●章成志 苏新宁

基于条件随机场的自动标引模型研究

摘 要 条件随机场(Conditional Random Fields,CRF)模型是一种概率图模型。为了有效利用标引对象的特征,并考虑到抽词标引可以转换为序列标注问题,本文提出基于条件随机场的自动抽词标引模型。实验结果表明,该模型在改善抽词标引的性能方面,要优于支持向量机、多元线性回归模型等其他机器学习方法,是到目前为止解决序列标注问题的最好方法。但是,该模型本身还不能解决由于样本中存在同义词和相近词带来的问题,需要进一步对训练集和标引过程中存在的词汇语义情况进行考虑,提高标引的质量。图 1。表 3。参考文献 32。

关键词 抽词标引 条件随机场 自动标引

分类号 G252

ABSTRACT CRF (Conditional Random Fields) model is a state - of - the - art sequence labeling method. The CRF model can use the features of documents more sufficiently and effectively. At the same time, keywords extraction can be considered as the string labeling. Keywords extraction model based on CRF is proposed and implemented. Experimental results show that the CRF model outperforms other machine learning methods such as support vector machine, multiple linear regression model etc. in the task of keywords extraction. 1 fig. 3 tabs. 32 refs.

KEY WORDS Keywords extraction. Conditional random field. Automatic indexing.

CLASS NUMBER G252

自动标引包括自动抽词标引与自动赋词标引两种类型。自动抽词标引是一种识别有意义且具有代表性的片段或词汇的自动化技术[1],在文本挖掘领域被称为关键词抽取,在计算语言学领域通常着眼于术语自动识别[2-3]。

关键词是表达文件主题意义的最小单位,大部分对非结构化文件的自动处理,如自动标引、自动摘要、自动分类、自动聚类等,都必须先进行抽词标引,再进行其他处理。可以说,抽词标引是所有文件自动处理的基础与核心技术[4]。

目前大多文档都不具有关键词,而手工标引费时费力且主观性较强,因此自动抽词标引是一项值得研究的技术[4]。自动抽词标引方法可以分为四类,即:①基于统计的方法。该方法不需要复杂的训练过程,简单易行,主要途径有 N-Gram[5]、词频[6]、TF * IDF[7]、字同现[8]、词共现[9]、PAT-tree[10]及特征组合[11]等。②基于语言学的方法。主要从词法分析[12]、句法分析[13]、语义分析[14-15]及篇章分析[16-17]等角度进行抽词标引。③基于机器学习的方法。通过对训练数据进行训练获得统计参数,进行样本的抽词标引,如 NB[18]、最大熵模型[4]、SVM[14]、Bagging[13]等,相关系统如 GenEx[19]、KEA[20]等。④其他方法。即上述方法的综合运用或集成一些启发式知识,如词位置[21]、词长、词排版规则、html 标记[22-23]等。

目前的自动抽词标引方法,大多不能有效利用文本中包含的多个特征,离实用化还有一定距离。为了有效利用标引对象的特征,并考虑到抽词标引可以转换为序列标注问题,本文提出基于条件随机场的自动抽词标引模型。实验结果表明,条件随机场模型在改善自动抽词标引的性能方面要优于支持向量机、多元线性回归模型等其他机器学习方法。

1 基于条件随机场的抽词标引模型

1.1 CRF 抽词标引模型的引入

1.1.1 CRF 简介

条件随机场(Conditional Random Fields,CRF)模型是一种概率图模型[24]。标注序列的结构可以看作一般的无向图。

对于一组长度为 n 的观察序列 $X(X_1X_2\ldots X_n)$,输出状态序列 $Y(Y_1Y_2\ldots Yn)$ 的概率定义如下:

$$P(Y|X)=\frac{1}{Z_X}\exp\left(\sum_i\sum_j\lambda_j f_j(y_{i-1},y_i,X,i)\right) \quad (1)$$

其中,Z_X 为归一化常量,它使得所有的状态序列的概率和为 1,Z_X 的计算公式如下:

$$Z_X=\sum_y\exp\left(\sum_i\sum_j\lambda_j f_j(y_{i-1},y_i,X,i)\right) \quad (2)$$

其中 $\sum_j\lambda_j f_j(y_{i-1},y_i,X,i)$,是对整个观察序列 X,标记位于 i 和(i-1)的特征函数。$\lambda=(\lambda_1,\ldots,\lambda_m)$ 是特征函数对应的权重。标注任务就是搜索概率最大的 Y *,使得:

$$Y^*=\arg\max_y P(Y|X) \quad (3)$$

训练时根据最大熵原理,对训练数据选择λj使$L(\lambda_1,\ldots,\lambda_m)$最大。一般通过迭代算法(GIS)或梯度下降求出逼近最优参数的近似解。

CRF的优点是能有效整合多种特征,即使有些特征之间存在交叉现象,CRF还是能发挥很好的性能。CRF能方便地在模型中包含领域知识,并且较好地解决了标注偏置问题[25]。

1.1.2 利用CRF进行抽词标引的出发点

人工标引要标引文本主题,首先必须对文本内容特征进行分析,确定需要揭示的主题概念。主题分析主要依据文本篇名、前言、目次、文摘及参考文献等,必要时可浏览正文[26]。

通常,每篇文章提取关键词3~5个,以题内关键词为主。题外关键词从文章的摘要、文章开头、分段标题、每段的开头和结尾、文章尾段等部分中提取。标引员进行标引时,一般会通览全文,对文章内容进行高度概括,提炼出能反映文章主题内容的关键词,有的时候还要借助于主题词表进行赋词标引。从关键词的选择过程可以看出,关键词的标引可看成文本的序列标注,即:对待标引文本,依据词语的各种特征,进行是否为关键词的判断。

在抽词标引中,概率图模型可以表达标注序列的结构。用随机变量序列$X(X_1X_2\ldots X_n)$表示待标注的语序列,$Y(Y_1Y_2\ldots Y_n)$表示可能的关键词标注序列,取样本集合为$\{(x_1y_1,x_2y_2,\ldots,x_ny_n)\}$,$x_i(1\leqslant i\leqslant N)$表示一个将被进行关键词标注的词,$y_i(1\leqslant i\leqslant N)\in\{-1,+1\}$,其中"-1"表示非关键词,"+1"表示关键词。本文利用CRF++[27]作为CRF关键词序列标注工具。

1.2 基于CRF模型的抽词标引方法

利用CRF模型进行抽词标引,要经历标引语料的训练与新样本的标注两个基本过程。本节对基于CRF模型的抽词标引方法进行详细描述。

1.2.1 CRF模型训练集的获取

本文以人大报刊复印资料[28]"人大2005年一季度经济类专题"库中经济类600篇论文作为数据集进行基于CRF的自动标引研究。数据集中的论文包括题名、摘要、关键词、段落和章节、图表标题信息以及参考文献等部分。这些数据具有丰富的语言特征,很适合用于研究序列标注问题。

1.2.2 CRF模型特征选择

根据影响关键词标注的各种因素,并且考虑到系统运行的时间和空间消耗问题,本文主要把特征窗口限制在当前词的前后两个词的大小,定义特征空间如表1所示。

(1) 全局特征。主要包括词语的词频与逆文档频率,词语首次出现位置及其全局位置特征。全局位置特征即当前词是否在文章题名、文章摘要、章节图表标题、文章第一段、文章最后一段位置出现过,根据词语在这些位置出现与否,全局位置特征分别取值为"0"或者"1"。

(2) 局部特征。主要包括当前词的词性、当前词的长度、出现的位置、前后两词的词性、前后两词的长度、前后两词出现的位置等。本文利用SegTag分词程序[29]进行分词和词性标注,而出现的位置有t(题名位置),a(文摘位置),c(正文位置),根据当前词所处的位置分别取"0"或者"1"。

(3) 混合特征。主要包括前后两词的词频与逆文档频率;前后两词的首次出现位置,前后两词是否在题名、文摘、章节图表标题、第一段、最后一段等位置出现。

表1给出了CRF模型及后面将介绍的其他5个标引模型中所用到的特征及特征的计算方法。

表1 抽词标引样本特征表

特征类型	特征序号	特征表示	特征意义	特征归一化处理方法
局部特征	1	Word	词语本身	Word
	2	Len	词语的长度	$\frac{Len(Word)}{Max\_Len}$
	3	POS	词或短语的词性,词或短语中的每个词的词性根据其是否含有名词成分,取值为0或1。	$\frac{\sum Phrase(Word_{ij})}{j}$
局部特征	4	t	词语当前所处位置是否为题名位置	{0,1}
	5	a	词语当前所处位置是否为文摘位置	{0,1}
	6	c	词语当前所处位置是否为正文位置	{0,1}

续表

特征类型	特征序号	特征表示	特征意义	特征归一化处理方法
全局特征	7	TF * IDF	出现频数 * 逆文档频率	$\frac{Freq(Word)}{Max\_Freq} \times \log_2 \frac{N+1}{n+1}$
	8	DEP	首次出现位置	$\frac{\#(Word)}{\sum word_i}$
	9	Title(简称 T)	词语是否在题名中出现过	{0,1}
	10	Abstract(简称 A)	词语是否在文摘中出现过	{0,1}
	11	Heading(简称 H)	词语是否在章节标题中出现过	{0,1}
	12	First_Para(简称 F)	词语是否在第一段中出现过	{0,1}
	13	Last_Para(简称 L)	词语是否在最后一段中出现过	{0,1}
	14	Reference(简称 R)	词语是否在参考文献中出现过	{0,1}
混合特征	15 - 16	TF * IDF_Pre_N(N = 1,2)	当前词语的前两词的 TF * IDF	同 TF * IDF 方法
	17 - 18	TF * IDF_Next_N(N = 1,2)	当前词语的后两词的 TF * IDF	同 TF * IDF 方法
	19 - 20	Dep_Pre_N(N = 1,2)	前两词的首次出现位置	同 DEP 方法
	21 - 22	Dep_Next_N(N = 1,2)	后两词的首次出现位置	同 DEP 方法
	23 - 24	Title_Pre_N(N = 1,2)	前两词是否在 Title 中出现	{0,1}
	25 - 26	Title_Next_N(N = 1,2)	后两词是否在 Title 中出现	{0,1}
	27 - 28	Abs_Pre_N(N = 1,2)	前两词是否在 Abstract 中出现	{0,1}
	29 - 30	Abs_Next_N(N = 1,2)	后两词是否在 Abstract 中出现	{0,1}

1.3　几个对照的抽词标引模型

基于机器学习的抽词标引的思想,就是将自动抽词标引看成一种分类问题。本节简要介绍已有的几个用于抽词标引的分类模型(后文简称"标引模型"),即支持向量机模型(Support Vector Machines, SVM)、多元线性回归模型(Multiple Linear Regression, MLR)、Logistic 回归模型(Logistic Regression,简写为 Logit)。此外,本文还给出两个常规的自动标引方法作为基准进行比较分析,分别记为基准模型 1、2(BasaLine1、BaseLine2,简写为 BL1、BL2)。

1.3.1　SVM 模型

SVM 由 Vapnik 在 1995 年提出,用于解决二值分类模式识别问题[30]。2004 年,曾华军等人在进行搜索结果聚类研究时,曾利用 SVM 进行显著短语的提取[31]。2006 年,张阔提出基于 SVM 的自动标引模型[14]。本文拟采用 SVM^{light}[32]进行自动标引。

1.3.2　多元线性回归模型

线性回归是最简单的回归形式。曾华军等人曾利用多元线性回归模型进行显著短语的提取,他们通过实验发现,在解决显著短语提取这一问题上,多元线性回归模型能取得较好的结果[31]。本文拟采用多元线性回归模型作为自动标引模型的一个参照。

1.3.3　Logistic 回归模型

当因变量为二值类型时,Logistic 回归更加适合样本标记的预测。曾华军等人也利用 Logistic 回归模型进行过显著短语的提取,并通过实验发现,在解决显著短语提取这一问题上,Logistic 回归模型也能取得较好的结果[31]。

1.3.4　基准模型 1

基准模型 1 中,将归一化的词语的词频(TF)、归一化的逆文档频率(IDF)、词语长度三个特征作为考虑词语权重的因素,权重计算公式如下:

$$Score = TF * IDF * Len \quad (4)$$

1.3.5　基准模型 2

与基准模型 1 不同,基准模型 2 除了考虑归一化的词语的词频(TF)、归一化的逆文档频率(IDF)、词语长度三个特征外,还加入了词语的首次出现位置(Dep)作为权重因素,权重计算公式如下:

$$Score = TF * IDF * Len * DEP \quad (5)$$

1.4 抽词标引模型训练集的建立

本节对MLR、Logit、SVM、CRF这四个需要训练的模型分别建立了训练集。由于MLR与Logit模型训练方法类似,因此放在一起描述。

1.4.1 MLR与Logit抽词标引模型训练集

对数据集进行特征选择,并根据表1中的特征归一化计算方法,得到每个文本的词语的归一化特征值,最后得到模型的训练集。定义关键词标注符号类型为:{+1,-1},其中"+1"表示当前词语在当前文本中是关键词,如"补贴政策 +1","-1"表示当前词语在当前文本中不是关键词,如"航空工业 -1"。

1.4.2 SVM抽词标引训练集的建立

本文利用SVM^{light}作为SVM分类工具,根据SVM^{light}的训练数据格式,对数据集进行特征选择,并根据表1中的特征归一化计算方法,得到每个文本的词语的归一化特征值,最后得到模型的训练集。

1.4.3 CRF抽词标引训练集的建立

本文利用CRF++[27]作为CRF关键词序列标注工具,根据CRF++的训练数据格式,对数据集进行特征选择,并根据表1中的特征归一化计算方法,得到每个文本的词语的归一化特征值,最终得到模型的训练集。

关键词识别可以通过特征标注来实现,即:对每个词语赋予一个特殊标记符号。该实验中,标注符号的构成主要考虑关键词的边界情况,共定义6种符号类型,分别为:{KW_Y,KW_B,KW_I,KW_O,KW_N,KW_S}。其中,KW_Y表示当前词语在当前文本中是关键词;KW_B表示当前词语在当前文本中是关键词的一部分,并且为关键词的开始;KW_I表示当前词语在当前文本中是关键词的一部分,且为关键词除了首词之外的部分;KW_O表示当前词语在当前文本中的关键词重叠情况,即:该词作为一个关键词的KW_I,又可作为下一个关键词的KW_I;KW_N表示当前词语在当前文本中不是关键词;KW_S表示当前词语是停用词。表2给出CRF抽词标引模型的训练集样例。

表2 CRF抽词标引模型训练集样例

Word	POS	t	a	c	TF*IDF	Len	DEP	T	A	H	F	L	R	Lable
贸易投资	1	1	0	0	0.0915	0.5714	0.0387	1	1	1	1	1	1	KW_B
一体化	1	1	0	0	0.0541	0.4286	0.0548	1	1	1	1	1	1	KW_I
与	0	1	0	0	0.0002	0.1429	0.0671	1	0	1	1	0	1	KW_S
就业增长	1	1	0	0	0.0265	0.5714	0.0775	1	0	1	0	0	0	KW_N
——	0	1	0	0	0.0022	0.2857	0.0866	1	0	0	0	0	0	KW_S
以	0	1	0	0	0.0006	0.1429	0.0949	1	0	0	0	0	0	KW_S
江苏省	1	1	0	0	0.0325	0.4286	0.1025	1	1	1	0	0	1	KW_N
为	0	1	0	0	0.0001	0.1429	0.1096	1	1	0	0	1	0	KW_S
案例	1	1	0	0	0.0077	0.2857	0.1162	1	0	0	0	0	0	KW_N
的	0	1	0	0	0.0000	0.1429	0.1225	1	1	1	1	1	1	KW_S
实证分析	1	1	0	0	0.0128	0.5714	0.1285	1	1	1	0	0	1	KW_Y

1.5 抽词标引模型的测试

对于基准模型1和基准模型2,本文取权重最大的前6个词作为标引关键词。根据MLR与Logit模型训练集分别进行多元线性回归和Logistic回归,得到基于MLR的和基于Logit回归的关键词权重计算公式。根据这两个公式,分别计算测试文本里每个词的权重,权重最大的前6个词作为标引关键词。用同样的方法对测试样本进行格式化,并根据SVM^{light}中的测试程序对待标引文本中的关键词候选集进行自动分类,将分类结果为"+1"的词语作为关键词。同理,利用CRF++的测试程序对待标引的文本进行词语序列的关键词标注,将标注为"KW_Y"的词语、"KW_B KW_I"词语序列或"KW_B KW_O KW_I"序列中"KW_B KW_O"与"KW_O KW_I"词语序列作为

关键词。例如,将标注序列"集群 KW_B 剩余 KW_O 剩余 实现 KW_I"中的"集群剩余"与"剩余实现"提取为关键词。

2　实验结果分析与讨论

2.1　标引模型的性能评价指标

假设测试集中词语总数为 n,自动标引系统标引结果如表 3 所示。人工标引的结果分为两种情况,分别为人工标引为关键词的情况(即:(a + c))与人工标引为非关键词的情况。人工标引为非关键词的情况,就是将人工标引关键词后文本剩下的词作为非关键词。同理,自动标引结果也可以分为这两种情况,其中,(a + b)为标引系统标引的关键词总数。

表 3　标引结果评价列联表

	人工标引为关键词	人工标引为非关键词
标引系统标引为关键词	a	b
标引系统标引为非关键词	c	d

本文主要利用信息检索领域经典的评价方法,即查准率(P)、召回率(R)以及 F1 值对标引模型的标引性能进行评价,对自动标引模型进行 10 折交叉验证,指标计算方法如下:

$$P=\frac{a}{a+b} \tag{6}$$

$$R=\frac{a}{a+c} \tag{7}$$

$$F_1(P,R)=\frac{2PR}{P+R} \tag{8}$$

2.2　标引模型的标引性能比较

通过通用词典与经济类词典对数据集进行分词,分别利用 BasaLine1、BasaLine2、MLR、Logit、SVM、CRF 这 6 个模型进行抽词标引实验,并利用 10 折交叉验证方法进行标引性能的评估。本实验的数据集规模为 600 个样本,分成 10 组,每组 60 个样本,进行交叉验证。图 1 中(a)、(b)、(c)分别为 6 个模型自动标引结果的 P、R、F_1 值结果比较图。

图 1(c)表明,综合 P 值和 R 值,CRF 模型的标引性能是最好的。CRF 模型标引结果的 F_1 值达到 0.5125,表明 CRF 模型用于自动抽词标引上,能提高自动标引性能。SVM 模型次优,其余的为:Logit > MLR > Baseline2 > Baseline1。

从自动标引的查准率来看,SVM 和 CRF 模型要大大优于其余四种标引模型。SVM 的查准率高达 80.17%,CRF 的查准率为 66.37%,Logit、MLR、BaseLine2、BaseLine1 的准确率分别为:32.48%、31.74%、27.78%、23.43%。

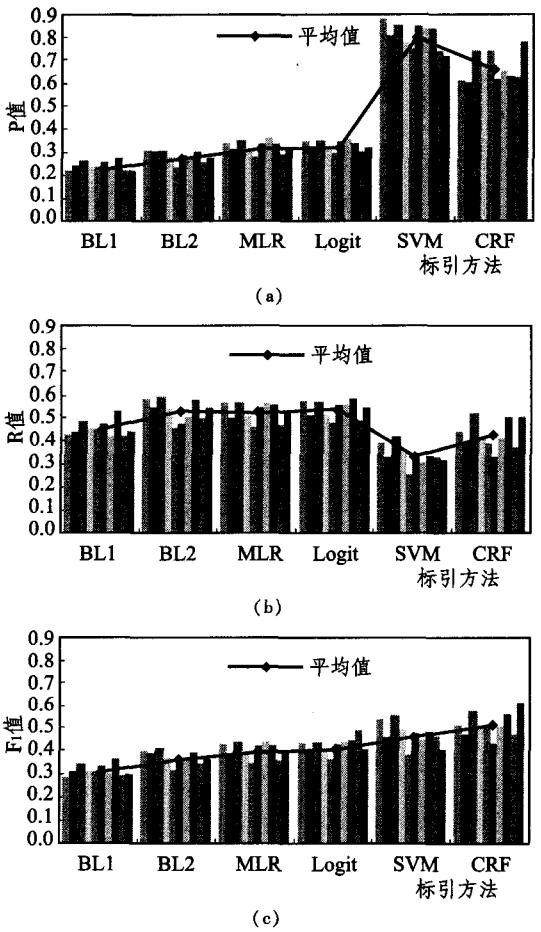

图 1　标引结果的 P 值(a)、R 值(b)、F_1 值(c)比较

图 1(a)表明 MLR 的 P 值稍高于 Logit,说明 Logistic 回归处理关键词标引这种二值分类问题,从查准率上来看,稍强于多元回归。BaseLine2 模型的 P 值高出 BaseLine1 模型 4.35%,说明了以 DEP 作为词语权重计算因素,能提高自动标引的查准率。图 1(b)表明 SVM 与 CRF 的 R 值都低于其余四个模型,Logit 模型的召回率最高,其次是 MLR 模型。

值得注意的是,在实验中我们发现,由于数据集中的关键词存在赋词标引情况,抽词标引性能比较依赖于训练语料。如何自动获得比较"干净"的抽词标引语料是个很有意义的研究课题。

2.3　CRF 抽词标引错误原因分析

我们对 CRF 模型标引的结果进行考察,发现错

误情况主要包括两个方面:①训练集存在的问题。在训练集中,本文没有考虑关键词存在同义词和相近词这一问题。例如,原文给出的关键词有"牧民",无"牧户",而正文中有"牧户"等同义词或近义词,在训练集则作为非关键词,这样会影响CRF等统计模型的训练准确度。②标引过程中存在的问题。由于测试样本会存在同义词或相近词的情况,而CRF等统计模型本身无法解决这些问题,降低了标引的质量。因此,今后进一步的工作是对训练集和标引过程中存在的词汇语义情况进行考虑,提高标引的质量。

3 结语

条件随机场模型是到目前为止解决序列标注问题的最好方法。由于CRF模型能有效融合多种信息,因此该方法是一种很有应用前景的方法,对提高自动抽词标引的实用化程度有积极的意义。下一步工作包括:利用词语间的语义关系提高抽词标引的质量;对网页、电子邮件等其他类型的文档进行抽词标引及应用研究;采用标准的语料库进行中英文文本抽词标引性能测试等。

参考文献:

[1] 曾元显. 关键词自动提取技术与相关词反馈[J]. 中国图书馆学会会报, 1997(59): 59-64.

[2] 王强军, 李芸, 张普. 信息技术领域术语提取的初步研究[J]. 术语标准化与信息技术, 2003(1): 32-33, 37.

[3] Xun E, Huang C, Zhou M. A Unified Statistical Model for the Identification of English baseNP[C]. In: Proceedings of 4th ACM Conference on Digital Libraries,. Beakeley, CA, USA, 2000: 254-255.

[4] 李素建, 王厚峰, 俞士汶, 辛乘胜. 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.

[5] Cohen J D. Highlights: Language and Domain-independent Automatic Indexing Terms for Abstracting[J]. Journal of the American Society for Information Science, 1995, 46(3): 162-174.

[6] Luhn H P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information[J]. IBM Journal of Research and Development, 1957, 1(4): 309-317

[7] Salton G, Yang C S, Yu C T. A Theory of Term Importance in Automatic Text Analysis[J]. Journal of the American society for Information Science, 1975, 26(1): 33-44.

[8] 马颖华, 王永成, 苏贵洋, 张宇萌. 一种基于字同现频率的汉语文本主题抽取方法[J]. 计算机研究与发展, 2004, 40(6): 874-878.

[9] Matsuo Y, Ishizuka M. Keyword Extraction from a Single Document Using Word Co-ocuurrence Statistical Information[J]. International Journal on Artificial Intelligence Tools, 2004, 13(1): 157-169.

[10] Chien L F. PAT-tree-based Keyword Extraction for Chinese Information Retrieval[C]. In: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR1997), Philadelphia, PA, USA, 1997: 50-59.

[11] 张庆国, 薛德军, 张振海, 张君玉. 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

[12] Ercan G, Cicekli I. Using Lexical Chains for Keyword Extraction[J]. Information Processing and Management, 2007, 43(6): 1705-1714.

[13] Hulth A. Improved Automatic Keyword Extraction Given More Linguistic Knowledge[C]. In: Proceedings of the 2003 Conference on Emprical Methods in Natural Language Processing, Sapporo, Japan, 2003: 216-223.

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

[15] 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.

[16] Dennis S F. The Design and Testing of a Fully Automatic Indexing-searching System for Documents Consisting of Expository Text[C]. In: G. Schecter eds. Information Retrieval: a Critical Review, Washington D. C.: Thompson Book Company, 1967: 67-94.

[17] Salton G, Buckley C. Automatic Text Structuring and Retrieval-Experiments in Automatic Encyclopaedia Searching[C]. In: Proceedings of the Fourteenth SIGIR Conference, New York: ACM, 1991: 21-30.

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

[19] Turney P D. Learning to Extract Keyphrases from Text[R]. NRC Technical Report ERB-1057, National Research Council, Canada. 1999: 1-43.

(下转第99页)

表》中的专指主题词和附属主题词,把其他参与兼容的分类法横向展示。将参与兼容的分类表类号与《中图法》的类号相对照,列出其等值兼容或近似兼容的概念。最终形成一个以《中分表》为核心的兼容体系,实现不同词表之间的兼容互换。

参考文献:

[1] 刘华梅. 基于情报检索语言互操作技术的集成词库构建研究——以教育词库为例[D]. 南京:南京农业大学信息管理系,2006.

[2] Marcia Lei Zeng, Lois Mai Chan. Trends and issues in establishing interoperability among knowledge organization systems[J]. Journal of the American Society for Information Science and Technology, 2004,55(5): 377 -395.

[3] 刘华梅,侯汉清. 近十年情报检索语言互操作研究进展[J]. 图书馆理论与实践,2006 (4):31 -33.

[4] 傅兰生. 我国叙词兼容两大方案的分析——兼论国家级叙词兼容词库的建立[J]. 情报学报,1991,10(4):257 -264.

[5] 朱岩. "国家叙词库"建库设计与分析[J]. 情报理论与实践,1991(4):28 -30.

[6] 朱岩. 对建立国家叙词库的认识与思考[J]. 科技情报工作,1991(2):15 -17.

[7] 洪漪. 我国国家叙词库建设中几个问题的探讨[J]. 情报学刊,1991,14(3):209 -211.

[8] 方陆明. 利用电子计算机建立农业叙词库及其管理系统——兼谈机编叙词表的几个问题[J]. 农业图书情报学刊,1989(1):61 -65.

[9] 侯汉清. 建立以《中国分类主题词表》为核心的检索语言兼容体系[J]. 北京图书馆馆刊,1998(4):35 -39.

[10] 张雪英,侯汉清. 叙词表词汇转换系统的设计[J]. 情报学报,2000,19(5):451 -457.

[11] 陆勇. 面向信息检索的汉语同义词自动识别[D]. 南京:南京农业大学信息管理系,2005.

刘华梅　南京农业大学信息管理系硕士研究生毕业,现在国家图书馆工作。通讯地址:北京中关村南大街33号。邮编100081。

侯汉清　南京农业大学教授,博士生导师。通讯地址:南京农业大学信息科技学院。邮编210095。

(收稿日期:2007-11-20)

(上接第94页)

[20] Witten I H, Paynter G W, Frank E, Gutwin C, Nevill-Manning C G. KEA: Practical Automatic Keyphrase Extraction[C]. In: Proceedings of the 4th ACM Conference on Digital Library (DL'99), Berkeley, CA, USA, 1999: 254 -256.

[21] 韩客松,王永成. 中文全文标引的主题词标引和主题概念标引方法[J]. 情报学报, 2001, 20(2): 212 -216.

[22] Keith Humphreys J B. Phraserate: An Html Keyphrase Extractor[R]. Technical Report, University of California, Riverside, 2002.

[23] 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, 24(1): 87 -92.

[24] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segementing and Labeling Sequence Data[C]. In: Proceedings of the 18th International Conference on Machine Learning (ICML01), Williamstown, MA, USA, 2001: 282 -289.

[25] 周俊生,戴新宇,尹存燕,陈家骏. 基于层叠条件随机场模型的中文机构名自动识别[J]. 电子学报, 2006, 34(5): 804 -809.

[26] 侯汉清,马张华. 主题法导论[M]. 北京: 北京大学出版社, 1991: 199.

[27] CRF + +: Yet Another CRF toolkit[OL]. [2005-12-20]. http://chasen.org/ ~taku/software/CRF + +.

[28] 人大报刊复印资料[OL]. [2007-12-01]. http://art.zlzx.org.

[29] 中文自然语言处理开放平台[OL]. [2005-12-25]. http://www.nlp.org.cn.

[30] Vapnik V. The Nature of Statistical Learning Theory [J]. New York: Springer-Verlag, 1995: 1 -175.

[31] Zeng H J, He Q, Chen Z, Ma W Y, Ma J. Learning to Cluster Web Search Results[C]. In: Proceedings of 27th Annual International Conference on Research and Development in Information Retrieval (SIGIR04), Sheffield, 2004: 210 -217.

[32] SVMlight[OL]. [2005-12-20]. http://svmlight.joachims.org.

章成志　南京理工大学信息管理系讲师,博士,中国科技信息研究所博士后。通讯地址:南京理工大学信息管理系。邮编210094。

苏新宁　南京大学信息管理系教授,博士生导师。通讯地址:南京大学信息管理系。邮编210093。

(收稿日期:2008-01-10)