。研究方向：专利分析、数据挖掘、信息可视化。E-mail: zhangzf@istic.ac.cn
桂婕，女，1976年生，博士，副研究员。研究方向：专利分析和科技创新管理。E-mail: guij@istic.ac.cn
张运良，男，1979年生，博士，副研究员，研究方向：知识组织、知识工程、自然语言理解、文本自动分类。E-mail: zhangyl@istic.ac.cn
刘喜文，男，1983年生，在读博士。研究方向：数据挖掘、本体技术。Email:liuxiwenhit@163.com

Research of Patent Indexing and Application Based on Chinese Scientific and Technical Vocabulary System

Zhang Zhaofeng / Institute of Scientific and Technical Information of China, Beijing, 100038
/ Nanjing University, Nanjing, 210093
Gui Jie, Zhang Yunliang / Institute of Scientific and Technical Information of China, Beijing, 100038
Liu Xiwen / Nanjing University, Nanjing, 210093

Abstract: This paper introduces a method on how to index patent based on Chinese Scientific & Technical Vocabulary System. Tactics and flow are designed according to the characteristics of the patent literature. And experiment is also made, then the authors analyze the result, which verifies the availability of the method. Lastly, special application features of the result are also mentioned.

Keywords: Indexing, Patent indexing, Subject indexing, Scientific & Technical Vocabulary System

（收稿日期：2013-10-14）