DOI: CNKI:11-2746/G2.20111220.1231.002

基于字角色标注的中文书目关键词标引研究*

邓三鸿　王　昊　秦嘉杭　苏新宁

摘　要　中文书目机器自动标引是数字图书馆建设中亟待解决的关键问题之一。本文试图将条件随机场(CRFs)序列标注机器学习算法引入到关键词抽取中,建立面向图书内容、基于字角色标注的中文书目关键词标引模型。将图书内容转化为字序列,进而提出构建关键词角色空间模型和综合利用字序列上下文特征的设计思路。通过实验,从题名和内容提要中分别自动抽取关键词,论证该模型的合理性和实用性。图6。表3。参考文献23。

关键词　中文书目　关键词标引　字角色　序列标注　自动标引

分类号　G25　TP391

Research on Keywords Indexing for Chinese Bibliography Based on Word Roles Annotation

Deng Sanhong, Wang Hao, Qin Jiahang & Su Xinning

ABSTRACT　Automatic indexing by computers for Chinese bibliography has become one of the most critical problems which should be solved immediately in digital library construction. This paper tries to introduce Conditional Random Fields (CFRs) algorithm into the keyword extraction of Chinese bibliography, and builds the model which faces book contents based on the word roles annotation. The model turns the book contents into sequences of words. Based on that, an idea which combines word roles space model building with context features of word sequence comprehensive utilization has been proposed. Moreover, the paper also verifies the rationality and practicality of the model by showing the experiment of automatically extracting keywords from titles and abstracts. 6 figs. 3 tabs. 23 refs.

KEY WORDS　Chinese bibliography.　Keywords indexing.　Word roles.　Sequence annotation.　Automatic indexing

1　引言

图书自动标引是指利用计算机从书目内容中自动提取出能够代表该图书主题标引词的过程,按标引词来源可分为关键词(抽词)标引和主题词(赋词)标引。前者以在图书内容中出现过的语言片段作为标引词[1];后者是以关键词经过规范化后所形成的领域知识(即主题词)作为标引词[2]。关键词标引作为图书馆工作自动化的重要组成部分,是实现海量图书内容检索的前提,也是实现图书其他自动处理,如图书自动分类和聚类、自动摘要、个性化推荐等工作的核心技术[3-4]。目前,中文图书的关键词标引多采用手工方式,由图书作者或图书编目人员给出关键词。由于缺乏规范,很多作者并没有给出关键词,而面对巨大的图书出版量,对书目内容并不十分了解同时也缺乏相应领域知识的图书编目人员在手工补充标引时显得力不从心,不仅严重影响了图书馆的工作效率,而且获取的关键词具有随意性。在这种情况下,将信息自动化技术引入到图书编目工作中,探索书目自动标引的方法和过程,提高图书馆工作效率,成为数字图书馆建设中亟待解决的关键问

*　本文系国家社科基金项目"面向语义网本体的知识管理研究"(编号:09CTQ010)的研究成果之一。
通讯作者:王昊,Email:ywhaowang810710@sina.com

题之一。

总结前人的研究成果,目前中文图书的关键词抽词标引主要有两种方法:一是传统的基于分词的统计方法,即首先对图书文本内容进行精确分词,在排除非用词后建立候选关键词集合,再采用各种统计算法如词频、TF-IDF 值、ATF × PDF 值等对候选关键词进行权重计算,进而根据权重筛选关键词[5-8];另一种则是基于词序列的语言学方法,即首先对图书文本进行粗分词,使文本转化为词汇序列,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,进而判断是否可将相邻词汇合并作为关键词[9],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,进而根据关键词角色模板将相关词序列合并为关键词[4,10-11]。上述两种方法均存在明显缺陷:①目前中文分词技术还没有达到非常精确,现有的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领域术语相冲突;②中文组词具有很强的灵活性,使得词汇数量非常庞大,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。

本文试图对词序列标注机器学习方法的标引模式进行改进,将关键词看作是汉字的组合,从而设计关键词的字角色空间,在此基础上构建基于字角色标注的书目关键词标引模型,针对图书内容(这里仅指题名和内容提要)及其关键词的特点和关系,对人工关键词标引的特征和规律进行机器学习,进而利用生成的标注模型对书目实现自动关键词标引,以解决图书人工标引投入大、效率低、准确性差等问题。笔者通过实验论证自动标引模型的正确性和合理性,并通过对比分析探索影响该模型的特征因素,以期获得最佳的序列标注模型,为书目自动标引系统的实际开发和具体应用提供事实依据。

2 基于字角色标注的书目关键词标引模型

本节在对现有图书标引数据统计和分析的基础上,总结书目关键词的来源特征,提出了采用机器学习中的序列标注技术、基于书目内容的上下文特征抽取关键词的基本方法,分析在自动标引过程中需要解决的关键问题,并针对中文分词准确性低以及中文汉字特征复杂的特点提出基于字角色标注的图书自动标引模型。

2.1 图书标引数据的统计和分析

图书作为文本信息,内容篇幅较长,重点较分散,因此从正文获取图书主要内容(关键词)既不可行也不必要,可以考虑从图书著录项题名和内容提要中抽取关键词进行标引。

为了探索图书自动标引的思路和方法,有必要对实验数据进行统计分析,从总体上了解图书标引数据的分布及特点。本文的实验数据来自某大学图书馆,至 2010 年 6 月底,该馆共有馆藏书目 187,213 种,其中给出标引词的有 150,850 种。属于关键词标引的,即标引词能够在书目内容(这里仅包括题目和内容提要)中抽取的则有 65,482 种,涉及关键词 6,511 个;题名标引中的有 52,279 种,涉及关键词 5,746 个;提要标引的 42,754 本,涉及关键词 5,167 个;标引词同时出现在题名和提要中的有 29,530 种,涉及关键词 4,159 个。本文重点探讨书目关键词标引的方法和过程,以上述关键词标引数据作为分析对象,此后如不加说明,文中所指关键词即为标引词。

根据对现有书目人工关键词标引数据的分析,笔者得到以下结论:①由于图书具有内容丰富和类目明确等特点,其关键词相对期刊文献一般较少,被限制在 1 ~ 5 个词语之间,实验数据中有 99.8% 以上的图书仅有 1 ~ 2 个关键词,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,而且相对来说具有较高准确率和较大实用性;②关键词是图书内容的高度浓缩,一般具有很强的类目特征,而在本文采用的实验数据中存在较为严重的图书类目分布不均衡现象,其中 B、C、D、F、G、H、I、J、K、O、T 等 11 类的书目数量占总数 94% 以上,这样为书目特征的抽取(即关键词的抽取)增加了难度,特别是在采用机器学习方法抽

取书目特征时,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;③笔者采用 ICTCLAS 分词系统对书目关键词进行了切分,发现约 57.61% 的关键词可以被切分成2~6个词汇,关键词的多词汇组合特征使其不适合采用传统的分词后词频统计方法进行抽取,可以考虑采用语言片段角色标注的思路判断语言片段是关键词的组成部分,进而认为连续的语言片段组合即为关键词;④组成关键词的语言片段可以是词语和汉字,经过统计笔者发现,组成关键词的汉字种类远远少于词语类型,那么字切分后获得的语言学特征将比词切分少,可以降低机器学习的复杂度,而且关键词的词组合类型远比字组合类型来得多,使得词角色空间比字角色空间要复杂得多,正是由于中文词语组合的多样性以及中文分词的不准确性等特点,采用字角色标注可能比词角色标注具有更高的准确率和稳定性。

2.2 中文书目关键词标引模型的分析和设计

根据关键词标引的特点和规律,笔者构建了一个基于字角色标注的中文书目关键词标引模型,基本思路是:首先建立一个角色空间模型,用于标识在图书内容中出现的所有汉字(包括符号串);将图书内容转化为字观察序列作为第一种语言特征,并对特征进行衍生,扩展观察序列以强化汉字的上下文语境规律;然后根据角色空间模型,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学习样本(训练语料);根据机器学习的思想,采用机器学习模型对学习样本的序列规律(纵向上下文特征)和标注规律(横向标注特征)进行学习,形成学习模型;最后将学习模型作用于仅由观察序列构成的测试样本(测试语料)进行序列标注,自动获得字序列所对应的角色序列,则符合关键词角色模板相应的字序列组合即为关键词,整个过程如图 1 所示。

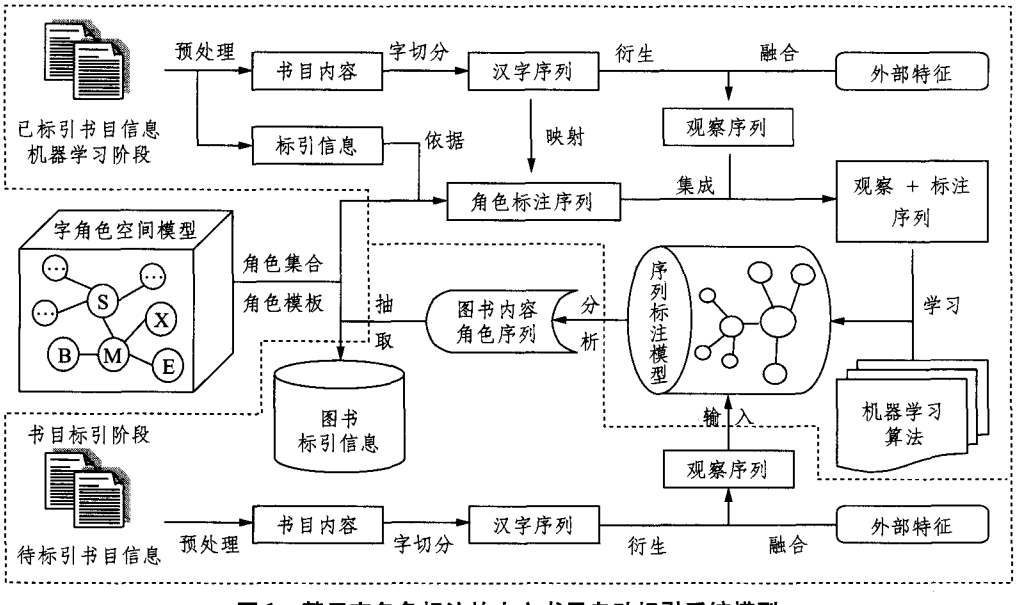


图 1 基于字角色标注的中文书目自动标引系统模型

图 1 中,笔者将整个图书自动标引过程分为两个阶段: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。在上述系统模型中,涉及字角色空间模型创建、观察序列衍生、机器学习算法选择以及特征模板设置等在具体应用中需要解决的问题。

(1)字角色空间模型创建

字角色空间在整个自动标引模型中占据核心地位,角色设定合理与否将直接关系到标引

模型的成败。在学习阶段,它提供角色集合参与字序列的角色标注;在分析阶段,它提供角色组合模板,用于判断连续角色符号所对应的字序列组合是否为关键词。根据关键词的字组成规律,笔者将标引模型中所有汉字(包括字符串)分成6种角色(L)(见表1)。

表1 汉字标注角色集合

角色(L)	说 明	示 例
B	关键词首字	如"汽车保险"之"汽"
M	关键词中字	如"汽车保险"之"车"、"保"
E	关键词尾字	如"汽车保险"之"险"
S	单字关键词	如"水与文化"之"水"
X	非关键词中的单字	如"水与文化"之"与"、"文"、"化"
F	符号串	如"Microsoft Project 2002 标准教程"之"Microsoft"

定义1:集合 $\Omega = \{R, P\}$ 为字角色空间,其中R为字角色集合,P为关键词角色组合模板集合。

定义2:集合 $R = \{B, M, E, S, X, F\}$ 为字角色集合。其中,单字关键词中汉字标注为S,多字关键词根据其组成汉字在词中位置分别标注为B、M、E;非关键词中的汉字标注为X;在中文书目中出现的符号串标注为F。根据R,可以将任意汉字转化为角色,完成字空间到角色空间的映射。

定义3:集合 $P = \{S, BE, BM_nE \mid n = 1, 2, \cdots, N\}$ 为关键词角色组合模板集合。其中,S表示单汉字关键词模板,BE为两个汉字关键词模板,BM_nE 为多汉字关键词模板(n为大于1的整数)。P中模板所对应的汉字序列即为关键词。

需要说明的是,在书目题名和内容提要中可能出现多个使用了相同语言片段的关键词,那么这些共享的语言片段就可能出现多种角色。例如"可持续发展"和"发展战略"是书目"可持续发展战略"的关键词,那么"发"和"展"均存在两种角色,分别为"M"、"B"和"E"、"M",本文用符号组合"MB"和"EM"对其进行标注。

(2)观察序列扩展

本文采用序列标注的方法实现书目关键词的抽取。序列标注的数据样本由两部分构成:观察(特征)序列和标注(角色)序列。标注序列仅出现在学习样本中,由关键词角色组成;而观察序列则是学习算法的重点考察对象,在学习样本和测试样本中均存在,由各种文本特征所组成。①字序列(B),最基本的观察序列,汉字自身的特点及其上下文语境特征决定了汉字的表现角色,由此来判断其所属角色是否最合理;②类目序列(C),每个类目都有相应的词汇集合,即每个类目都有相应的汉字集合,通过类目可以在一定程度上反映出汉字特征;③位置序列(P),指字在其所在词语中的位置,可以在一定程度上反映出当前汉字与其前后汉字之间的关系以及其所在词汇的长度等特征,本文采用4词位标记[12]:若汉字在其所在词汇中居于首位则记为M,居于末尾则记为N,其他位置记为W,单字词记为V;④姓氏序列(N),用于记录当前汉字是否为汉语常用姓氏,如果是则记为Y,不属于则记为N;⑤音译外来字序列(T),有些汉字经常用于对外文单词的翻译,称之为音译外来字,这类汉字具有不同于其他汉字的一些特征,可将属于该集合的汉字记为Y,不属于则记为N。上述观察序列的标记、取值及其具体描述参见表2。

表2中示例部分为某一书目的学习语料,其中前五列为5类观察序列值,例如"中",其在词中位置为词首字"M",为非姓氏字"N",是常用的音译外来字"Y",在当前语境下属于经济学"F"类;最后一列为标注的角色序列,如"中"的角色标注为"X",为非关键词成分;根据角色模板,不难发现"BME"为关键词的角色组合,其所对应的汉字组合"增值税"即为关键词,于是书目关键词抽取问题就被转化为汉字角色标注问题。

表2 观察序列标记、取值及其描述

观察序列	取值情况	描述	示例					
字序列(B)	具体汉字或连续字符串	字形特征	B	P	N	T	C	L
			中	M	N	Y	F	X
类目序列(C)	中图法大类号	类目特征	国	N	N	N	F	X
位置序列(P)	M	所在词词首	增	M	N	Y	F	B
			值	W	N	N	F	M
	N	所在词词尾	税	N	N	Y	F	E
	W	所在词词中	转	M	N	N	F	X
			型	N	N	Y	F	X
	V	单字词	可	M	N	Y	F	X
姓氏序列(N)	Y	姓氏字	行	W	N	Y	F	X
			性	N	N	N	F	X
	N	非姓氏字	实	M	N	Y	F	X
音译字序列(T)	Y	音译外来字	证	N	N	N	F	X
			研	M	N	N	F	X
	N	非音译外来字	究	N	Y	N	F	X

(3)机器学习算法选择

所谓机器学习,就是根据对已知情况状态及其可能原因的分析和学习,来判断未知情况可能状态的方法和过程,常用于状态确定之类问题的解决。机器学习方法分为两类:①根据对象自身特征判断对象状态,通常将对象用特征向量来描述,例如表2中用五元特征向量<中,M,N,Y,F>来描述"中",机器据此自动将其判断为"X"。这实际上是分类问题,常用的算法有决策树[13]、人工神经网络[14]、支持向量机[15]、朴素贝叶斯[16]等;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来描述,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。例如表2中,不仅可以用向量<国,N,N,N,F>从横向上描述汉字"国",还可以通过其前字<中,M,N,Y,F>以及后字<增,M,N,Y,F>从纵向关系上描述。这实际上是一个序列标注问题,常用的算法有隐马尔科夫模型(Hidden Markov Model, HMM)[17]、最大熵模型(Max Entropy Model, MEM)[18]、条件随机场(Conditional Random Fields, CRFs)[19]等。

在两类机器学习方法中,前者适用于对象的特征较多而且特征值具有较强区分度的情况,后者适用于上下文语境对当前对象具有较强影响的环境。以表2示例中的汉字<增,M,N,Y,F>为例,其被标注为什么角色,不仅可以由本身的向量值决定,而且和其所处的上下文存在很大关系,在表2中被标注为"B",而在"国有资产增值难题研究"中被标注为"X",在"经济增长理论"中又被标注为"M",因此在本文构建的模型中采用序列标注的机器学习方法;又由于HMM存在生成模型所固有的独立性假设和无法融合多种特征的缺陷,MEM则因局部归一化导致标记偏置(label bias)问题,CRFs则是目前序列标注效果最佳的机器学习算法[20]。本文采用CRFs算法实现训练语料的机器学习,并以开源软件CRF ++ 0.51[21]作为CRFs算法的运行平台。

(4)特征模板设置

CRFs算法不仅可以利用对象本身的特征,还可以利用对象上下文特征,将观察序列和上下文约束结合在一起,可以设置CRFs的特征模板。单字的上下文约束通过字长窗口反映。常用的上下文信息主要包括远程上下文信息和局部上下文信息,前者指与当前对象具有一定文本距离的对象所提供的长距离约束,如"词语触发对"[22];后者指以当前汉字为中心,向前或(和)向后连续选取一定长度范围的上下文作为当前汉字的约束,这个局部连续范围称为字长窗口,常用的有3字长和5字长窗口[12]。本文采用后者来纵向约束当前字对象。

选择不同的观察序列组合以及不同的上下

文约束,笔者建立了如表3所示的5组特征模板(feature template)[23]。表中B表示字序列,N表示姓氏特征,T表示音译外来字特征,P表示词位标注特征,C表示类目特征,L表示标注角色特征;B_n等表示观察对象自身特征,n为字长,即考察当前对象前后2个字长范围内的语言特征;$B_{n-1}B_n$等表示当前对象与其他对象之间的二元关系特征,$B_{n-2}B_{n-1}B_n$等为当前对象与其前两个对象之间的三元关系特征;$L_{-1}L_0$则表示前一个字的标注角色对当前字角色标注的影响。

表3 书目关键词字角色标注的特征模板

模板名称	观察特征	标注角色	n-gram	特征模板
TMPT1	B	L	1-gram	$B_n, n=-2,-1,0,1,2$
			2-gram	$B_{n-1}B_n, n=-1,0,1,2$; $B_{n-2}B_n, n=0,1,2$; $L_{-1}L_0$
			3-gram	$B_{n-2}B_{n-1}B_n, n=0,1,2$
TMPT2	BP	L	1-gram	$B_n, P_n, B_nP_n, n=-2,-1,0,1,2$
			2-gram	$B_{n-1}B_n, P_{n-1}P_n, n=-1,0,1,2$; $B_{n-2}B_n, P_{n-2}P_n, n=0,1,2$; $L_{-1}L_0$
			3-gram	$B_{n-2}B_{n-1}B_n, P_{n-2}P_{n-1}P_n, n=0,1,2$
TMPT3	BPNT	L	1-gram	$B_n, P_n, N_n, T_n, B_nP_nN_nT_n, n=-2,-1,0,1,2$
			2-gram	$B_{n-1}B_n, P_{n-1}P_n, N_{n-1}N_n, T_{n-1}T_n, n=-1,0,1,2$; $B_{n-2}B_n, P_{n-2}P_n, N_{n-2}N_n, T_{n-2}T_n, n=0,1,2$; $L_{-1}L_0$
			3-gram	$B_{n-2}B_{n-1}B_n, P_{n-2}P_{n-1}P_n, N_{n-2}N_{n-1}N_n, T_{n-2}T_{n-1}T_n, n=0,1,2$
TMPT4	BPNTC	L	1-gram	$B_n, P_n, N_n, T_n, C_n, B_nP_nN_nT_nC_n, n=-2,-1,0,1,2$
			2-gram	$B_{n-1}B_n, P_{n-1}P_n, N_{n-1}N_n, T_{n-1}T_n, C_{n-1}C_n, n=-1,0,1,2$; $B_{n-2}B_n, P_{n-2}P_n, N_{n-2}N_n, T_{n-2}T_n, C_{n-2}C_n, n=0,1,2$; $L_{-1}L_0$
			3-gram	$B_{n-2}B_{n-1}B_n, P_{n-2}P_{n-1}P_n, N_{n-2}N_{n-1}N_n, T_{n-2}T_{n-1}T_n, C_{n-2}C_{n-1}C_n, n=0,1,2$
TMPT5	BPNTC	L	1-gram	$B_n, P_n, N_n, T_n, C_n, B_nP_nN_nT_nC_n, n=-1,0,1$
			2-gram	$B_{n-1}B_n, P_{n-1}P_n, N_{n-1}N_n, T_{n-1}T_n, C_{n-1}C_n, n=0,1$; $B_{n-2}B_n, P_{n-2}P_n, N_{n-2}N_n, T_{n-2}T_n, C_{n-2}C_n, n=1$; $L_{-1}L_0$
			3-gram	$B_{n-2}B_{n-1}B_n, P_{n-2}P_{n-1}P_n, N_{n-2}N_{n-1}N_n, T_{n-2}T_{n-1}T_n, C_{n-2}C_{n-1}C_n, n=1$
TMPT6	BPNTC	L	同TMPT4,仅除去$L_{-1}L_0$	

3 面向题名的书目关键词标引实验分析

题名作为对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。在所有图书标引数据中,共有关键词156,772个,其中53,217个来自书目题名,约占总数的1/3强。由此可见,从题名中抽取关键词进行图书标引是合理的,也是常见的人工标引方式。本节以题名关键词标引数据作为实验对象进行分析,探讨影响书目标引的主要特征,论证前述自动标引模

型的正确性和合理性。

本文采用经典的评价指标正确率(P)、召回率(R)和F1值对标引模型进行性能评价,并以关键词识别数(C)、正确识别数(RC)以及标注正确率(RI,即单字被正确标注的概率)作为辅助指标,如下所示:

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

$$R=\frac{RC}{N}(\text{其中 } N \text{ 表示人工标引词的个数}) \quad (2)$$

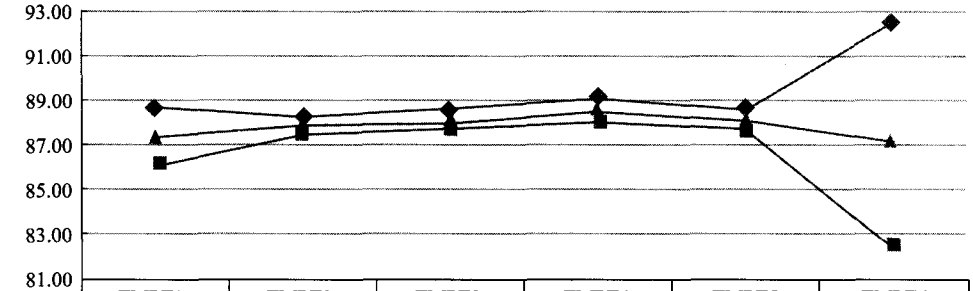
$$F_1=\frac{2\times P\times R}{P+R} \quad (3)$$

$$RI=\frac{\text{机器正确标注的字个数}}{\text{人工正确标注的字个数}} \quad (4)$$

在本次实验中,53,217个关键词共来自52,279本图书。笔者随机选择其中47,000本图书的标引信息作为训练数据,共有关键词47,938个;以剩余的5,279本图书作为测试数据,共有关键词5,279个。

3.1 基于不同特征模板的标引结果比较分析

笔者以集成了CRFs算法的CRF++0.51作为运行平台,分别以表3所列的6个特征模板进行机器学习实验,测试结果如图2、图3所示。



	TMPT1	TMPT2	TMPT3	TMPT4	TMPT5	TMPT6
P	88.56	88.29	88.42	89.07	85.50	92.46
R	86.19	87.42	87.69	88.03	87.78	82.48
F1	87.36	87.85	88.05	88.55	88.14	81.18

图2 基于不同特征模板的书目关键词标引主要测评指标结果比较(%)

图2显示了在不同特征模板的作用下,即随着书目单字横向和纵向特征的变化,书目自动标引正确率、召回率以及F1值等的变化规律。①当以字本身作为唯一观察序列时(即TMPT1),书目自动标引的正确率P值达到88.56%,离标引效果最好的TMPT4(F1值最高)的正确率89.07%只差0.51个百分点,其召回率R值也超过了85%,可见单字本身所具有的特征起到了重大作用。②随着观察序列的扩充(TMPT1~TMPT4),单字及其扩展值所表现出来的特征越来越多,然而P值提高不多,在加入某些观察序列之后反而出现了P值下降的情况;但是R值产生了明显的变化,并带动综合指标F1值出现了较大幅度的提升。由此可知,在书目关键词抽取中,观察序列的合理扩充能够导致书目特征的增加,可以有效地提高R值和F1,但对P值则影响甚微。③对比TMP1~TMPT4的F1值,图2中从左到右分别增加了0.49、0.20和0.50个百分点,增加词位和类目观察序列的识别效果相对较好,可见对书目关键词的自动抽取而言,字的词位、类目特征比姓氏、音译外来字特征具有更强的影响效果。④TMPT4和TMPT5的观察序列相同,不同之处在于后者比前者的字长窗口变小了,后者仅考虑前后字对当前对象的影响,字长宽口为3,由此也导致了P、R、F1值3项指标都出现了较明显的下滑,可见字长窗口越大,可利用的上下文特征也越多,书目自动标引的效果也越好。⑤TMPT6与TMPT4的区别仅在于前者没有以前字的标注结果作为上下文特征,结果出现了较

大的异常:P 值达到 92.46%,为所有模板最高,但是高 P 值是以低召回为代价的,R 值仅为 82.48%,是唯一一个低于 85% 的特征模板,综合结果 F1 值为 81.18%,相比 TMTP4 下降了 1.37 个百分点,可见前字的标注结果对后字的标注具有决定性的影响,是最重要的上下文特征。

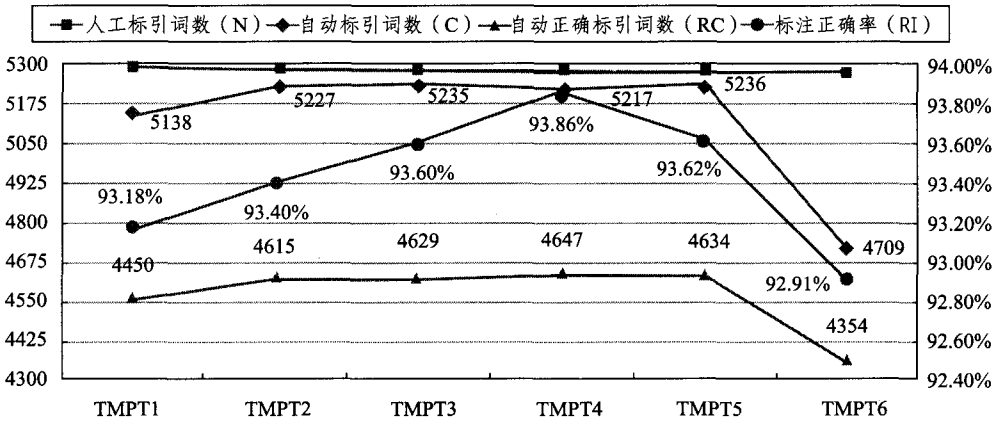


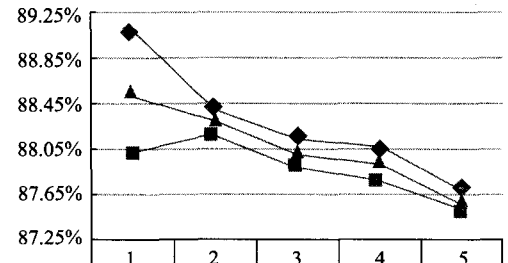
图 3 基于不同特征模板的书目关键词标引辅助测评指标结果比较

图 3 列出了以不同特征模板进行实验时,可以获得的自动关键词数(C)、自动正确关键词数(RC)以及单字的标注正确率(RI)等辅助测评指标,同时列出人工关键词数进行比较。①TMPT4 的自动标引效果是最好的,相对于 TMTP1 ~ TMPT3,其观察序列最多,可利用的字特征也最多;相对于 TMPT5,其字长窗口较大,可利用的上下文信息就越多;相对于 TMPT6,其多利用了前一个字的标注信息。正因为其充分利用了各种特征,在 C 值不明显占优的情况下,RC 值最高,直接导致了 P 和 R 值最高,若本文提出的书目自动标引模型最终用于实际工作,那么通过该模板获得的序列标注模型是最合理和最有效的。②从 TMPT1 到 TMPT4,RC 值逐渐升高,这是致使 R 值逐渐提高的直接原因,而 RC 值的升高又是由于 C 值变大引起的,不难想象,随着自动识别的关键词个数变多,其中被正确识别的关键词也会随之变多。③由于 RC 的增加速度低于 C 值的变化,因此在 TMPT2 和 TMPT3 比 TMPT1 增加了观察序列之后,P 值出现了不升反降的态势,可见在书目自动标引中,观察序列扩充的主要作用在于提高了 C 值,进而带来了 RC 值、R 值以及 F1 值升高的连锁反应。④图 2 中 TMPT6 的 P 值达到了最高,比识别效果最佳的 TMPT4 高出了 3.39 个百分点,结合图 3 可以发现,TMPT6 的高正确率是由于 C 值下降的幅度远大于其 RC 值的下降幅度所造成的,因此 R 值相对较低。可见,正确率并不是衡量识别效果的唯一标准,还应该结合召回率进行综合分析。⑤本文采用了字标注方法抽取关键词,要抽取出正确的关键词,首先要保证字的角色被正确标注了,图 3 中列出了不同特征模板下单字被正确标注的概率,融合了最多特征的 TMPT4 的 RI 值达到了 93.68%,而标注效果最差的 TMPT6 也达到了 92.91%,可见字角色标注模型在 CRFs 算法作用下具有很强的实用价值。

3.2 基于不同控制参数的标引结果比较研究

CRFs 算法有两个重要的可调节参数:特征函数阈值(f)和软边界参数(c)[23]。前者用于控制参加计算的特征数,即在训练数据中出现 f 次以上的特征是可参考的,f 值越大,可利用的特征必然越少,反之可参加计算的特征越多;后者主要用来调节 CRFs 算法中数据欠拟合(underfitting)和过拟合(overfitting)之间的平衡。

在利用训练数据进行学习中，可以调节 f 和 c 值生成不同的序列标注模型，而不同标注模型的预测能力也不相同。笔者采用了综合效果最佳的 TMPT4，在不同的 f 值(1～5)和 c 值(1～5)作用下分别进行了书目关键词自动抽取实验，结果见图 4、图 5。



	1	2	3	4	5
P	89.07%	88.40%	88.15%	88.04%	87.66%
R	88.03%	88.20%	87.93%	87.82%	87.50%
F1	88.55%	88.30%	88.04%	87.93%	87.58%

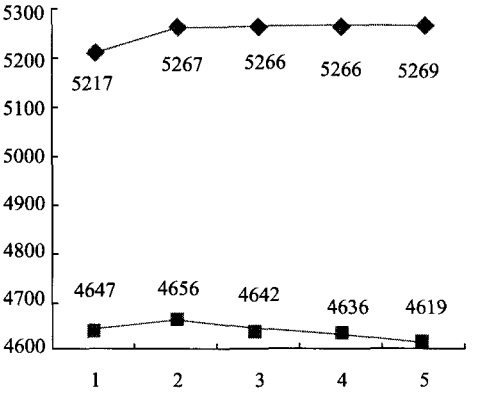
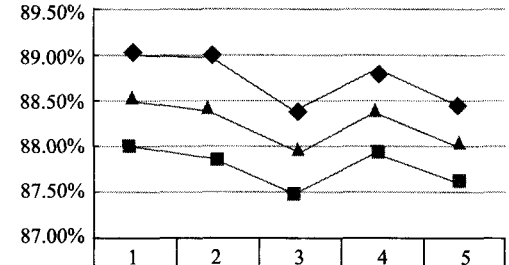


图 4　基于不同特征函数阈值(f=1…5)的书目关键词标引测评指标结果比较

图 4 中清楚显示了当软边界参数 c=1，特征函数阈值 f=1…5 时各项测评指标值的比变化情况。①总整体上看，随着特征函数阈值(f)的逐渐增大，P、R 以及 F1 值均发生明显的连续下降，可见 f 值下降带来的训练语料上下文特征减少使得各标注角色的区分度变小，字角色被误标注的概率增大，导致标注的正确率急剧下降，被正确抽取的关键词无论在个数上，还是在比率上都变小；②随着 f 值的变化，P 值和 F1 值一直下跌，特别是 P 值，降幅很大，然而 R 值和 RC 值则是先升后降，在 f=2 时，两个指标值均达到最大，可见随着文本特征的减少，被正确标注为关键词角色的概率增加，关键词被召回的可能性也变大。但是，随着 f 值的进一步增大，被误标注为关键词角色的概率也急速变大，被召回的关键词虽然增加了，但正确召回的却减少了。因此，若应用的重心在于尽可能多地抽取出关键词，那么可以适当增大 f 值(f=2)以提高关键词的召回。



	1	2	3	4	5
P	89.07%	89.05%	88.40%	88.89%	88.50%
R	88.03%	87.82%	87.44%	87.91%	87.59%
F1	88.55%	88.43%	87.92%	88.40%	88.04%

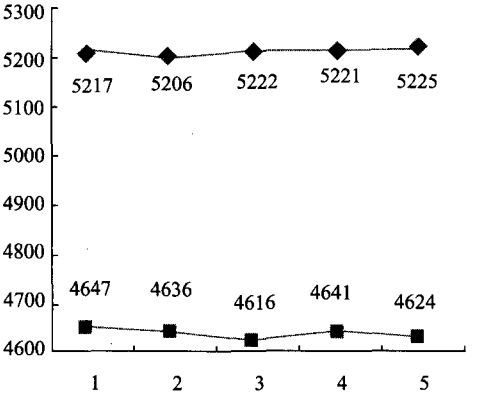


图 5　基于不同软边界参数(c=1…5)的书目关键词标引测评指标结果比较

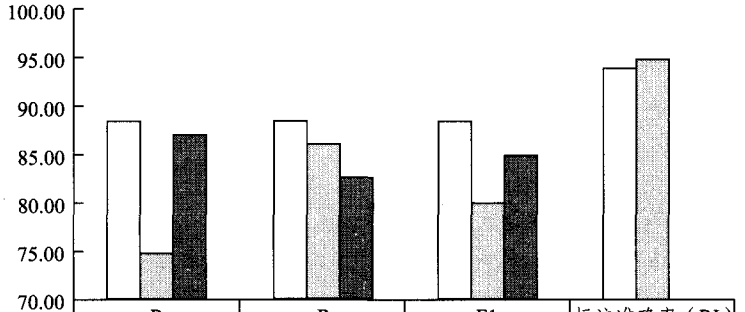
图 5 显示的是当特征函数阈值 f=1，软边界参数 c=1…5 时各项测评指标值的变化情况。

①从总体来看,随着软边界参数 c 的增大,识别的关键词个数基本保持平衡,在 5200 ~ 5230 附近振荡,但是正确召回的关键词持续减少,导致自动标引的效果越来越差;②当 c = 1 时,自动标引模型的综合识别效果为最佳,F1 达到最高的 88.55%。因此,在 f = 1 并且 c = 1 的情况下,TMPT4 应用效果是最好的,可参与实际应用;③在 c = 3 时,出现了一个拐点,正确召回的关键词数降至最低的 4616 个,由此导致 P、R 和 F1 值出现了急速下滑,但当 c = 4 时,识别效果又有所恢复,此后继续保持下落态势。

4 面向内容提要的书目关键词标引实验分析

前文指出,在图书馆馆藏的所有书目中,大约有 1/3 的关键词直接来自题目,通过机器自动抽取,大约能识别 88% 左右的关键词;除此之外,还有 42,967 个关键词来自 42,754 本书目的内容提要,约占关键词总数的 27.41%,可以认为提要是图书关键词的第二大来源,从提要中抽取关键词可以作为合理补充。笔者对此也进行了实验分析。

本文随机选择 38,000 本图书标引数据作为训练数据,共有关键词 38,213 个;剩余的 4,754 本图书作为测试数据,共有关键词 4,754 个。通过实验测试,在 c = 1,f = 2 的参数控制下,在测试数据中共获得关键词(C)5475 个,其中正确识别(C,即与人工标引匹配)的有 4087 个,可得主要测评指标 P、R 以及 F1 值分别为 74.65%、85.97% 和 79.91%。图 6 列出了在相同实验参数下,面向题名和提要的关键词抽取测评指标比较的情况。



	P	R	F1	标注准确率(RI)
□题名标引	88.40	88.20	88.30	93.75
□提要标引	74.65	85.97	79.31	94.64
■提要词语	87.12	82.52	84.76	

图 6 面向主题和提要的书目关键词标引测评指标结果比较(%)

从图 6 发现:①在图书提要中抽取关键词的 P、R 和 F1 值都相对较低,综合效果下降了 8.39 个百分点,降幅如此之大的主要原因是标引的准确率下降明显,比题名标引少了 11.75 个百分点,而召回相对较高,达到 85% 以上;②在标注准确率(RI)一项上,摘要标引比题名标引准确率高出 0.89 个百分点,标注准确率提高了,而标引的 F1 值反而下降了,笔者认为主要是由于 RI 值描述的是单字的角色标注;而 F1 值描述的是从书目中抽取出关键词的总体效果,并没有反映出单个关键词词语的抽取状况,而在书目提要中有的关键词则是多次出现;③笔者以关键词词语作为对象,进一步考察抽取效果,结果如图 6 中各指标末列所示,在 4,754 本测试书目中,有人工关键词词语 11,593 个,通过机器可抽取出 10,982 个词语,属于正确抽取的有 9,567 个,P、R 和 F1 值均恢复到 80% 以上。

相对书目的题名,书目内容提要的篇幅较

长,关键词可能在其中多次出现,因此笔者从关键词以及关键词词语两个角度对实验结果进行了评价。根据实验结果,笔者得出结论:面向书目内容提要抽取关键词效果较面向题目抽取相对较差,但各项测评指标均达到了70%,综合指标F1值接近80%,可应用于实际工作。

5 结语

本文在对实践数据进行分析的基础上,论证了书目关键词自动标引的可行性,提出了采用机器学习序列标注方法实现书目关键词抽取的基本思路和可行方案,构建了面向图书内容基于字序列标注的关键词标引模型,并以某大学图书馆书目标引数据作为研究对象,对模型进行了实验论证和分析。笔者认为:①对图书题名和内容提要采用字序列标注思想抽取关键词的方法是可行的;②在合理的观察序列环境下,CRFs算法的标注准确率能够达到92%以上,甚至接近95%,可见CRFs算法在书目内容的字角色标注中具有实用价值;③可通过扩展观察序列、加大上下文窗口以及改进字角色空间模型等方式提高角色标注准确率,进而增强书目自动标引的综合效果;④从题名中抽取关键词的综合效果可达到88.55%,从提要中抽取则可以达到79.91%,超过或接近80%,说明本文提出的模型在书目自动标引中具有一定的实用价值。

在本文实验基础上,笔者认为可以进一步建立实用的以题名抽取为主,提要抽取为辅的联合书目关键词标引平台,用于图书馆自动编目工作以增强书目关键词标引的效率和客观性,提高图书馆工作的自动化程度;序列标注是机器学习技术的一个重要分支,是利用上下文语境特征实现特定短语抽取和语言片段身份标注的常用方法,图书馆中的很多工作都可以采用序列标注实现自动化,例如领域专业术语的抽取以实现领域知识库、自动问答系统中关键语言片段抽取以实现虚拟参考咨询服务、期刊自动标引系统开发等等;本文建立了6角色空间模型,然而在实验中发现各角色的分布很不平衡,其中X角色的字数量远远超过其他角色,而S角色正好相反,不同角色字数量的巨大落差容易引起数据稀疏,进而带来标注偏差等问题,可以通过增加角色,提高空间模型的复杂度来平衡各角色之间的数量差异,如增加关键词前缀字、后缀字等角色。完善书目关键词标引模型,进而开发出实用的标引平台将是今后研究的方向。

参考文献:

[1] 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(5):89-94,99. (Zhang Chengzhi, Su Xinning. Automatic indexing model based on conditional random fields [J]. Journal of Library Science In China, 2008 (5):89-94,99)

[2] Chu C M, O'Brien A. Subject analysis: The first critical stages in indexing[C]. Journal of Information Science, 1993, 19(6):439-454.

[3] 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,2011(5):28-39. (Wang Hao, Yan Ming, Su Xinning. Research on automatic classification for chinese bibliography based on machine learning[J]. Journal of Library Science In China, 2011 (5): 28-39.)

[4] 李素建,王厚峰,俞士汉,等. 关键词自动标引的最大熵模型应用研究[J]. 计算机学报,2004:1192-1197. (Li Sujian, Wang Houfeng, Yu Shihan, et al. Research on maximum entropy model for keyword indexing[J]. Chinese Journal of Computers, 2004:1192-1197.)

[5] 张雪英,Jürgen Krause. 中文文本关键词自动抽取方法研究[J]. 情报学报,2008(4):512-520. (Zhang Xueying, Jürgen Krause. An approach to automatic keyword extraction in Chinese text[J]. Journal of the China Society for Scientific and Technical Information, 2008(4):512-520.)

[6] 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(2):298-302. (Xu Wenhai, Wen Youkui. A Chinese keyword extraction algorithm based on TFIDF method[J]. Information Studies: Theory & Application, 2008(2):298-302.)

[7] 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(5):587-593. (Zhang Qingguo, Xue De-

jun, Zhang Zhenhai. Automatic keyword extraction from massive data sets based on feature combination[J]. Journal of the China Society for Scientific and Technical Information, 2006(5):587 - 593.)

[8] 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(6):75 - 79. (Yang Jie, Ji Duo, Cai Dongfeng, et al. Keyword extractionin multi-document based on Joint Weight[J]. Journal of Chinese Information Processing, 2008(6):75 - 79.)

[9] 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(2):161 - 164. (Wang Canhui, Zhang Min, Ma Shaoping, et al. Chinese keyword extraction algorithm based on neighbour words [J]. Journal of Guangxi Normal University: Natural Science Edition, 2007(2):161 - 164.)

[10] 章成志. 基于集成学习的自动标引方法研究[J]. 情报学报,2010(1):3 - 8. (Zhang Chengzhi. Automatic indexing method based on ensemble learning[J]. Journal of the China Society for Scientific and Technical Information, 2010(1):3 - 8.)

[11] Zhang K, Xu H, Tang J, et al. Keyword extraction using support vector machine[C]. Proceedings of the Seventh International Conference on Web-Age Information Management (WA IM2006), Hong Kong, China, 2006:85 - 96.

[12] 黄昌宁,赵海. 由字构词——中文分词新方法[C]. 中国中文信息学会二十五周年学术会议,2006:53 - 63. (Huang Changning, Zhao Hai. Character-based tagging: A new method for Chinese word segmentation[C]. Proceedings of the 25th anniversary conference of Chinese Information Processing Society of China, 2006:53 - 63.)

[13] Quinlan J R. Induction of decision tree[J]. Machine Learning, 1986, 1(1):81 - 106.

[14] Hecht-Nielsen R. Theory of the back propagation neural network[C]. Proceedings of International Joint Conference on Neural Networks, IEEE, 1989, 1:593 - 603.

[15] Cortes Corinna, Vapnik V. Support-vector network [J]. Machine Learning, 1995(20):273 - 297.

[16] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers[J]. Machine Learning, 1997 (29):131 - 163.

[17] Zhou Guodong, Su Jian. Named entity recognition using an HMM-based chunk tagger[C]. Proceedings of the 40th Annual Meeting of the ACL, Philadelphia, July 2002:473 - 480.

[18] Olover B, Franz J O, Hermann N. Maximum entropy models for named entity recognition[C]. Proceedings of the Conference on Natural Language Learning at HLT-NAACL. Edmonton, Canada, 2003:148 - 151.

[19] Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets [C]. Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Application (NLPBA). Geneva, Switzerland, 2004:104 - 107.

[20] 王昊,苏新宁. 基于 CRFs 的角色标注人名识别模型在网络舆情分析中的应用[J]. 情报学报,2009(1):88 - 96. (Wang Hao, Su Xinning. Model for person name recognition based on role labeling using CRFs and its application to web opinion Analysis[J]. Journal of the China Society for Scientific and Technical Information, 2009(1):88 - 96.)

[21] Kudo T. CRF ++: Yet another CRF toolkit[OL]. [2011 - 08 - 07]. http://crfpp.sourceforge.net/.

[22] 赵健,王晓龙,关毅,等. 中文名实体识别:基于词触发对的条件随机域方法[J]. 高技术通讯,2006(8):795 - 801. (Zhao Jian, Wang Xiaolong, Guan Yi, et al. Chinese named entity recognition: A CRF approach based on word triggers information[J]. Chinese High Technology Letters, 2006 (8):795 - 801.)

[23] 王昊,邓三鸿. HMM 和 CRFs 在信息抽取应用中的比较研究[J]. 现代图书情报技术,2007(12):57 - 63. (Wang Hao, Deng Sanhong. Comparative study on HMM and CRFs applying in information extraction[J]. New Technology of Library and Information Service, 2007(12):57 - 63.)

邓三鸿 王 昊 南京大学信息管理系副教授,情报学博士。通讯地址:南京市汉口路22号。邮编:210093。

秦嘉杭 南京财经大学副教授,情报学博士、图书馆馆长。通讯地址:南京市栖霞区文苑路3号。邮编:210046。

苏新宁 南京大学信息管理系教授、博士生导师。通讯地址:南京市汉口路22号。邮编:210093。

(收稿日期:2011-08-12)