文章编号：1003-0077(2016)03-0111-07

中文专利文献术语自动识别研究

杨双龙[1]，吕学强[1]，李卓[1]，徐丽萍[2]

（1. 北京信息科技大学 网络文化与数字传播北京市重点实验室，北京 100101；
2. 北京城市系统工程研究中心，北京 100089）

摘 要：中文专利文献中含有大量领域术语，对这些术语进行自动识别是信息抽取、文本挖掘等领域的重要任务。该文提出了基于专利文献标题的术语词性规则自动生成方法以及针对候选术语排序的 TermRank 算法。该方法首先从大量的中文专利文献标题中自动生成词性规则；然后利用生成的词性规则对中文专利文献正文部分进行规则匹配获得候选术语表；再利用提出的 TermRank 排序算法对候选术语表排序，最终得到术语列表。通过在 9 725 篇中文专利文献数据上实验，证实了该方法的有效性。

关键词：术语自动识别；专利文献；信息抽取；文本挖掘

Automatic Recognition of Terms in Chinese Patent Literature

YANG Shuanglong[1], LV Xueqiang[1], LI Zhuo[1], XU Liping[2]

(1. Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing 100101, China;
2. Beijing Research Center of Urban System Engineering, Beijing 100089, China)

Abstract: Chinese patent literatures contain abundant domain-specific terms, and automatic recognition of terminology is an important task in information extraction and text mining. In this paper, we propose an approach of automatic generation of term formation rules and a novel TermRank algorithm. Firstly, we focus on generating a set of term formation rules automatically through a large number of patent titles and then applied those rules to patent texts for term candidates. Finally, the TermRank algorithm decides the final terms. Experimental results on 9725 Chinese patent literatures demonstrate the effectiveness of the proposed approach.

Keywords: automatic term recognition; patent literature; information extraction; text mining

1 引言

自动术语识别(Automatic Term Recognition, ATR)是信息抽取研究领域的重要组成部分。它是指通过无人工干预或尽量少的人工干预方法，从自由文本中自动识别出能够代表某个专业领域中一般概念的词汇串的过程。通过术语自动识别技术构建的术语库是非常重要的基础数据资源，为中文分词、本体构建、词典编撰与更新、自动标引、信息检索以及机器翻译等提供不可或缺的数据支持。此外，伴随着信息技术的高速发展，数字化信息资源与日俱增，对这些资源进行术语的自动识别对于及时把握领域最新发展状况及未来发展趋势具有十分重要的意义。

中文专利文献是重要的数字化信息资源，它们记载着各学科领域的最新发明成果，其中存在着大量的专业术语。结合对中文专利文献的观察分析与前人[1-2]的研究，发现专利文献中的术语具有如下几个明显特点：(1)专利文献中的术语嵌套现象较为常见；(2)专利文献中的术语具有较强的领域相关性，即高频率出现在某一领域的术语在另外的领域

收稿日期：2014-03-20 定稿日期：2014-05-16

所属课题：国家自然科学基金(61271304)；北京市教委科技发展计划重点项目暨北京市自然科学基金 B 类重点项目(KZ201311232037)；北京市属高等学校创新团队建设与教师职业发展计划项目(IDHT20130519)

中低频出现甚至不出现；(3)专利文献中的术语具有重复出现的特点，即术语在整个专利文献集中的多篇文献中出现；(4)专利术语长度较长，通常由2—5词构成；(5)专利术语大多是由名词或复合名词构成。以上术语的特点是对中文专利文献进行术语自动识别的重要依据。

本文针对中文专利文献中术语的特点，结合目前主流的术语自动识别方法，提出了基于专利标题的词性规则自动生成方法，利用这些规则再从专利文献中匹配出候选术语。根据得到的候选术语，提出TermRank方法对其进行排序，并确定最终术语表。

2 相关研究工作

目前，国内外研究者在术语自动识别研究领域，通常采用两种不同的研究方法。第一种为传统的规则与统计相结合的术语识别方法。在生成候选术语集的过程中，先对中文文本进行分词和词性标注处理。通过观察标注好的语料总结出构成术语的词性规则集，利用这些词性规则在语料中匹配生成候选术语集。Frantzi[3]、Dagan[4]等人通过观察总结了各自的词性规则，如表1所示。

表1 前人总结的词性规则

规则编写者	词性规则
Frantzi	Noun＋Noun (Adj|Noun)＋Noun ((Adj|Noun)＋|((Adj|Noun)＊(NounPrep)?(Adj|Noun)＊)Noun
Dagan	Noun(＋)
Fahmi	((Adj|N) ＋ |(((Adj|N)＊(N Prep)?(Adj|N)＊

依靠人工编写词性规则的方式虽然识别精度较高，但对编写者的语言学知识依赖性太大，不同人对同一个语料编写的词性规则并不一致。Yang[5]等人采用去除句子中功能词的办法，对句子进行粗切分得到候选术语集。闫兴龙[6]等人对语料中的句子进行切分，得到候选多字集合，并将其作为下一步过滤算法的输入。虽然在得到候选术语阶段这些方法不需要利用词性规则，但是在对句子进行粗切分时对外部的资源依赖性太大，外部资源的质量往往决定了得到的候选术语集的质量。索红光[7]等人将文本通过先组织成词汇链，再结合词频、区域特征等抽取关键词，该方法在召回率和准确率方面均有所提高，但是受到知识库质量以及分词准确率的很大限制。

在对候选术语集进行排序方面，国内外许多研究者提出了不同的排序算法。其中贡献最大的是由Frantzi提出的C-value/NC-value[3]算法，它们对于识别词串较长的术语取得了较好的效果。但是，C-value/NC-value对于识别长度较短的术语或者出现频率较低的术语并不太理想。因此，许多研究者提出了不同的基于C-value改进方法[8-9]，改进后的方法在一定程度上比原始C-value更具优势。徐川[10]等人通过计算候选词串间的结合强度，在中文专利文献中识别术语的平均正确率达到80.24%，但也存在一定的误识别率。杨洁[12]等人提出ATF×PDF的术语权重计算方法，该方法综合考虑了词频、词性以及词语之间语义相似性等信息，取得了一定的实验效果，但是对分词效果和外部资源依赖较大。目前，术语自动识别研究领域的主流趋势是对多种排序方法的融合[11-12]，融合后的方法具有一定的识别效果。

第二种识别术语的方法是采用近年来在信息抽取领域逐渐趋于研究热点的机器学习算法。Fethi选择浅层语言学知识作为CRF机器学习模型的特征，在医学领域语料库上进行术语自动识别研究。贾美英[13]等选择了词本身、词性、左右信息熵、互信息、TF/IDF等特征，利用CRF机器学习算法对军事情报领域进行术语自动识别研究，证明了CRF的有效性。机器学习算法虽然综合利用了较多的语言学知识和统计学参数，较之传统方法具有其独特优势，但是对训练语料的规模和质量要求较高，并且需要人工标注大量数据，语料的训练也需要花费较长的时间。

本文提出的方法属于以上第一种方法的范畴，但是所用到的语言学词性规则并不是通过人工编写，而是通过对专利标题中的术语进行统计自动生

成。此外，针对目前主流的候选术语排序算法对长度较短术语识别不理想的缺点，提出对长术语和短术语都适用的 TermRank 排序算法。

3　候选术语的自动生成

传统的术语识别方法在对文本进行分词和词性标记预处理后，研究者利用人工总结的词性规则进行候选术语的抽取。为了避免人工总结词性规则不完备，本文提出一种能够从专利文献标题中自动生成术语词性规则方法。

3.1　基于专利标题的词性规则自动生成

专利文献一般是对发明、实用新型、外观设计的记载，其标题是对整个文献的高度概括，因此往往会直接给出所要描述的对象。

观察发现，专利文献的标题中都至少包含一个正确术语。表 2 列举了几篇经 ICTCLAS[14] 分词及词性标注处理后的专利标题以及其中所包含的术语。

表 2　专利文献标题所含术语举例

序号	专利文献标题	所包含术语
1	快速/d 联接/v 压力/n 传感器/n	压力传感器
2	用于/v 减小/v 暗/a 电流/n 的/udel 图像/n 传感器/n 及其/cc 制造/vn 方法/n	暗电流、图像传感器
3	一/m 种/q 电动/b 汽车/n 的/udel 电量/n 显示器/n 装置/n	电动汽车、电量显示器
4	一/m 种/q 音乐/n 催眠/v 保健/n 枕/v	音乐催眠保健枕
5	足/a 底/f 软组织/n 检测/vn 系统/n	足底软组织
6	可/v 检测/vn 道岔/n 压力/n 的/udel 立式/b 杆/ng 架/v	道岔压力、立式杆架

根据中文专利标题的以上特点，将标题形式化地表示成如图 1 所示。

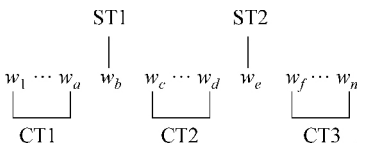

图 1　中文专利标题形式化表示

其中，$w_i(i = 1,2,\cdots n)$ 表示专利标题被 ICTCLAS 切分出的词，$w_1 \cdots w_a$，$w_c \cdots w_d$ 以及 $w_f \cdots w_n$ 为标题中的术语，分别表示为 CT1CT2CT3；w_b 和 w_e 是不属于任何术语构成部分的词，本文称其为停用词，其构建方法在 3.2 节介绍。

以停用词 ST1ST2 为分隔符，将子串 CT1CT2CT3 的词性规则提取出，即可作为下一步生成候选术语的词性规则。例如，专利标题："一/m 种/q 电动/b 汽车/n 的/udel 电量/n 显示器/n 装置/n"中包含术语："电动/b 汽车/n"、"电量/n 显示器/n"。提取出它们的词性规则："b＋n"、"n＋n"，并将它们添加至词性规则集中，作为下一步生成候选术语的词性规则。

3.2　停用词表构建

3.1 节提到的停用词是从专利标题中自动生成词性规则的重要资源。本文选择手工构建停用词表，而不是直接采用现成的通用停用词表，是因为现成的通用停用词表内的某些停用词在专利文献中有可能是术语的组成部分。例如，"排/v"在通用停用词表中存在，但在"全自动/b 排/v 纸/n 机/ng"中，它又是构成术语的一部分，因此不能将其加入停用词表。类似"排/v"这类在通用停用词表中存在，但在中文专利文献中又是构成术语的部分的词在语料中大量存在。

本文构建的停用词表中的停用词来源于以下三个方法。方法一：对专利标题分词后进行词频统计，将出现频率高于 20 的停用词加入停用词表；方法二：将明显不会出现在术语中的词性加入停用词表，如/vyou、/m、/wkz、/ulr 等词性；方法三：应用方法一和方法二步骤生成的停用词表对标题进行过滤后，对剩余词串进行人工观察，若再发现新的停用词，也将其加入到停用词表中。

3.3　候选术语的生成

对生成的词性规则按照所含词性的个数进行分类。由于本文中只识别 2—5 词术语，故将词性规则分为四类：2—5 词词性规则。自动生成的词性规则数量较多，无法将它们全部应用到文献中进行术

语匹配，因此需要有选择地从中挑选出部分词性规则。本文对每一类词性规则按照出现频率降序排列，并只取 Top 5 条规则应用到中文专利文献的正文部分[①]进行词性匹配，即可生成候选术语集合。

抽取出的候选术语也按照所包含词的个数进行分类，即分类为：2—5 词候选术语。这样分类的目的是为了让每一类长度的术语都单独构成一张候选术语表，在对其利用第四节中的排序算法进行排序时能够不受其他长度的术语的影响，从而排序结果更公平。

4 TermRank 排序算法

对候选术语排序的目的是为了确定最终术语表。一个好的排序算法能够将候选术语列表中分散的正确或错误的术语重新排序，使正确的术语的权重增大，排名位置尽量靠前，反之亦然。

本文提出的 TermRank 算法是受 Page 和 Brin 提出的 PageRank[15] 算法思想启发。PageRank 在 Web 信息检索领域应用广泛且效果显著。PageRank 的核心思想是：若有多个网页链向某一网页，则表明该网页质量较高，故其 PageRank 值也高；而某一网页的 PageRank 值被其外链数平均分配给它所链向的网页。

统计发现，中文专利文献中也存在类似情况：若某候选术语来自多篇专利文献，则该候选术语是真正术语的可能性也越大。例如，“编程/vn 控制器/n”在 163 篇专利文献中出现，“液晶/n 显示器/n”在 331 篇专利文献中出现。如此高文档频率出现表明它们并非偶然出现。

但是也存在并不是正确术语的候选术语在多篇文献中出现。例如，“传感器/n 包括/v”出现在 472 篇文献中，但它并不是一个正确术语。通过对此类非术语的候选术语分析，发现其中通常包含一个或多个停用词。因此，当发现候选术语中存在停用词时，应该降低其排序权重。基于以上统计和分析，提出针对中文专利候选术语的 TermRank 排序算法，如式(1)所示。

$$TR(T_i)=\sum_{j=1}^{M}\frac{TF_{T_i}(d_j)}{C(d_j)}-|T_i|\times count(T_i) \tag{1}$$

其中，T_i 为候选术语，$TR(T_i)$ 为候选术语 T_i 的 TermRank 值；M 为包含候选术语 T_i 的专利文献数量；$TF_{T_i}(d_j)$ 为包含候选术语 T_i 的专利文献 d_j 中 T_i 的词频；$C(d_j)$ 为专利文献 d_j 中抽取出的候选术语数量。$|T_i|$ 为候选术语 T_i 的长度(词粒度)，$count(T_i)$ 为候选术语 T_i 中包含的停用词数量。

通过分析式(1)发现，第一项和第二项并不一定在同一数量级上，当 M 值较大或者较小时，对候选术语的 TermRank 值影响并不大，因此需要对它们分别进行归一化处理。本文选择线性变换归一化方法，对其中第一、第二项归一化，公式分别如式(2)和式(3)所示。

$$\left\{\sum_{j=1}^{M}\frac{TF_{T_i}(d_j)}{C(d_j)}-\min\sum_{j=1}^{M}\frac{TF_{T_i}(d_j)}{C(d_j)}\right\}/\left\{\max\sum_{j=1}^{M}\frac{TF_{T_i}(d_j)}{C(d_j)}-\min\sum_{j=1}^{M}\frac{TF_{T_i}(d_j)}{C(d_j)}\right\} \tag{2}$$

$$\{|T_i|\times count(T_i)-\min|T_i|\times count(T_i)\}/\{\max|T_i|\times count(T_i)-\min|T_i|\times count(T_i)\} \tag{3}$$

由式(1)可知，候选术语 T_i 的 TermRank 不仅被出现在多篇专利文献中这一现象增强，而且还被它在该专利文献中的词频增强。即，若某候选术语在某篇专利文献中出现频率越高，则该候选术语越有可能是正确术语。候选术语 T_i 的 TermRank 被其中出现的停用词所抑制，且若其中出现的停用词数量越多，则抑制作用越明显。对候选术语列表中的每一个候选术语都按照以上公式计算其 TermRank 值，经排序后，取 Top-N 条作为最终术语表。

5 实验及结果分析

5.1 实验设计

本文实验数据由国内某专利公司提供，共有 9 725 篇专利文献。去除其中的表格和图片，保存为纯文本后的语料大小为 123M。采用 ICTCLAS 对专利文献进行分词及词性标注处理。词性标注采用中科院计算所二级词性标注集，可参见《ICTPOS3.0

① 中文专利文献通常由以下几部分组成：专利标题、技术领域、背景技术、发明内容、附图说明、具体实施方式。本文认为除“专利标题”外，其余部分皆属于专利文献正文。

汉语词性标记集》[①]。

采用 3.2 节介绍的构建停用词表方法，最后构建的停用词表中包含停用词共 246 个。表 3 列出了其中部分停用词。

表 3　人工构建的停用词表中部分停用词

序号	停用词	序号	停用词
1	所	6	这种
2	之	7	它们
3	或	8	采用
4	它	9	确定
5	了	10	用于

5.2　评价方法

采用人工方式对实验结果进行判断。为避免人的主观性和领域知识的局限性，对于明显正确或错误的术语直接标记相应标记，而对于很难辨别正确性的候选术语则利用 Google 搜索引擎进行判断。只要符合以下情况的任何一条，则将该候选术语标记为正确术语，否则标记为错误术语：1)在 Wikipedia、百度百科、互动百科等知识网站存在对应词条；2)在专利检索系统存在此词条；3)Google 搜索引擎未对候选术语中任何成分进行过滤或打乱次序等处理。

由于实验结果集太大，难以对整个排序后的列表进行整体评估，因此采用 P@N 评价方法，即判断最终术语表中前 N 条的准确率(Precision)，其计算公式如式(4)所示。

$$\text{正确率} = \frac{\#\ \text{前}\ N\ \text{条候选术语中正确的术语}}{\#\ \text{前}\ N\ \text{条候选术语}} \times 100\% \tag{4}$$

5.3　实验结果及分析

利用 3.1 节所述自动生成词性规则方法，从专利文献标题中共生成 2 832 条无重复词性规则。表 4 列出按照频率排序后的 Top5 条。该统计结果从实验数据上验证了大部分术语是由名词或复合名词构成的特点。

表 4　自动生成词性规则 Top 5 条举例

序号	词性规则	出现频次
1	N＋N	1 410
2	N＋VN	647
3	VN＋N	435
4	N＋N＋N	375
5	N＋V	292

续表

表 5 是对词性规则按照不同长度分类后，其出现频次所占总频次(2 832)百分比的统计信息。其中长度为 4 和 5 的词性规则共占 71.5%，验证了专利文献中术语长度偏长的特点。

表 5　不同长度的词性规则比例

长度(词粒度)	出现频次	所占百分比/%
2	196	6.9
3	609	21.5
4	995	35.1
5	1032	36.4

这种通过从专利文献的标题中自动总结词性规则的方法相对于传统的词性规则生成方法，具有以下两方面的优势：1)大幅度减少冗余信息：相对于从专利正文总结词性规则，从标题中总结词性将大幅度减少冗余的词性规则；2)对分词和词性标注工具的精度依赖减小：不管标题中的术语被正确地或错误地分词和词频标注，它的词性规则模式都将被加入词性规则集中。在抽取候选术语时，若候选术语被错误切分和标注，也将被抽取出。

由于自动生成的词性规则较多，将所有规则都应用到专利文献中抽取候选术语并不必要。因此对于每一类长度的词性规则，按照出现频次的高低，只取 Top 5 条。表 6 是不同长度词性规则的 Top 5 条。

应用表 6 中列出的词性规则，再对专利文献正文进行抽取。抽取出 2 词候选术语 493 286 条；3 词候选术语 152 274 条；4 词候选术语 31 809 条；5 词候选术语 3 966 条。表 7 是抽取出的部分候选术语及对应匹配的词性规则。

利用词性规则抽取出的候选术语质量较高，但也存在部分噪音。例如，候选术语“结合/v 附图/n”虽然匹配“V＋N”词性规则，但本身并不是真正术语；候选术语“位移/v 传感器/n”中的“位移”的词性应该为 n，“语音/n 式微/v 型/k 乳腺/n 检查仪/n”正确的分词和词性标注应该为“语音/n 式/k 微型/

① http://ictclas.org/news_ictclas_files.html

a 乳腺/n 检查仪/n”。虽然这些词串被错误地分词或词性标注，但本身仍然为术语，且被正确地识别出来，这正是本文所采用的自动生成词性规则的优势之处，即对分词和词性标注的精度依赖性较小。

表6 不同长度词性规则 Top5 条

2词词性规则 TOP@5		3词词性规则 TOP@5		4词词性规则 TOP@5		5词词性规则 TOP@5	
规则	频次 f	规则	频次 f	规则	频次 f	规则	频次 f
N+N	1410	N+N+N	375	N+N+N+N	70	N+N+N+N+N	19
N+VN	647	N+N+VN	233	N+N+NV	58	N+N+N+N+VN	18
VN+N	435	N+N+N	225	NN+VN+N	57	B+N+N+N+VN	12
N+V	292	VN+N+N	123	B+N+N+N	39	N+N+N+VN+N	11
V+N	275	N+V+N	123	N+VN+N+N	31	N+V+K+N+N	9

表7 部分候选术语及匹配的词性规则

候选术语	词长	匹配词性规则
温度/n 传感器/n	2	N+N
位移/v 传感器/n	2	V+N
数字/n 信号/n 处理器/n	3	N+N+N
结合/v 附图/n	2	V+N
水温/n 水位/n 感应/n 电路/n	4	N+N+N+N
低压/n 电磁/n 换向/vn 阀/n	4	N+N+VN+N
牛/n 小肠/n 碱性/n 磷酸/n 酶/n	5	N+N+N+N+N
语音/n 式微/v 型/k 乳腺/n 检查仪/n	5	N+V+K+N+N

将候选术语按照不同词长划分到不同候选术语表中，由于本文只识别长度为2—5词术语，因此得到四张候选术语表。对候选术语的排序是在每一张候选术语表上单独进行，是为了避免由于某类长度的候选术语识别较多从而对整体排序造成不公正的现象出现。为了验证本文提出的 TermRank 方法的有效性，选取 TF 和 C-Value 作为对比方法。表8为对最终候选术语排序结果采用 P@N 评价方法的统计信息，其中 N 依次取值 100，200，400，800，1 000。

表8 对候选术语排序结果的 P@N 评价

长度	排序算法	P@100	P@200	P@400	P@800	P@1000
2词	TF	0.590	0.615	0.598	0.596	0.591
	C-Value	0.690	0.650	0.680	0.691	0.685
	TermRank	0.890	0.845	0.827	0.801	0.795
3词	TF	0.780	0.765	0.763	0.753	0.750
	C-Value	0.760	0.745	0.765	0.751	0.746
	TermRank	0.930	0.895	0.887	0.882	0.879
4词	TF	0.680	0.670	0.663	0.670	0.669
	C-Value	0.690	0.675	0.663	0.670	0.659
	TermRank	0.860	0.820	0.805	0.796	0.778

续表

长度	排序算法	P@100	P@200	P@400	P@800	P@1000
5 词	TF	0.680	0.660	0.655	0.635	0.629
	C-Value	0.720	0.705	0.695	0.681	0.677
	TermRank	0.850	0.825	0.805	0.797	0.775

由表 8 中的实验结果可以看出，本文提出的 TermRank 方法对不同长度的候选术语排序效果都显著优于其他两种排序方法。在 P@1000 上，Term-Rank方法对 3 词长度术语的识别正确率均达到 80%以上。从 P@100～P@1000 上的正确率逐渐递减的规律也印证了 TermRank 具有较好地将术语和非术语区分开的能力。

6　结论与展望

术语自动识别研究是信息抽取和文本挖掘等领域的重要研究课题。本文首先利用统计学方法从专利标题中自动学习出构成术语的词性规则，解决了人工总结术语词性规则的不足。对候选术语集的排序算法的优劣反应在最终识别出的术语的质量上，本文提出的 TermRank 排序方法综合考虑了专利文献中语言学和统计学特征，能够较好的区分术语和非术语，在 P@1000 级别上的准确率验证了其较高的可靠性。文中对每一类长度的词性规则模板依据统计频率选取 Top5 条的方式，存在一定的局限性。因此，在下一步的研究工作中，需要设计出一种更好的选取词性模板策略，进一步提高自动识别术语的效果。

致谢

感谢中国科学院计算技术研究所提供的 ICTCLAS 分词及词性标注工具，让本文实验得以顺利完成。

参考文献

[1] 游宏梁，张巍，沈钧毅，等. 一种基于加权投票的术语自动识别方法[J]. 中文信息学报，2011，25(3)：9-16.

[2] 岳金媛，徐金安，张玉洁等. 面向专利文献的汉语分词技术研究[J]. 北京大学学报(自然科学版)，2013，49(1)：159-164.

[3] Frantzi K，Ananiadou S，Mima H. Automatic recognition of multi-word terms：the C-value/NC-value method[J]. International Journal on Digital Libraries，2000，3(2)：115-130.

[4] Dagan I，Church K. Termight：Identifying and translating technical terminology[C]//Proceedings of the fourth conference on Applied natural language processing. Association for Computational Linguistics，1994：34-40.

[5] Yang Y，Lu Q，Zhao T. Chinese term extraction using minimal resources[C]//Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics，2008：1033-1040.

[6] 闫兴龙，刘奕群，方奇等. 基于网络资源与用户行为信息的领域术语提取[J]. 软件学报，2013，24(9)：2089-2100.

[7] 索红光，刘玉树，曹淑英. 一种基于词汇链的关键词抽取方法[J]. 中文信息学报，2006，20(6)：25-30.

[8] 李超，王会珍，朱慕华，等. 基于领域类别信息 C-value 的多词串自动抽取[J]. 中文信息学报，2010，24(1)：94-98.

[9] 韩红旗，朱东华，汪雪锋. 专利技术术语的抽取方法[J]. 情报学报，2011，30(12)：1280-1285.

[10] 徐川，施水才，房祥等. 中文专利文献术语抽取[J]. 计算机工程与设计，2013，34(6)：2175-2179.

[11] 杨洁，季铎，蔡东风，等. 基于联合权重的多文档关键词抽取技术[J]. 中文信息学报，2008，22(6)：75-79.

[12] 梁颖红，张文静，周德富. 基于混合策略的高精度长术语自动抽取[J]. 中文信息学报，2009，23(6)：26-30.

[13] 贾美英，杨炳儒，郑德权，等. 采用 CRF 技术的军事情报术语自动抽取研究[J]. 计算机工程与应用，2009，45(32)：126-129.

[14] Zhang H P，Yu H K，Xiong D Y，et al. HHMM-based Chinese lexical analyzer ICTCLAS[C]//Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17. Association for Computational Linguistics，2003：184-187.

[15] Brin S，Page L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer networks and ISDN systems，1998，30(1)：107-117.

(下转第 124 页)

extraction from free texts[C]//Proceedings of the Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on. IEEE, 2012: 1290-1293.

[12] Kim S, Yoon J. Experimental Study on a Two Phase Method for Biomedical Named Entity Recognition[J]. IEICE Transactions on Information and Systems, 2007, E90-D(7): 1103-1110.

[13] Chan S K, Lam W, Yu X F. A cascaded approach to biomedical named entity recognition using a unified model[C]//Proceedings of the 7th IEEE International Conference on Data Mining, Omaha, Nebraska, USA, 2007: 93-102.

[14] Gu B, Popowich F, Dahl V. Recognizing biomedical named entities in Chinese research abstracts[M]. Advances in Artificial Intelligence. Springer Berlin Heidelberg, 2008: 114-125.

[15] 蒋锦文,于鹏. 浅谈中医学术语的特点和研究方法[J]. 天津中医学院学报,2000,3: 023.

孙水华(1962—),博士研究生,副教授,主要研究领域为自然语言处理与机器翻译。
E-mail: sunsh@mail.dlut.edu.cn

黄德根(1965—),博士,教授,博士生导师,主要研究领域为自然语言处理与机器翻译。
E-mail: huangdg@dlut.edu.cn

牛萍(1988—),硕士研究生,主要研究领域为自然语言处理与机器翻译。
E-mail: 425204127@qq.com

(上接第117页)

杨双龙(1989—),硕士研究生,主要研究领域为中文信息处理、网络数据挖掘。
E-mail: yslgoodboy@gmail.com

吕学强(1970—),博士,教授,主要研究领域为中文信息处理、多媒体信息处理。
E-mial: lxq@bistu.edu.cn

李卓(1983—),博士,讲师,主要研究领域为分布式计算,社交网络。
E-mial: lizhuo@bistu.edu.cn