汉语科技词系统在文献自动赋词标引中的应用研究*

闫莹莹 许德山 张运良 李鹏 / 中国科学技术信息研究所 北京 100038

摘要：文章首先介绍了汉语科技词系统的体系结构和功能，其次设计了自动赋词标引研究的整体思路，完成了自动赋词标引的系统功能实现，包括标引知识库的格式转换、算法实现和系统实现，并收集语料进行测试。最后对自动赋词标引的结果进行了分析，并且总结了该自动赋词标引研究的特点和不足，介绍了未来的工作设想。

关键词：自动标引，赋词标引，汉语科技词系统，标引知识库，词系统应用，D2RQ
DOI：10.3772/j.issn.1673—2286.2013.11.002

引言

《汉语科技词系统》是中国科学技术信息研究所在“十一五”科技支撑计划资金的支持下牵头研发的领域词系统。目前已建成包括新能源汽车、重大自然灾害监测与防御、新一代工业生物技术、新能源、智能材料与智能结构五个不同领域的词汇组织系统。建设初衷是希望能够通过词系统的相关建设支持我国在相关领域的自动信息分析处理，并进一步支持战略决策、科研发展和科技创新。

本文在汉语科技词系统的基础上，研究中文自动赋词标引系统。自动标引包括关键词自动提取和自动赋词标引两种。自动赋词标引是一种通过规范化的词语来描述文献主题的方法，特点是借助统一的词表，对文本的主题加以限定，这种方法能使相同主题的各种异构文献相对集中，从而对文本进行更加有序化和规律化的组织。

本文涉及《汉语科技词系统》的体系结构和功能介绍，并以该词表知识库为基础，设计了自动赋词标引的整体思路和算法，完成自动标引系统的开发。

1 词系统的体系结构和功能

汉语科技词系统[1,2]是吸收叙词表和本体思想的一种知识系统，它基于丰富的科技文献资源和知识工程师的努力，提供中英文对照、定义、关系、属性、多维分类和形式化概念描述等多层面的知识。目前汉语科技词系统包含新能源汽车、重大自然灾害监测与防御、新一代生物技术、新能源、智能材料与智能结构5个领域。

汉语科技词系统以词条（Term）为基本组织对象，在整体的知识结构包括：1）词条基本信息；2）词条定义及注释知识；3）词条之间的关系知识；4）词条的属性知识；5）词条的多维分类知识；6）词条形式化概念描述知识。其中，词条的基本信息包含词条的中文词形、对应的英文翻译、对应的拼音、词汇类型（即核心词/基础词区分）等知识要素。词条的定义主要是核心词，也就是那些在领域中处于核心骨干地位的词条，定义通常来自教科书、百科全书、科技期刊以及互联网。除了定义以外，还可以为词条添加有关变化、历史信息和知识工程师或者专家编辑审核体会的注释。词汇之间的关系从宏观讲仍然是等同关系、层级关系和相关关系，并对以上关系类型做了细化，尤其是对相关关系。细化既有通用的部分，也有针对新能源汽车特定的部分。属性用来表征一些依

* 本文系国家“十二五”科技支撑计划课题“科技知识组织体系共享服务平台建设”（编号：2011BAH10B03-2）、中国科学技术信息研究所重点工作项目“汉语科技词系统建设与应用工程”（编号：ZD2012-3-2）的研究成果之一。

附于主体存在的属性和属性的具体值，从而更全面地描述词汇（或者概念）。分类既提供了一个范畴或者粗分类表来管理词汇，又给出词汇与真实文本分类的相关关系，支持多维分类，包括中国图书馆分类法CLC和国际专利分类法IPC以及团队自己研制的针对新能源汽车的分类法。形式化概念描述采用HNC的概念符号体系描述，通过HNC概念描述，可以把有相同概念基元的词条聚成一个群落，并根据实际需求进行扩检和缩检，也可以计算词条之间的相似度，还可以进一步用于词空间构建。词系统的知识结构从总体来看，包含两种词汇组织方式：词汇定义组织和词间关系组织。

目前，汉语科技词系统已经通过Vocgrid网络平台（http://www.vocgrid.org/）对外提供服务。系统提供了基本的注册、登陆、认证、密码及注册码找回等基本功能，用户可以通过该平台访问获取词条的全部知识。经过认证的教育和科研领域的注册用户可以获得一个唯一的注册码，通过这一注册码还可以免费下载汉语科技词系统提供的数据、工具、程序、说明文档、演示程序等。截至目前，新能源汽车领域词系统中包含54,831条词条，其中5,712条为核心词，其余49,120条为基础词，包含推荐关系类型在内的76种关系类型以及57,821个关系实例，有52种属性类型，并建设了18,362个实例。面向新能源汽车的NEV分类法有4层154个类目，并且构建了5,431个类目实例。每一个核心词都包含对应的英译，系统中包含有5,431条定义。所有的5,712条核心词和另外的4,548条重要基础词拥有HNC概念描述。

2 标引整体思路设计

基于人工标引的语料库，依据新能源汽车词系统作为标引知识库，辅以机器学习方法，完成了一个自动赋词标引研究的整体思路设计。整体思路包括两部分：第一，将以MySQL数据库形式存在的汉语科技词系统转化成SKOS数据格式，传统数据库存储的词汇知识不便于结构化提取和利用，将汉语科技词系统转化成SKOS格式作为标引知识库来应用。第二，自动赋词标引流程设计，包括确定文献文本的标引候选词、计算候选词的特征权值、使用训练文本建立模型、应用模型进行标引[3]。自动赋词标引整体流程设计图如图1。

（1）确认候选词

该模块完成的功能是对加工后的文献文本进行处理，得到候选词汇序列，这些词汇序列表示整个文本的主题内容。序列中的词汇均为汉语科技词系统中的核心概念。

本文将主题标引看作从一堆主题词汇集合中挑选合适的主题词汇，分别进行文本分词、去停用词来减少数据噪音、文本词汇匹配到标引知识库词汇的操作，得到候选词汇序列，同时保存这些词汇的频率、位置等信息。

（2）计算候选词的权值

本文确定的候选词权值计算的指标包括几类：词频信息，位置信息，候选词本身领域代表性。

其中，“tf×idf”用于对比候选词在特定文献中出现的频次与该词的一般出现频次，以从一个角度测算该词代表文献主题的概率。“首次出现位置”和“末次出现位置”或者由两个指标决定的“跨度”（词汇在文本中首次出现和末次出现的位置的跨度大小）可用于从另一个角度来确定该词的代表性，一般出现在一篇文献的文本开始或结束部分的词相对比较重要。“节点度”是指

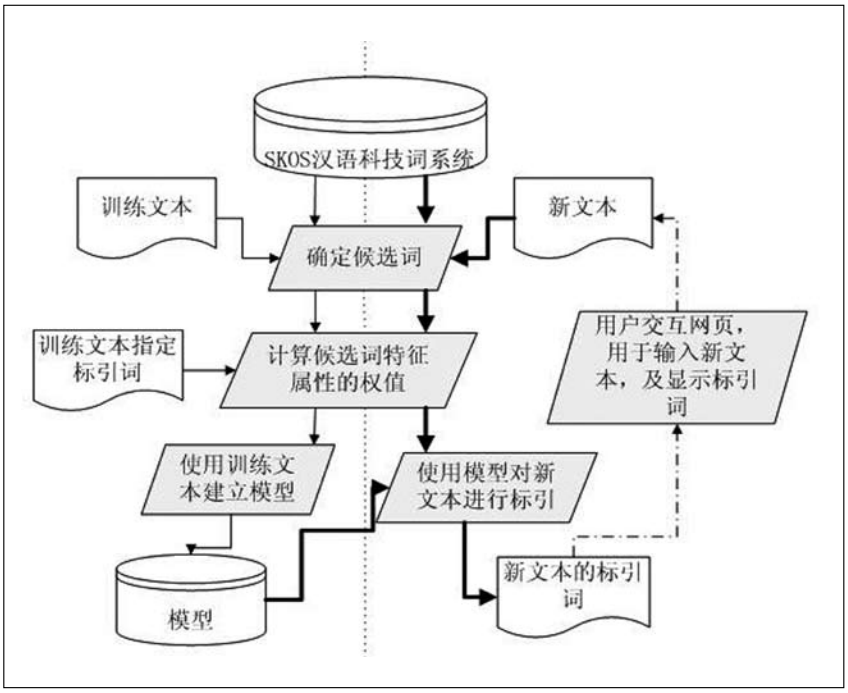

图1 自动赋词标引整体流程

在标引知识库的图中，节点度就是在图中有关联的词的个数。节点度高的词更能反映领域主题。

（3）构建模型

构建的模型为：将人工标引的文献主题词作为训练集，将候选词与其进行比对，采用贝叶斯分类算法，将符合人工标引结果的作为正集，不符合的作为反集。

（4）应用模型进行标引

通过计算候选词权值，对需要标引的新文献按模型进行计算，标引的词汇个数可以人为选定，得出最终的标引词汇。

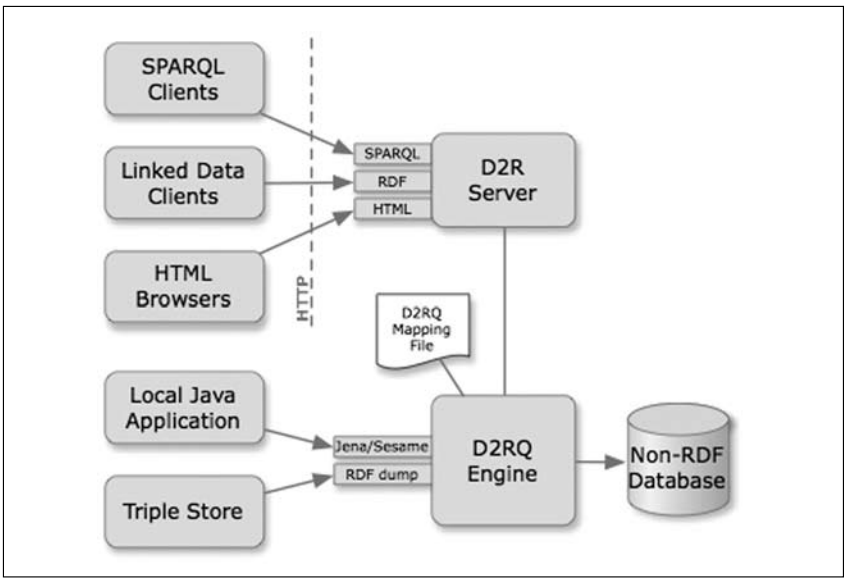

图2 D2R的主体架构

3 功能实现

3.1 标引知识库格式转化

Linked data的推动者们开发了一系列实用的工具，来帮助完成传统数据向Linked data的转换。D2R便是其中一个非常流行的工具[2]。它的作用是将关系型数据库发布为Linked data。D2R主要包括D2R Server、D2RQ Engine以及D2RQ Mapping语言。D2R Server是一个HTTP Server，它的主要功能提供对RDF数据的查询访问接口，以供上层的RDF浏览器、SPARQL查询客户端以及传统的HTML浏览器调用。D2RQ Engine的主要功能是使用一个可定制的D2RQ Mapping文件将关系型数据库中的数据换成RDF格式。D2RQ engine并没有将关系型数据库发布成真实的RDF数据，而是使用D2RQ Mapping文件将其映射成虚拟的RDF格式。该文件的作用是在访问关系型数据时将RDF数据的查询语言SPARQL转换为RDB数据的查询语言SQL，并将SQL查询结果转换为RDF三元组或者SPARQL查询结果。D2RQ Engine是建立在Jena（Jena是一个创建Semantic Web应用的Java平台，它提供了基于RDF、SPARQL等的编程环境）的接口之上。D2RQ Mapping语言的主要功能是定义将关系型数据转换成RDF格式的Mapping规则。图2呈现了D2R的主体架构。

本文将汉语词系统的RDF格式文件用在领域文献关键词提取中，使用D2RQ工具直接将关系型数据库中的数据包装成真实的RDF文件。将关系型数据库转化成RDF文件的步骤为：

（1）手工编制关于数据库schema的映射文件

生成真实的RDF文件最主要的内容在手工编制关于数据库schema的映射文件（mapping file）。Mapping语言中最重要的是两个概念，一个是d2rq:ClassMap，另一个是d2rq:PropertyBridge。

《汉语科技词系统》数据库Schema建立映射中用到的SKOS元素为skos:Concept；skos:prefLabel、skos:altLabel、skos:broader、skos:narrower。用到的《汉语科技词系统》中的数据库、数据表及表字段为Database:vocabulary、Table:Concept,relation、Concept:CID（概念ID字段）、CCN（概念名称字段）、relation:CID1（概念1ID字段）、CID2（概念2ID字段）、REL（关系名称字段）。映射文件编写要点示例如下：

（2）运行导出命令

```
片段一：
# Table  concept
: classmap_concept a d2rq: ClassMap;
d2rq:dataStorage:database;
d2rq:uriPatten"Http://www.vocgrid.org/nev#c_@@concept.CID@@";
```

```
: concept_CCN a d2rq:PropertyBridge;
d2rq:belongsToClassMap:classmap_concept;
d2rq:property skos:prefLabel;
d2rq:column"concept.CCN";
d2rq:lang"zh";
片段二：
# Table relation
: Classmap_relation1 a d2rq:PropertyBridge;
d2rq:belongsToClassMap: classmap_concept;
d2rq:property skos:broader;
d2rq:refersToClassMap:classmap_concept;
d2rq:condition"relation.REL";
d2rq:join"relation.CID1=concept.CID";
d2rq:join"relation.CID2=conceptcopy.CID";
d2rq:alias"concept AS conceptcopy";
```

图3 RDF格式汉语科技词系统示例

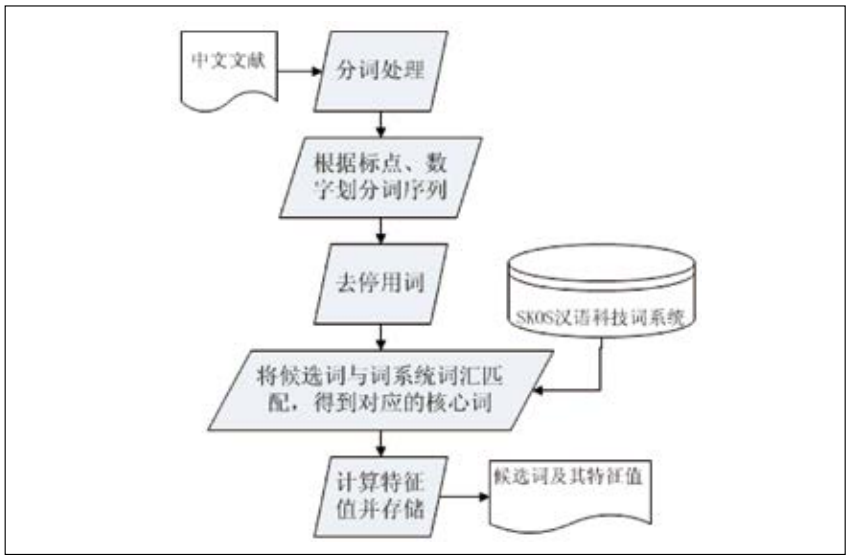

图4 识别文献候选词流程图

进入dump-rdf所在目录，在命令行中键入：

dump-rdf -f RDF/XML -o vogrid.rdf mapping-vocgrid.ttl即可。

（3）生成的RDF格式的汉语科技词系统示例（如图3）

3.2 标引算法实现

（1）识别候选词

识别候选词算法流程见图4，经过了四个主要步骤的处理。包括中文分词处理；分词后做去除停用词处理，得到词汇序列；对得到的词汇序列做文献文本词汇到标引知识库词系统词汇的映射；记录词汇的词频、位置及标引词集合信息。

算法中将处理后的文献词汇映射到词系统标准词汇是个重要的处理过程。该过程算法为：首先在词系统中寻找是否有与文献词汇相对应的标准词汇，如果有，判断是否为核心概念，若不是核心概念，则通过词系统中词汇的关联关系找到对应的核心概念名称描述，作为识别出来的一个候选词，直到完成这篇所有的文献词汇的匹配为止。同时存储得到的候选词的词频位置及标引词集合信息。

（2）定义候选词特征，计算权值

本系统把关键词标引看作从候选词集合中挑选关键词，挑选过程考虑以下一些可能因素，如词频、逆向文档频率、TF*IDF、第一次出现的位置、最后一次出现的位置、出现的跨度、节点度等因素。综合这些因素，建立了一个特征模板，如表1所示。

（3）利用训练语料生成模型

采用贝叶斯分类模型完成监督

学习过程。在确定特征集合后，对训练数据进行估计。从训练文档中获取候选词权值集合，对每一候选词分别依据在文献中的权值计算成为关键词或非关键词的概率值，并生成并保存模型。

（4）利用模型确定标引词，指定最终的标引集

利用模型，根据新文献的概率值来确定新文献的关键词集合。指定关键词的个数为5、10、15，选出概率在前的候选词作为文献的自动标引词。

3.3 系统实现

自动标引系统界面如图5所示。

用户选择输入文本后，可以选择两种方式进行文本的关键词自动标引，包括自由标引和赋词标引。标引结果会直接作为文本关键词显示给用户。

4 实验及结果分析

为了验证标引系统的效果，本文选用了新能源汽车领域50篇学位论文作为数据集，数据集中的论文包括题名、摘要、关键词，带有段落和章节、图表标题信息以及参考文献等部分。每篇论文平均55000中英文字符，范围从40000到85000不等（共4.5M）。标引词为人工对每篇论文进行标引的词汇，人工标引词每篇平均7.1个，各篇论文标引词从4个到13个不等。共355个指定标引词。

测试的方法采用ten-fold cross-validation（十折交叉验证），具体方法是将50篇论文分为10组，分别取1组为测试语料，其余9组为训练语料进行效果测试。

本文采用最常用的评测标准精确率Precision（P）、召回率Recall（R）和F-Measure（F）值对自动标引模型进行评价。计算公式为：

表1 特征模板

模板号	意义	名称	取值类型
1	TF*IDF	TF*IDF	Double
2	第一次出现位置	FirstOccur	Double(0-1)
3	出现跨度	spreadOccur	Double(0-1)
4	节点度	nodeDegree	Int

表2 各模板函数描述

TF*IDF	=TF*IDF 词频TF：候选词汇出现在某文献中的次数与文献中词汇总数的比值。 逆向文档频率IDF：出现候选词的文献数比语料中文献总数的LG值。
第一次出现的位置	候选词第一次出现的词汇的位置与文献词汇总数的比值。
词汇跨度	最后一次出现的位置与第一次出现的位置的差值。
节点度	在词汇知识库（汉语科技词系统）中候选词汇关联的词汇个数。

图5 系统界面

$$P=\text{自动标引与人工标引一致个数}/\text{自动标引个数}$$

$$R=\text{自动标引与人工标引一致个数}/\text{人工标引个数}$$

$$F=2PR/(P+R)$$

（1）标引结果示例如表3。

表3 试验结果1（标引示例）

标引效果较好的文献示例
题名：燃料电池电动汽车能量管理系统优化控制与动态仿真研究
人工标引关键词：电动汽车；燃料电池电动汽车；混合动力系统；能量管理； 燃料电池；模糊神经网络；优化控制；动态仿真；仿真
摘要：能源问题和环境问题目前在世界范围内已经成为一个被广泛关注的话题，各国在研究和开发新能源方面加快了步伐。燃料电池作为一种新型的、清洁高效的二次能源有着广泛的应用前景。燃料电池电动汽车更因其能有效降低尾气排放量，改善大气污染程度，成为汽车技术发展的新方向。燃料电池输出特性偏软决定了其单独作为车载能量源并不合适。因此，为提高车辆的动态性能及燃料效率，一般配置辅助能源与主能源燃料电池共同构成燃料电池电动汽车的混合动力系统。本文以研发燃料电池电动汽车为背景， 以混合动力能量管理系统为研究对象，开展能量管理优化控制与动态仿真研究，其主要研究内容如下： 通过分析燃料电池和镍氢电池等多种车载能源的特性，研究了燃料电池电动汽车(FHEV)动力系统的各种拓扑结构，提出并设计了燃料电池与镍氢电池组并联直连的混合方案。在此基础上，结合能量管理的要求，设计了一种分别基于驱动状态和制动状态的FHEV能量管理系统结构的结构。基于提出的混合动力能量管理系统结构，进行能量管理控制策略的研究。……
自动赋词标引关键词：能量控制；能量管理；电动汽车；电动汽车/驱动器；混合动力系统;汽车;控制策略；马达；镍氢电池；电池；发电装置；车载能源；燃料电池；续航里程；CTL

表4 试验结果2（P、R、F）

组号	5个标引词			10个标引词			15个标引词		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
1	20.00	12.50	15.38	10.00	12.50	11.11	6.67	12.50	8.70
2	40.00	27.78	32.79	24.00	33.33	27.91	16.00	33.33	21.62
3	36.00	20.00	25.71	18.00	20.00	18.95	13.33	22.22	16.67
4	20.00	13.89	16.39	18.00	25.00	20.93	12.00	25.00	16.22
5	16.00	13.33	14.55	16.00	26.67	20.00	10.67	26.67	15.24
6	20.00	14.71	16.95	16.00	23.53	19.05	10.67	23.53	14.68
7	16.00	9.76	12.12	8.00	9.76	8.79	6.67	12.20	8.62
8	12.00	8.57	10.00	6.00	8.57	7.06	5.33	11.43	7.27
9	12.00	9.68	10.71	20.00	32.26	24.69	13.33	32.26	18.87
10	8.00	7.41	7.69	8.00	14.81	10.39	5.33	14.81	7.84
AVG	20.00	13.76	16.23	14.40	20.64	16.89	10.00	21.39	13.57

（2）使用当前学位论文的语料，标引结果个数为5、10、15时，精确率、召回率和F值为表4。

（3）结果分析

在示例《燃料电池电动汽车能量管理系统优化控制与动态仿真研究》中，自动赋词标引的结果返回了“能量控制；能量管理；电动汽车；混合动力系统；汽车；控制策略；镍氢电池；电池；燃料电池”等，包括完全匹配与近似匹配的词汇，标引的正确率非常明显。即使与人工标引关键词不同的词汇，也大都描述了论文的内容。这表明，利用《汉语科技词系统》为文章做自动赋词标引的工作是很有效果的。

表4给出了50篇博硕士论文的自动赋词标引取不同个数的结果，召回率的平均值分别是13.76%、20.64%、21.39%，当标引词为10时，P、R、F的值分别是14.40%、

20.64%、16.89%。理论上影响评价指标的原因主要在于如下几个步骤：1）人工标引词的误差；2）中文分词结果；3）汉语科技词系统的收词；4）赋词标引的候选词识别算法；5）自动标引的学习算法；6）文献标引结果的主观性。

试验结果表明自动赋词标引工作还存在着很大的改进空间。我们针对具体的文献数据及标引流程做了详细分析，发现有些文献在候选词识别阶段的结果不理想，有些文献中抽取出来的候选词较少，甚至在候选词集合中就没有包含人工指定标引词中的词汇。产生这个现象的主要原因在于：（1）分词误差。本文采用最大正向匹配分词算法，无法切分出词典中没有的词汇。（2）词系统内容与人工标引的用词侧重点不同，如“模糊神经网络”“仿真”等词汇在新能源汽车词系统中并没有认定其为领域核心词。（3）在候选词识别中，从文献词汇到词系统核心词汇映射中的词汇相似度计算算法不完善。系统的完善是自动赋词标引研究工作今后的一部分内容。

5 结语

本文完成了将关系数据库模式的词系统转化成灵活易用的SKOS格式，选择文献自动标引技术方法并进行改进，完成了利用《汉语科技词系统》进行文献自动赋词标引的整体技术流程，并发布了文献自动赋词标引系统的在线系统。

文献自动赋词标引研究可以促进通过领域标准词汇进行文献资源整合。自动赋词标引能够利用词系统中的标准词汇，将代表文献的主题从自由词汇转换到标准词汇。通过词系统中已定义的标准词汇的结构和词汇间的关联，来完成异构文献的聚类和关联整合。

利用《汉语科技词系统》对文献进行中文关键词自动赋词标引研究，是《汉语科技词系统》在文献标引上的一个初步应用研究。目前由于中文标引知识库类资源的不足，中文赋词标引类的研究还不多。基于《汉语科技词系统》自动标引为中文赋词标引做了一些实验工作。接下来我们首要的工作还将继续对中文自动赋词标引的流程算法进行改进补充，以期使《汉语科技词系统》以及自动赋词标引的研究工作在文献标引及领域异构文献的资源整合中发挥更好的作用。

参考文献

[1] 中国科学技术信息研究所.汉语科技词系统:新能源汽车卷[M].北京:科学技术文献出版社,2011:1.
[2] d2rq工具页[EB/OL]. [2013-09-26]. http://d2rq.org/.
[3] MEDELYAN O. Human-competitive automatic topic indexing [D]. Hamilton: University of Waikato, 2009.
[4] 张远良,徐硕,朱礼军,乔晓东.汉语科技词系统:一种可用于科技信息资源深度内容分析的语义资源[J].图书情报工作,2011,55(4):100-105.

作者简介

闫莹莹（1981-），女，中国科学技术信息研究所，硕士。研究方向：知识组织，自动标引。E-mail: yanyy@istic.ac.cn
许德山（1979-），男，中国科学技术信息研究所，博士。研究方向：知识组织，知识表示，自动标引技术。E-mail: xuds@istic.ac.cn
张远良（1979-），男，中国科学技术信息研究所，博士。研究方向：知识组织，词系统自动构建，自然语言处理，词系统应用。E-mail: zhangyl@istic.ac.cn
李鹏（1979-），男，中国科学技术信息研究所，硕士。研究方向：智能信息处理。E-mail: lipeng_cn@istic.ac.cn

Research of Automatic Assignment Topics Indexing Using Chinese Scientific and Technical Vocabulary Systems

Yan Yingying, Xu Deshan, Zhang Yunliang, Li Peng / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: The author introduced the system structure and function of Chinese science and technology system firstly, then designed the whole idea of automatic assignment topics indexing, completed the function of automatic assignment topics indexing system, including knowledge base format conversion, algorithm and system implementation, and collected data to test the result. Finally, author analyzed the result of the automatic assignment topics indexing, summarized the characteristics and the insufficiency of automatic assignment topic indexing, and introduced the future work.

Keywords: Automatic topic indexing, Term assignment indexing, Chinese Scientific and Technical Vocabulary Systems, Indexing knowledge base, Application of vocabulary systems, D2RQ

（收稿日期：2013-10-09）