基于字序列标注的中文关键词抽取研究

王 昊 邓三鸿 苏新宁
(南京大学信息管理系 南京 210093)

【摘要】以某大学图书馆的所有馆藏书目为研究对象,在对图书关键词标引信息进行分析的基础上,总结中文关键词的基本特点及其抽取规律,构建一个基于字序列标注的中文关键词抽取模型,提出中文关键词抽取的基础思路和实现方案,并通过实验论证模型的合理性、正确性和实用性,认为字序列标注方法优于词序列标注,基本上可以解决不分词情况下的中文关键词抽取问题。

【关键词】序列标注 条件随机场 关键词抽取 机器学习 字序列 词序列
【分类号】TP391.1

Research on Chinese Keywords Extraction Based on Characters Sequence Annotation

Wang Hao Deng Sanhong Su Xinning
(*Department of Information Management, Nanjing University, Nanjing 210093, China*)

【**Abstract**】Based on the whole Chinese booklist of a certain university library as well as the analysis of its book indexing information, the paper summarizes the features and extracting laws of Chinese keywords, and establishes a Chinese keywords extraction model based on characters sequence annotation, which proposes the basic idea and implementation scheme for extracting keywords. It verifies the feasibility, rationality and practicality of the model by large - scale experiments, and basically solves the problems of Chinese keywords extraction without executing words segmentation, which shows that characters sequence annotation is better than words sequence annotation.

【**Keywords**】Sequence annotation Conditional random fields Keywords extraction Machine learning Characters sequence Words sequence

1 引 言

关键词抽取是指利用计算机从文本中自动提取出能够代表该文本主题的词汇或短语集合以实现文本表示的过程[1]。它是实现海量文本内容检索的前提,也是实现如文本标引、自动分类聚类、自动摘要、个性化推荐等工作的核心技术[2-4]。中文语法的特殊性及组词的复杂性加大了中文关键词抽取的难度。

目前中文关键词抽取主要有两种方法:基于分词的统计方法,先对文本进行精确分词,在排除非用词后建立候选关键词集合,再采用统计算法如词频、TF - IDF 值、ATF × PDF 值等,结合词语的词性、位置、形态等特征对候选关键词进行权重计算,筛选关键词[5-9];基于词序列的语言学方法,首先对图书文本进行粗分词,使其转化为词汇序列,然后根据词汇的上下文语言学特征确定相邻词汇之间的语义关系,判断是否可将相邻词汇合并作为关键词[10],或对词汇的上下文语言学特征进行机器学习,再用训练后形成的学习模型判断词汇的角色,进而根据关键

收稿日期:2011 - 10 - 08
收修改稿日期:2011 - 11 - 13

词角色模板将相应词序列合并为关键词[11-14]。

上述方法均存在如下明显缺陷：

(1)现有效果较好的中文分词系统均倾向于将文本切分为长度较小的词汇，与关键词一般为长度较大的名词性领域术语相矛盾，因此需要对分词后获得的词汇进行再处理；

(2)机器学习需要大量标注数据作为训练样本建立标注模型，这需要大量的人力资源及领域知识[5]；

(3)中文组词具有很强的灵活性，使得词汇数量巨大，特征丰富而不易学习，而且将关键词看作是词汇组合使得词汇角色非常复杂，例如关键词的组成部分可能被切分到其他非关键词中。

基于机器学习的方法在训练样本足够大、覆盖范围足够广的情况下，针对同领域文本的关键词抽取具有很好的准确性和召回率。因此，本文试图利用现有的图书标引数据，对词序列标注方法进行改进，将关键词看作是汉字的组合，设计关键词的字角色空间，构建基于字角色标注的书目关键词抽取模型，对人工标引关键词的特征和规律进行机器学习，进而利用生成的标注模型对未标引书目实现关键词抽取，以解决词序列标注稳定性差、词语转变为关键词困难等问题。笔者通过实验对比论证了字序列标注模型的正确性和合理性，为模型的进一步优化、应用提供了事实依据。

2 标引数据的统计和分析

采用某大学图书馆的馆藏书目数据作为研究对象，解决了标引数据难获取的问题，而且获得的标注模型具有很强的实用价值；关键词词组合的思想具有一定应用价值，但是存在切分不准确、角色多样性以及成分不稳定等缺点，需要进一步完善。为了探索关键词抽取的思路和方法，有必要从总体上了解图书标引数据的分布及其特点。

至2010年6月底，该馆共有馆藏书目187 213种，其中具备标引词的有150 850种。在关键词标引中，标引词能够在书目题名或内容摘要中抽取的有65 482种，涉及关键词6 511个；题名标引的有52 279种，涉及关键词5 746个；摘要标引的有42 754种，涉及关键词5 167个；标引词同时出现在题名和摘要中的则有29 530种，涉及关键词4 159个。

2.1 书目关键词个数的统计和分析

在关键词标引数据中，图书关键词个数的分布情况如表1所示：

表1 书目关键词个数的分布情况

关键词个数 \ 关键词位置	题名或摘要	题名关键词	摘要关键词	题名摘要关键词
5	1	1	0	0
4	10	6	4	1
3	84	58	25	6
2	881	800	151	75
1	64 506	51 414	42 575	29 448
书目合计	65 482	52 279	42 755	29 530

不难发现，书目关键词的个数被限制在了1-5个词语之间，其中绝大部分书目均只有一个关键词，部分书目有两个，两者合计占总书目的99.8%以上。书目关键词个数相对较少的特点决定了机器自动抽取书目关键词的方法是合适的，而且将具有相对较高的准确率和实用性。

2.2 图书类目的分布情况分析

在书目的关键词标引数据中，排除错误分类现象(主要包括没有分类和大类目丢失)，图书类目的分布情况如图1所示：

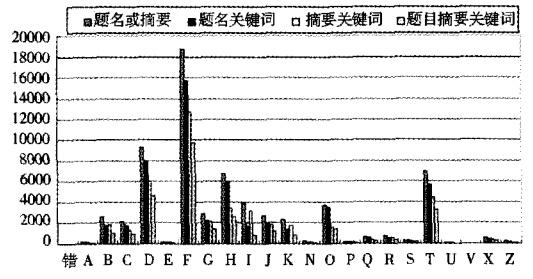


图1 图书关键词的类目分布情况

不难发现，F类(经济学)图书占据了绝对领先地位，其次是D(政治理论)、H(语言学)和T(工业技术理论等)等类，U(运输)和V(航空航天)类最少；从总体上来看，类目分布极不均衡，其中B、C、D、F、G、H、I、J、K、O、T等类的书目数量远远超过其他11个类目，占总数的94%以上。关键词一般具有很强的类目特征，在书目分类研究中通常占据最大的权重比重。而图书分类的不均衡现象为书目特征(即关键词)的抽取增加了难度，特别是在采用机器学习方法时，会导致学习不充分、标注偏差等问题，最终致使模型的标引准确率下降，可以考虑将部分稀缺类目进行合并，以平衡类目数据间的差异。

2.3 关键词组成成分分析

传统的图书自动标引方法对书目文本进行分词，

进而统计词频，选择有意义的高频词作为关键词。随着标引工作的逐渐深入，研究人员开始认识到：关键词通常是一个短语或词组，通过分词直接获得关键词存在较大缺陷。借助中国科学院计算技术研究所研制的ICTCLAS分词系统[15]，对题名和摘要中的6 511个关键词进行分词统计，结果如图2所示：

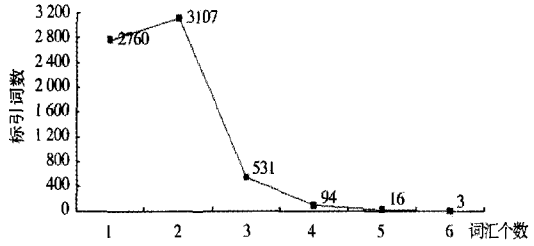


图2 由相应词汇个数组成的关键词数量统计

在所有关键词中，未被切分的有2 760个，约占总数的42.39%；大部分关键词可以被切分成2－6个词汇，其中超过半数被切分成了2个词汇；被切分成词汇的个数越多，此类关键词数量越少。因此，笔者认为书目关键词大多由若干个词汇构成，对书目内容分词后统计词频确定关键词的方法并不合适，对词汇进行角色标注进而构造关键词的抽取方法值得探讨。

2.4 词组合和字组合

任一关键词均可看作是词或字等片段的组合，若能够在语言片段中识别出连续的标记符号，那么其所对应的语言片段组合即为关键词。对65 482本图书的题名分别进行词切分和字切分，前者采用ICTCLAS系统，对分词结果不做任何人工修正；后者则将文本切分为单个汉字（包括连续符号）。切分结果如图3所示：

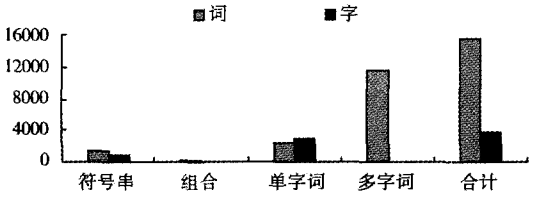


图3 书目题名的词切分和字切分对比情况

词切分共获得包括连续符号串、符号汉字组合、单字词和多字词等在内的15 442个不同词语，其中多字词最多，占75.08%；字切分则获得连续符号串和单字共3 725个，远少于切分后的词语种类，笔者认为：

（1）若以汉语片段作为语言学特征，那么字切分后获得的特征将比词切分少，机器学习的复杂性会降低；

（2）汉语中字与字之间具有灵活的组合性，在不同上下文语境下，同一个字可能会出现不同的前后组合，歧义现象将极大增加组成关键词成分的角色数量，给词语的角色标注带来极大不确定性和不稳定性；而字是汉语片段的最小单位，不会出现组合的多样性和分词的错误性，将降低角色空间的复杂性，并相对地增加字角色标注的准确性和稳定性。

3 中文关键词抽取模型的分析和设计

根据中文书目关键词标引的特点和规律，构建了一个基于字角色序列标注的中文书目关键词抽取模型，基本思路如下：

（1）建立一个角色空间，用于标识汉字（包括符号串），使得每个汉字都对应一个符号角色；将图书内容转化为字观察序列作为第一种语言特征，并扩展观察序列以强化汉字的上下文语境规律；

（2）根据角色空间，将汉字映射成角色符号作为标注序列，观察序列和标注序列一起构成了学习样本；采用机器学习算法（如条件随机场等）对学习样本的序列规律和标注规律进行学习，形成学习模型；

（3）将学习模型作用于仅由观察序列构成的测试样本进行序列标注，自动获得角色序列，则符合关键词角色模板的相应的字序列组合即为关键词。

整个过程如图4所示：

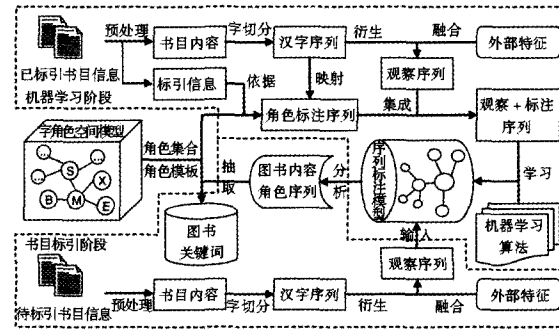


图4 基于字角色标注的中文书目关键词抽取模型

定义1：集合Ω＝{R，P}被定义为字角色空间，其中R为字角色集合，用于机器学习阶段完成汉字标注；P为关键词角色组合模板集合，用于书目标引阶段完成关键词的抽取。

定义2：集合R＝{B，M，E，S，X，F}，其中单字

关键词被标注为S,多字关键词则根据其组成汉字在词中位置分别标注为B、M、E;非关键词中的汉字则标注为X;在中文书目中出现的符号串标注为F。

根据R,可将任意汉字转化为角色,完成字空间到角色空间的映射,极大降低了序列的复杂度。需要说明的是,可能出现多个关键词共享语言片段,那么这些共享的语言片段就会有多种角色。例如在题名为"可持续发展战略"的书目中,存在两个关键词"可持续发展"和"发展战略",那么"发"和"展"就均存在两种角色,分别为"M"、"B"和"E"、"M"。本文把这种可能有多种角色的汉字用符号组合进行标记,例如把上例中"发"和"展"分别标记为"MB"和"EM"。

定义3:集合 $P=\{S, BE, BM_nE \mid n=1, 2, \cdots, N\}$。其中,S表示单汉字关键词的角色组合模板,BE为两个汉字关键词的角色组合模板,BM_nE则为多汉字关键词模板。P中模板所对应的连续汉字序列组合即为关键词。

定义4:字序列[16] $B=\{B_1, B_2, B_3, \cdots, B_n \mid n>0\}$,其中$B_n$为汉字或符号串。字序列用于描述语言片段的上下文语境特征。

例如题名"中国增值税转型可行性实证研究"对应的字序列为{中,国,增,值,税,转,型,可,行,性,实,证,研,究}。字序列是最基本的观察序列,正是由于汉字的自身特点及其固定排列,使得序列中的每个汉字都表现出一定的角色特征。

定义5:标注序列 $L=\{L_1, L_2, L_3, \cdots, L_n \mid n>0$ 且 $L_n \in R\}$,其中L_n为汉字所对应的角色。

例如题名"中国增值税转型可行性实证研究"对应的标注序列即为{X,X,B,M,E,X,X,X,X,X,X,X,X}。标注序列仅出现在训练语料中,即让机器学习某汉字被标注为指定角色的上下文特征,并将其作为一种知识(模型)存储。

定义6:特征模板 $TMPT=\{B_n(n=-2,-1,0,1,2), B_{n-1}B_n(n=-1,0,1,2), B_{n-2}B_n(n=0,1,2), B_{n-2}B_{n-1}B_n(n=0,1,2), L_{-1}L_0 \mid B_n \in B$ 且 $L_n \in R\}$,其中B_n为一元模板,描述汉字本身的特征;$B_{n-1}B_n$和$B_{n-2}B_n$为二元模板,描述前后汉字之间和字与前前字之间的关系特征;$B_{n-2}B_{n-1}B_n$为三元模板,描述字与前后汉字之间的三元关系特征;$L_{-1}L_0$描述前字的标注角色对后字标注的影响;n极大与极小值之间的间隔被称为字长窗口[16],用于描述上下文约束距离,本文采用5字长窗口。

如图4所示,笔者将图书关键词抽取过程分为两个阶段:先学习已人工标引的书目字序列上下文特征,再分析出未标引书目字序列对应的角色序列,进而抽取关键词。所谓机器学习,就是根据对已知情况状态及其可能原因的学习,来判断未知情况的可能状态的方法。目前常用的机器学习方法可分为两类:

(1)根据对象自身特征判断对象状态,也称为分类,常用的算法包括人工神经网络、支持向量机、朴素贝叶斯等,适用于对象特征较多而特征值具有较强区分度的情况;

(2)根据对象自身及其上下文环境判断对象状态,不仅需要将对象用特征向量描述,而且还需要设定当前对象和前后对象之间存在的语义关系类型,也称为序列标注,常用的算法包括隐马尔科夫模型、最大熵模型、条件随机场(CRFs)等,适用于上下文语境对当前对象具有较强影响的情况。

一个汉字的标注角色不仅由其自身特征决定,而且与其所在的上下文环境,例如前后字特征、前字的标注角色等都有关系。因此在本文构建的关键词抽取模型中采用CRFs序列标注算法,并以开源软件CRF++0.51[17]作为运行平台。

4 面向题名的书目关键词抽取实验分析

题名作为图书内容的高度浓缩,能够在很大程度上反映图书主要内容。在本文使用的图书标引数据中,共有关键词156 772个,其中53 217个来自52 279本图书的题名,约占总数1/3。可见从题名中抽取关键词是合理的,也是目前常见的人工标引方式之一。以题名关键词标引数据作为实验对象,以正确率P、召回率R和F1值作为性能评价,从训练样本规模以及标注方法等方面来论证基于字序列标注抽取关键词模型的正确性和合理性。具体公式如下:

$$P=\frac{RC}{C} \quad (1)$$

其中:C为关键词识别数,RC为正确识别数。

$$R=\frac{RC}{N} \quad (2)$$

其中:RC为正确识别数,N表示人工标引关键词的个数。

$$F1 = \frac{2PR}{P+R} \quad (3)$$

4.1 基于不同训练样本规模的关键词抽取结果比较

一般认为,机器学习方法在训练样本越充分的情况下,获得学习模型的识别效果越佳。为此,笔者分别选择 27 000,32 000,37 000,42 000,47 000 本图书的标引信息作为训练数据,并以其他 5 279 种图书(含有关键词 5 279 个)作为测试数据,来探讨随训练样本规模扩大关键词抽取效果的变化情况。结果如图 5 所示:

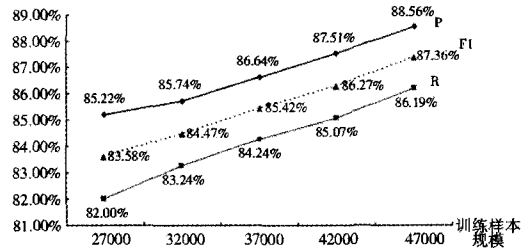


图 5 基于不同训练样本规模的关键词抽取结果比较

从图 5 可以看出:

(1)以单字序列作为观察对象,即仅考虑汉字自身以及上下文语境特征,在包含 6 种单字角色{B, M, E, S, X, F}和 4 种组合角色{EM, EB, MM, MB}的角色空间约束下,关键词抽取的 P、R 和 F1 值最高分别达到了 88.56%、86.19% 和 87.36%,可见基于字序列标注的书目关键词抽取模型具有一定的实用价值。

(2)随着训练样本规模的逐渐扩大,关键词抽取的效果明显提高,在书目样本量为 27 000 时,F1 值仅为 83.58%,当将样本量扩大到 47 000 后,F1 值增加了近 4 个百分点,进一步验证了对于基于机器学习的方法,在学习足够充分的情况下,能够发挥其最大的优势。可以推测:在本文构建的实验环境中,如果进一步加大训练样本,可以使关键词抽取的效果达到最佳。

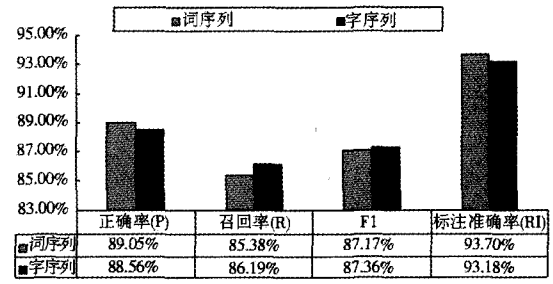
(3)当训练样本量达到 47 000 时,从测试数据中可以抽取 5 138 个关键词,其中正确的为 4 550 个,与完全抽取还有相当一段差距,除了可以通过扩充训练样本规模的方式加强机器学习之外,还可以增加学习样本的特征来提高样本区分度,包括扩展观察序列和增加上下文字长窗口等。

(4)本文将关键词抽取的应用限定在书目标引中,而目前在各级图书馆中存在大量可利用的关键词人工标引数据,能够保证充分的机器学习,因此在书目关键词抽取中引入序列标注方法是可行的、合理的。

4.2 基于词序列和字序列标注的关键词抽取结果比较

中文分词的不准确以及汉语组词的多变性,使得将文本作为词序列组合可能会带来词角色的复杂性和特征学习的不稳定性,为此本文提出了修正思路,即使用字序列来代替词序列,以提高关键词识别的综合效果,并对词序列和字序列标注的关键词抽取方法进行了全方位的比较。

为了具有可比性,在两次实验中采用了相同的训练和测试样本,其中训练样本为 47 000 种书目中包含的 47 938 个关键词,测试样本为 5 279 种图书含有人工标引关键词 5 279 个;均以字或词本身作为观察序列;字序列标注中采用了 10 角色(包括 6 种单角色和 4 种组合角色)空间,而在词序列标注中,为了能够从粗分词中进一步分离出关键词,笔者采用了 35 种角色(包括 5 种单角色和 30 种组合角色)对词汇进行区分。两次实验的结果对比如图 6 所示:



	正确率(P)	召回率(R)	F1	标注准确率(RI)
词序列	89.05%	85.38%	87.17%	93.70%
字序列	88.56%	86.19%	87.36%	93.18%

图 6 基于词序列和字序列标注的关键词抽取结果比较

从图 6 可以发现:

(1)两种方法的关键词抽取效果均较佳,F1 值均超过了 85%,可见基于序列标注实现关键词的抽取具有一定的合理性和实用性;

(2)基于字序列的关键词抽取方法在 F1 值上比词序列高出了 0.19 个百分点,可见前者的抽取效果比后者更佳;

(3)字序列方法优于词序列的主要原因是高召回率(高出 0.81 个百分点),即前者比后者能够识别更多的正确的关键词;

(4)词序列在识别正确率 P 和标注正确率 RI 两项上均优于字序列,一方面是由于词序列方法总体上识别的关键词数较少,使得其 P 值较高,这可能是由于其具有较多特征;另一方面由于相同文本的词总数远远

小于字总数，因为基数较小而使得 RI 值偏高。

为了进一步对比两种方法，图 7 列出了其他实验参数，包括关键词识别数、正确识别数、CRFs 算法特征数以及训练时间等。

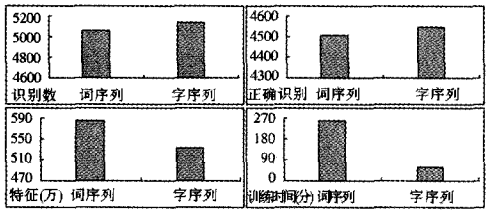


图 7　基于词序列和字序列标注的相关实验参数比较

在同等实验环境下，字序列标注方法在关键词识别数以及正确识别数两项指标上均列前茅，这是其高召回率的直接原因，因此若为了尽可能多地识别出文本关键词，字序列标注方法更为可行。从特征选择上来看，词序列标注能够采用的特征明显高于字序列标注，一方面在汉语中，词语数量远远高于字数量（见图 3），使得词语本身表现出来的特征明显多于汉字；另一方面，在词序列标注中使用了 35 种标注角色，远多于字序列，这也使得词序列表现出了更强烈的特征复杂性。特征数的增多虽然加大了学习的复杂性，但也使得学习更有效果，加强了标注模型的区分度，P 值升高。训练样本特征多，使得训练的时间变长，词序列标注学习所需时间远远高于字序列标注，当然角色空间的复杂化也是训练时间变长的一个重要因素。

两种关键词抽取方法均具有较强的实用性。字序列标注方法的角色空间较为简单，训练速度较快，可利用的语言特征变少，能够召回更多的关键词，具有较高的综合识别效果。而词序列标注方法能够利用更多的语言特征约束，需要更长的训练时间，关键词识别准确率较高，但是召回的关键词变少，综合识别效果不如字序列标注；同时由于汉语组词的复杂性，关键词的一部分或全部有可能被歧义地成为另一个非关键词汇的一部分，这也是导致其角色空间相当复杂的主要原因，相应地从词语中分离出关键词组成成分的算法也变得非常复杂。

5　结　语

本文认为采用序列标注机器学习方法实现关键词抽取是可行而且是合理的，并指出目前该方法存在大规模学习语料难以获取、传统的词序列标注存在不稳定性和复杂性等弊端。为了消除障碍，笔者将序列标注方法应用于存在大量人工标引数据的中文书目关键词抽取领域，并通过对现有标引数据的详细分析，对词序列标注方法进行了简化和修正，构建了基于字序列标注的关键词抽取模型，提出了该模型的基本思路和实现方案，并以某大学图书馆馆藏书目作为实验对象，论证了模型的合理性、正确性和实用性。在仅以单汉字作为观察序列的情况下，F1 值达到了 87.36%，可见该模型具有很强的实用价值。

然而，本文并没有对影响该模型的特征因素以及实验参数进行详细论证，没有计算出最佳的标注模型，因此还有待于今后进一步开展研究，具体包括训练样本规模的合理优化、融合多特征的观察序列扩展、特征模板的选择、字角色空间的优化、CRFs 参数的控制以及不同序列标注机器学习算法的比较等；此外，序列标注是一种实用的机器学习方法和自然语言处理技术，它同样可用于从文本中抽取其他语言片段，例如从专利文献中抽取领域术语、标注领域术语间关系等。

参考文献：

[1] Hulth A. Combining Machine Learning and Natural Language Processing for Automatic Keyword Extraction[D]. Stockholm: Stockholm University, 2004.

[2] 王昊，严明，苏新宁. 基于机器学习的中文书目自动分类研究[J]. *中国图书馆学报*，2010,36(6)：28－39.

[3] 章成志，苏新宁. 基于条件随机场的自动标引模型研究[J]. *中国图书馆学报*，2008,34(5)：89－94，99.

[4] Chu C M, O'Brien A. Subject Analysis:The Critical First Stage in Indexing[J]. *Journal of Information Science*, 1993, 19(6): 439－454.

[5] 邓箴，包宏. 改进的关键词抽取方法研究[J]. *计算机工程与设计*，2009,30(20)：4677－4680，4769.

[6] 张雪英，Krause J. 中文文本关键词自动抽取方法研究[J]. *情报学报*，2008,27(4)：512－520.

[7] 徐文海，温有奎. 一种基于 TFIDF 方法的中文关键词抽取算法[J]. *情报理论与实践*，2008,31(2)：298－302.

[8] 张庆国，薛德军，张振海，等. 海量数据集上基于特征组合的关键词自动抽取[J]. *情报学报*，2006,25(5)：587－593.

[9] 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.

[10] 王灿辉，张敏，马少平，等. 基于相邻词的中文关键词自动抽取[J]. *广西师范大学学报:自然科学版*，2007,25(2)：161－164.

[11] 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.

[12] Frank E, Paynter G W, Witten I H, et al. Domain－Specific Keyphrase Extraction[C]. In: *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden. Morgan Kaufmann, 1999: 668－673.

[13] 章成志. 基于集成学习的自动标引方法研究[J]. *情报学报*，2010，29(1)：3－8.

[14] Zhang K, Xu H, Tang J, et al. Keyword Extraction Using Support Vector Machine[C]. In: *Proceedings of the 7th International Conference on Web－Age Information Management* (*WAIM2006*), Hong Kong, China. 2006: 85－96.

[15] 中国科学院计算技术研究所. ICTCLAS汉语分词系统简介[EB/OL]. [2011－08－13]. http://ictclas.org/ictclas_introduction.html.

[16] 黄昌宁，赵海. 由字构词——中文分词新方法[C]. 见：*中国中文信息学会二十五周年学术会议报告*，2006：53－63.

[17] Kudo T. CRF++: Yet Another CRF Toolkit[EB/OL]. [2011－08－07]. http://crfpp.sourceforge.net/.

（作者E－mail：ywhaowang810710@sina.com）

IMLS资助各单位进一步进行虚拟参考咨询服务合作研究

博物馆和图书馆服务研究所（IMLS）已授予OCLC研究部和Rutgers大学传播和信息学院一项国家领导计划项目（National Leadership Grant），以资助他们合作研究基于图书馆的虚拟参考咨询服务（Virtual Reference Service，VRS）。

OCLC的高级研究科学家Lynn Silipigni Connaway将和Rutgers大学传播和信息学院教员Marie L. Radford和Chirag Shah一起，作为联合首席研究员，研究能够提供更加具有合作性和可持续性的虚拟参考咨询服务的新模型。IMLS授予的这25万美元的国家领导计划项目为期两年。

项目名为“网络协同：通过虚拟参考和社会化问答网站的合作来寻求可持续发展”，是建立在前期一个IMLS资助的Rutgers和OCLC的合作项目之上，需要调查依赖于图书馆员和学科专家之间更广泛的合作模型。前期的合作项目“寻求同步性：从用户、非用户和馆员三个角度来评测虚拟参考咨询服务”由Radford和Connaway领导，持续了5年，研究的结果最近总结并由OCLC公开发表：《寻求同步性：虚拟参考咨询的启示和建议》（http://www.oclc.org/reports/synchronicity/）。

在过去的10年中，许多图书馆成功地推出了在线聊天和即时通讯参考咨询服务，以补充传统的面对面的服务。这些服务受到了广大用户的欢迎，但是在目前这种资金不断减少的环境下，很难维持下去。这个新项目会得到一些建议措施，以帮助图书馆界在实施下一代虚拟参考咨询服务（VRS）时更好地理解为何如此选择。

该项目提出了一种新模型，使得在目前这种资源不断减少的情况下，虚拟参考咨询服务仍然可以继续生存。该项目将会研究知识机构，如图书馆和社会化问答（SQA）社区无缝整合的可能性。统计资料显示虚拟参考咨询服务仍然继续增长，现在大多数图书馆都提供了虚拟参考咨询服务，来替代传统的面对面参考服务。

该项目分为三个阶段，将识别出虚拟参考咨询服务需要改进的地方，使得虚拟参考咨询服务能够更加可持续地发展并利用学科知识来满足用户的需求。第一阶段（记录内容分析）将纵向分析500个随机选择的虚拟参考咨询服务记录和1 000个社会化问答站点记录。第二阶段（电话采访和分析）将对150个虚拟参考咨询服务的核心用户和社会化问答站点的核心信息提供人员进行电话采访和深度分析。第三阶段（构建设计规范）将专注于创建链接虚拟参考咨询服务和社会化问答站点的设计规范，以探索虚拟参考咨询服务可持续发展的解决方案。

关于博物馆和图书馆服务研究所（IMLS）

博物馆和图书馆服务研究所是为全美123 000家图书馆和17 500家博物馆提供联邦支撑的主要机构。IMLS的使命是建立强大的图书馆和博物馆，为广大公众提供信息和思想。IMLS是国家级的单位，和地方分支机构协调一致地工作，维护社会文化遗产、文化和知识，加强社会大众的学习和创新，并为他们的职业发展提供支撑。

关于Rutgers大学传播和信息学院

传播和信息学院成立于1982年，其教育、学术和公共使命的基本前提是：在传播和信息过程中，必须把用户放在第一位。传播和信息学院的教师和毕业生塑造了传播、新闻、媒体、图书馆和信息科学等领域，这些领域的专业知识是传播和信息学院教学的核心。学院教师致力于研究以下问题：新媒体和民主，社交网络，虚拟环境和协同工作，健康和保健，以及领导力和政策。

（编译自：http://www.oclc.org/news/releases/2011/201156.htm）

（本刊讯）