●王 昊，苏新宁（南京大学 信息管理系，江苏 南京 210093）

基于模式匹配的中文通用本体概念抽取模型

摘 要：本文重点阐述了中文通用本体概念实例的机器抽取过程，建立了基于模式匹配的通用本体概念识别模型，以此作为领域本体自动构建的基础。此后探讨了该模型在情报学各研究领域中的应用前景，并通过实验检验模型在实际应用中的识别效果。

关键词：通用本体；概念抽取；模式匹配；隐马尔可夫模型

Abstract: This article mainly expatiates the instance extraction process of Chinese general ontology concept by machine, and constructs a model for identification of general ontology concept based on pattern matching, which is the basis for automatic building of domain ontology. The article also discusses the application prospect of this model in various research fields of information science, and verifies the identification effect of this model in practical application by experiment.

Keywords: general ontology; concept extraction; pattern matching; hidden Markov model

自 Berners-Lee 和 Fischetti 在 1999 年首次提出语义网的概念[1]，语义网作为一个使用语义元素描述信息，以满足智能代理对异构分布信息的有效访问和准确检索的公开环境越来越受到人们的关注。本体机制则是语义网中实现人机语义交互、解决语义异构问题的关键技术。

本体提供丰富原语描述领域的概念模型，能够对领域知识进行推理和验证。一般认为，按照对领域的依赖程度，本体可分为通用（General）或顶层（Top-level）、领域（Domain）、任务（Task）、应用（Application）本体。目前在国内情报学领域，关于本体的研究主要集中于领域本体的构建和应用。其中领域本体的构建则尝试采用手工方式，如武汉大学董慧教授等研究历史领域本体构建[2]，或基于现存叙词表构建领域本体[3-4]，在本体概念获取以及概念关系建立等方面都缺乏自动化手段，无法满足快速有效构建或修改本体的要求。要实现中文本体学习自动化，首先要解决中文通用本体概念实例的识别问题。通用本体是指最普遍的概念及概念之间的关系，是其他类型本体的基础，经常引入到其他类型本体中以属性的形式与其他类型概念保持关联。Guarino 认为，通用本体包括空间、时间、物体、性质、行为、事件等概念及其关系[5]。通用本体概念的识别目前主要有两种途径：

1）基于词典获得登录术语。基于词典采用正向最长匹配分词算法，获得词典词，再根据隐马尔可夫模型（Hidden Markov Model，HMM）[6]或反向动态规划和正向 A * 解码算法相结合[7]进行词性标注，实现词典词的概念归类。

2）未登录命名实体的识别。命名实体（Named Entity，NE）是指文本中具有特定意义的专有名称和数量短语，主要包括数字表达式、日期表达式、人名、地名和机构名。命名实体是基本的信息元素，能够准确指示文本内容（语义），是通用本体概念的重要组成部分。目前中文命名实体（CNE）识别主要有 3 种方法：①基于规则的方法[8-10]：即人为构建能够表达中文命名实体组成规律的规则知识，识别文本中满足指定规则的命名实体。②基于统计的方法：对经过标注的语料进行训练，获得语言学规律知识，采用一定的统计算法，如 HMM、最大熵模型（Maximum Entropy Model，MEM）[11]、决策树（Decision Tree，DT）[12]和条件随机场（Conditional Random Fields，CRFs）[13]等，识别满足指定阈值的命名实体。③规则和统计相结合的方法：规则需要人为编制，依赖具体的领域，命名实体的开放性和随意性使得规则难以构建。而统计方法则在很大程度上受训练语料规模和领域的影响。因此，更好的做法是结合规则和统计[14-16]，提高命名实体识别的准确性。本文研究重点在于阐述中文通用本体概念实例中未登录命名实体的抽取过程，建立基于模式匹配的通用本体概念抽取模型（Concepts Extraction Model based Pattern Match，PMCEM），为中文本体概念识别提供自动化手段，作为领域本体自动构建以及情报学其他领域研究的基础。最后通过实验探讨模型识别效果。

1 基于模式匹配的概念抽取模型

笔者在 Guarino 认识的基础上进一步将通用本体归纳

为时间、空间、自然人、物体、行为、性质等6种类型概念，并提出一种基于模式匹配的概念抽取模型（PMCEM）用于识别中文文本中属于上述6种概念的术语，PMCEM主要由4个匹配模块组成：基于规则匹配，基于词典匹配，基于HMM链匹配以及基于词性匹配。模型基本算法如图1所示。

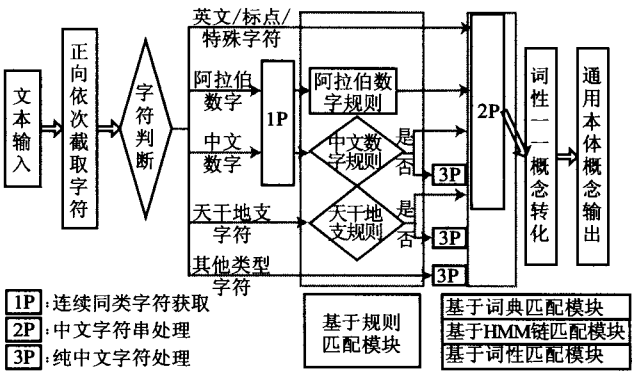

图1 基于模式匹配的通用概念抽取模型（PMCEM）

其中：①1P表示连续同类字符获取：从测试字符开始，抽取与测试字符同类的连续字符串。②2P表示中文字符串处理：将未处理词串粗分，采用Viterbi算法确定粗分词角色，再根据命名实体构成模式在角色链中进行实体匹配，最后根据HMM实现词性标注和歧义处理。③3P则表示纯中文字符处理：将测试字符或字符串加入到未处理中文字符串中等待处理。

2 基于规则匹配的CNE获取

2.1 规则集的构建

一个完整的数字表达式由4部分构成：前置约数＋数字核心词＋后置约数＋约数后缀。根据上述模板，以数字核心词即连续阿拉伯数字或连续中文数字为中心，寻找与核心词匹配的前置、后置约数和约数后缀，即可获得数字表达式。中文时间表达式则较为复杂，此处仅讨论未登录时间词的规则匹配。一个完整的时间表达式模板包括：前置时间词＋时间核心词＋后置时间词＋时间后缀。其中时间核心词又可以分为两类：数字表达式和干支表达式。

笔者在对语料库中中文数字时间表达式的语法规则进行总结的基础上，共定义了14类规则字符串，分别是阿拉伯数字、中文数字、阿拉伯数连接符、中文数字连接符、前置约数、后置约数、约数后缀、前置时间词、后置时间词、时间后缀、天干词、地支词、干支后缀和地支后缀。基于这些规则字符串可以创建三大类12条用于匹配数字和时间表达式的规则。

1）数字核心词为阿拉伯数字。①前置规则：前置时间词＋前置约数词，其中任何一部分都是可选。②连续阿拉伯数字规则：阿拉伯数字＋任意位阿拉伯数字或连接符或约束后缀，其中“＋”前部分内容必选，后面部分可选。③后置规则：又可以分为：数字规则：后置约数＋约数后缀，两部分均可选；时间规则：后置时间词＋时间后缀，两部分均可选。④约束规则：即用于排除满足上述规则但不符合中文语义或不成立的歧义短语的特定规则。例如“客户数5年翻4番”中“数5”的识别导致语义错误，“3分45秒”是数字词和量词的组合，而非时间词等。⑤歧义规则：即用于排除满足上述4条规则却和前词或后词形成歧义的命名实体的特定规则。在数字和时间表达式中涉及的中文字词很有可能与其相邻的字词构成词语，形成歧义短语。为此本文规定如果前置规则（或后置规则）和前相邻字词（或后相邻字词）形成词典词，则排除前置规则（或后置规则）。⑥修正规则：将部分识别出来的数字表达式改为时间表达式的特定规则。例如，“初5”、“2006/12/24”等被识别为数字，根据实际情况有必要将其改为时间表达式。

2）数字核心词为中文数字。其匹配规则与阿拉伯数字核心词基本相同，在此仅讨论两者不同的情况。①连续中文数字规则：中文数字＋中文数字或中文数字连接符或约束后缀。②约束规则：对中文数字的约束，如连续中文数字“三三”不符合中文数字语法“三十三”，不能作为中文数字。③单数字规则：中文数字常常出现在词语中，因此规定单个中文数字不在规则匹配阶段识别。

3）核心词为干支表达式。时间表达式的核心词还可能是（天）干（地）支表达式。中国古代经常使用天干地支计时，例如“戊戌政变”、“辛亥革命”、“子时”等等。①干支规则：天干词＋地支词＋干支后缀＋时间后缀。②地支规则：地支词＋地支后缀根据上述规则可以派生出若干具体的匹配模式用于识别词典未登录的含有数字的数字和时间表达式。至于单汉字数字可以在根据词典分词后判断，登录时间词则可以根据词语词性进行识别。此外，上述规则并没有包含所有的数字和时间表达式语法，因此可以根据实际情况对上述规则集进行扩充，从而能够更全面更准确地识别数字和时间表达式。

2.2 规则匹配算法

将上述各类规则进行适当调整、编排以形成合适的规则流，找到规则匹配的最佳计算方法。规则匹配算法的基本思路是：根据前置规则、连续数字规则、后置规则以及干支规则，识别满足规则的所有数字和时间表达式，再应用约束和歧义规则排除不符合中文语法或成文习惯以及歧义短语，最后根据修正规则将部分数字表达式转化为时间表达式。规则匹配具体过程如图2所示。

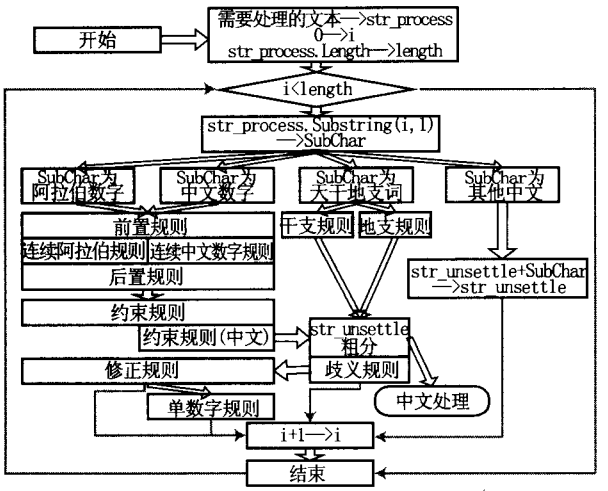

图2 基于规则的数字/时间表达式识别框图

3 基于层叠隐马尔可夫模型的模式化 CNE 抽取

考察语料库中标注好的中国人名、地名以及机构名的构成，发现这类命名实体都是由更小的词语单元所组成，而根据词语单元在命名实体中所承担角色不同以及中国人名、地名和机构名之间的嵌套层次，可以采用层叠隐马尔可夫模型[17]来模拟此类命名实体的识别过程，见图3。

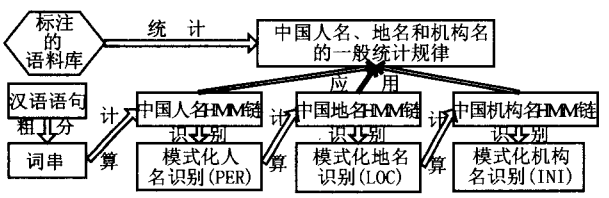

图3 基于层叠隐马尔可夫模型的模式化 CNE 识别模型

3.1 隐马尔可夫模型

隐马尔可夫模型（HMM）是指支持为观察序列从状态集合中选择具有最大可能性状态序列过程的模型。一个隐马尔可夫模型可以记做 HMM = (S, W, T, A, B, π)，其中：①$S = \{s_1, s_2, \cdots, s_n\}$，表示模型中所有可能状态的集合；②$W = \{w_1, w_2, \cdots, w_m\}$，表示模型输入的观察符号序列集合（词串）；③$T = \{t_1, t_2, \cdots, t_m\}$，表示模型输出的与观察序列相对应的状态序列集合（状态串）；④$A = a_{ij}$，表示状态转移矩阵，是指从 i 状态转移到 j 状态的概率；⑤$B = bj(k)$，表示符号状态分布矩阵，是指符号 w_k 处于各种状态下的概率分布情况；⑥$\pi = \{\pi_i\}$，表示初始符号的状态分布概率集合。

HMM 是在 A、B、π 等概率统计量支持下，为词串 W 选择最大概率的状态串 T。为此引入条件概率 $P(T \mid W)$ 表示词串 W 在状态串 T 下的概率，则 HMM 的目的即为求使 $P(T \mid W)$ 值最大的 T，记为：

$$T' = \arg\#\max P(T \mid W) \tag{1}$$

对 $P(T \mid W)$ 使用条件概率公式，并根据贝叶斯（Bayes）公式对其进行化简，可得：

$$P(T \mid W) = \frac{P(T, W)}{P(W)} = \frac{P(T)\,P(W \mid T)}{P(W)} \approx P(T)\,P(W \mid T) \tag{2}$$

根据 HMM 独立性假设和二元语义关系，在公式（2）中：

$$P(T) \approx P(t_1 \mid t_0)\,P(t_2 \mid t_1) \cdots P(t_i \mid t_{i-1}) \cdots P(t_m \mid t_{m-1}) \tag{3}$$

$$P(W \mid T) \approx P(w_1 \mid t_1)\,P(w_2 \mid t_2) \cdots P(w_i \mid t_i) \cdots P(w_m \mid t_m) \tag{4}$$

公式（1）最终可化简为：

$$T' = \arg\#\min \{\sum_{i=1}^{m} [-\ln P(w_i \mid t_i) - \ln P(t_i \mid t_{i-1})]\} \tag{5}$$

其中词语特定状态下概率和状态之间转移概率可以近似地从大规模语料库中统计估算：

$$P(w_i \mid t_i) \approx \frac{\text{训练语料中 } w_i \text{ 的状态为 } t_i \text{ 的次数}}{\text{训练语料中状态 } t_i \text{ 出现的总次数}} = \frac{C(w_i, t_i)}{C(t_i)} \tag{6}$$

$$P(t_i \mid t_{i-1}) \approx \frac{\text{训练语料中状态 } t_i \text{ 之前的状态是 } t_{i-1} \text{ 的次数}}{\text{训练语料中状态 } t_{i-1} \text{ 出现的总次数}} = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})} \tag{7}$$

于是公式（7）即可以转化为：$T' = \arg\#\min \{-\sum_{i=1}^{m} [\ln \frac{C(w_i, t_i)}{C(t_i)} + \ln \frac{C(t_{i-1}, t_i)}{C(t_{i-1})}]\}$ (8)

根据公式（8）可知通过模拟 HMM 可以获得与符号串 W 相对应的概率最高的状态串 T。

3.2 基于核心词典匹配的中文文本粗分

为了使用 HMM 来模拟人名、地名以及机构名的识别过程，首先要做的是将输入的中文文本转化为 HMM 中的词串观察序列（W）。本系统采用基于核心词典（Core Dictionary）的正向最大匹配算法来实现文本粗分。例如“南京大学信息管理系苏新宁教授参加了在南京金陵饭店举行的学术会议”经过文本粗分后就形成了词串序列“南京 大学 信息 管理 系 苏 新宁 教授 参加 了 在 南京 金陵 饭店 举行 的 学术 会议”。笔者就以此句为例说明三类实体的识别过程。

3.3 模式化人名识别

在人名识别 HMM 中，首先根据词串中各词所担任的角色，确定模型中的状态集合，再根据公式（3）和公式（4）对语料库进行训练获得各角色的统计数据，根据公式（8）计算具有最大概率的人名 HMM 链，最后进行人名角色模式匹配，识别中文文本中的人名词串。

1）角色定义。考察语料库中人名构成，发现中国人名在结构上具有一定的规律：一般最多由4个字（姓+姓+名+名）组成。表1即为总结获得的人名识别HMM的构成角色集合。

表1 中国人名识别HMM中的状态集合（构成角色集合）

序号	角色	解释	示例	序号	角色	解释	示例
1	B	姓氏	李世民，东方朔	9	Z	CD成词	李成才，王民航
2	C	双名首字	李世民	10	K	人名的上文	老师称赞胡平
3	D	双名尾字	李世民	11	L	人名的下文	李世民发动玄武门之变
4	E	单名	王昊	12	M	两个人名之间的词	王昊和王婷是兄妹
5	F	姓前缀	小王，老张	13	U	KB成词	教官对白展堂说
6	G	姓后缀	李总，王博	14	V	DL或EL成词	邓颖超生前用过的东西
7	X	BC成词	齐国明，查明权	15	W	其他姓名词	唐太宗
8	Y	BCD或BE成词	罗盘，李四光	16	A	其他词	参加 迎新晚会

2）角色词典生成算法。根据公式（8），需要对训练语料库进行统计，总结中文人名的一般语法规律，以便使用这些统计数据估算观察词串的最佳角色链。需要统计的数据包括：各词担任各角色的次数，各角色在训练语料中出现的总次数，以及角色集合中任意两角色之间相邻（有顺序）次数。具体过程见图4。

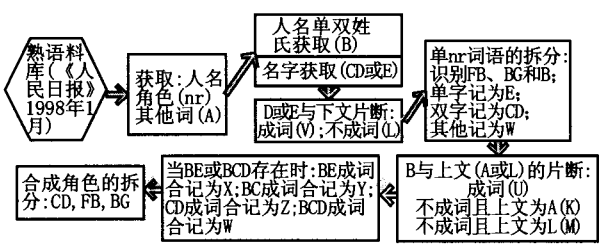

图4 人名角色信息获取过程

最后统计生成的<词语，角色>有序对，得到3个人名识别词典：统计词语w_i作为t_i角色出现的次数$C(w_i, t_i)$作为词语角色词典；累计所有不同角色的出现次数$C(t_i)$作为角色词典；以及统计任意两角色有序相邻的次数$C(t_{i-1}, t_i)$作为角色转移词典。

3）模式匹配算法。前文根据特殊到一般的归纳法，从大规模语料中发现中国人名的一般语法特征；此后即可采用一般到特殊的演绎推理法，认为遵循上述语法特征的角色组合即为中国人名，具体步骤如下：①应用Viterbi算法[18]，计算经过粗分获得的中文词串相对应的最佳（概率最大或公式（8）的值最小）角色串（HMM链），如前例的HMM链为“AAAAABZ-LAAAAAAAAAA”。②U和V角色是人名角色和一般角色的组合，需要将其及其对应的词串词断裂。③根据人名角色定义，认为HMM链中任何匹配BBCD，BBE，BBZ，BCD，BXD，BE，BG，BZ，CD，FB，XD，Y，Z，B，W等角色模式其对应的词语组合即为中国人名。④将原词串的人名词组合，并将其角色标记为PER，以便对新词串进行迭代式地模式化地名识别。即在上例中将“苏”和“新宁”组成一个词，而将其对应HMM链中“BZ”替换为“PER”。

3.4 模式化地名识别

在已经识别人名的文本上进行迭代式地名识别，其基本过程同人名识别类似，然而地名的结构与人名大相径庭，因而其角色构成也大不相同。

1）角色定义。表2列出了构成地名的角色定义。地名识别在人名识别基础上进行，因此人名词只能以PER和I（作为地名词的一部分）两个角色出现，其他词则充当除PER外的其他角色。

表2 中国地名识别HMM中的构成角色集合

序号	角色	解释	示例	序号	角色	解释	示例
1	H	特征词	蓓蕾村，义乌市	6	T	地名连接词	北京和上海都是直辖市
2	I	内部词	大陈镇	7	P	IJ与上文成词	看中国的变化
3	J	I和H成词	鞍山，澳门	8	Q	HI与下文成词	中华门前，昆仑山上
4	R	I的上文	从山东引进	9	A	其他词	瞻仰烈士遗容
5	S	HIJ的下文	台湾是中国的领土	10	PER	人名词	周恩来 邓颖超纪念馆

2）角色词典生成算法。在熟语料库中，地名以ns字符串进行标注。根据图5所示过程生成的<词语，角色>有序对，再对其进行统计，生成类似人名识别词典的3个地名识别词典。

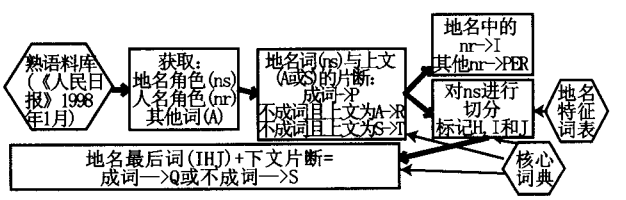

图5 地名识别HMM角色信息获取过程

3）模式匹配算法。角色集合决定角色模式，地名角色模式可以表示为：｛?I，*PER?I｝［*｛?I，*PER?I｝］［｛?J,?H｝］。其中*表示任意多个字符，?表示任意一个字符，［］表示内容可选，｛｝表示选择其

中一项。通过下述过程可实现中文文本中地名实体的识别：地名 HMM 链的生成“ISAAAPERAAARIIHSAAA”→PQ 角色拆分→地名模式匹配→地名模式替换“LOCSAAAPERAAARLOCSAAA”及地名词语组合“南京”“南京金陵饭店”。

3.5 模式化机构名识别

类似地名构成角色的定义，机构名的构成角色集合包括：特征词（h），内部词（i），i 和 h 成词（j），i 的上文（r），h 或 i 的下文（s），机构名连接词（t），i 与上文成词（p），h 或 i 与下文成词（q），其他词（A），人名词（PER）以及地名词（LOC）。其中人名词角色可为 PER 和 i，地名词角色可为 LOC 和 i，而其他词可以是除 PER 和 LOC 以外的其他角色。机构名识别词典生成及其模式匹配算法基本类似于地名识别，不再详述。机构名角色模式如下所示：{? {I, LOC}, *PER? i} [* {? i,? LOC, *PER? i}] [{? j,? h}]。前例的机构名 HMM 链为“LOCiiihPERAAAALOCAAAA”，“南京大学信息管理系”被识别，地名“南京”作为其一部分。

4 基于词性标注的歧义消除

相同中文字词在不同的语义环境中可能呈现不同的词性。以词性集合作为 S（状态）集合，PMCEM 采用 HMM 来模拟中文字词的词性标注[18]，其具体过程类似第 2.2 节介绍的内容。于是任何中文词串都可以生成对应的词性 HMM 链。由于歧义，同一中文片断可能存在多种切分，歧义消除就是选择其中最合理的词串。根据不同词串具有不同的词性 HMM 链的特点，PMCEM 采用计算（根据公式（8））和比较词串的词性 HMM 链值的方法来选择合理词串，消除歧义。笔者仅考虑链长为 1 的交集型歧义消除，表 3 显示了两个中文片段的词性标注，其中链值越小则说明该词性 HMM 链更符合实际情况。

表 3 两个歧义中文片段的词性标注和歧义消除

正向切分	词性 HMM 链	链值	反向切分	词性 HMM 链	链值
研究生 命 的 起源	n-n-u-n	36.602063	研究 生命 的 起源	v-n-u-n	30.638636
这 是非 常 重要 的	r-n-d-a-u	32.928924	这 是 非常 重要 的	r-v-d-a-u	25.905126

HMM 中公式（3）是根据独立性假设和二元语义关系获得，若考虑三元语义关系，即认为任意一个词的状态与其前两个词状态相关。公式（3）可进行修正，如下所示：

$$P(T) \approx P(t_1 \mid t_0)\, P(t_2 \mid t_1, t_0) \cdots P(t_i \mid t_{i-1}, t_{i-2}) \cdots P(t_m \mid t_{m-1}, t_{m-2}) \quad (9)$$

由于三元语义关系更详细揭示了词语所在的语义环境，可获得更好的识别效果。最后 PMCEM 根据词性（或实体标记）——概念转化实现通用本体概念识别：时间实体和时间词作为时间概念，地名和机构名实体作为空间概念，人名实体则作为自然人概念，名词作为物概念，形容词、副词和数词实体等可作为性质概念，动词为行为概念。

5 PMCEM 在情报学各领域中的应用

PMCEM 可以从任意中文文本中抽取空间、时间、物体、性质、行为等作为句子主干词的概念，对于研究中文句法结构、文本语义，乃至构建领域知识网、挖掘文本隐含信息都具有基础作用。

1）本体构建自动化。笔者认为本体构建可以分为 5 个层次：①命名实体识别；②本体要素（Ontology Element）识别；③本体关系（Ontology Relation）识别；④同指关系（Co - reference）处理；⑤本体模板（Ontology Template）。它是根据应用目标定义的本体框架，用于特定领域的信息识别和组织。PMCEM 识别通用本体概念，实际上是完成了本体构建的前两个层次，是本体概念关系识别的基础。利用概念在文本中的相对位置同时应用数据挖掘、数理统计、模式匹配等信息技术可以自动建立概念之间的直接关系，推理概念隐藏关系。具体的做法如下：①应用 PMCEM 识别中文文本中的概念集合（Concept Set）。②建立本体模型，定义本体的属性（类属性），作为概念关系建立的模板。③从标准、规范的领域词典中获取所需的概念关系模式，构建模式库（Mode Set）。④判断概念集合中的概念之间是否存在模式库中规则所表示的关系，从而构建本体概念关系。

2）智能信息检索。PMCEM 中提到的命名实体识别及其在本体构建中的应用可以从两个方面提高信息检索准确率，在一定程度上实现智能信息检索：①在基于关键字匹配的全文检索中，引入命名实体可以适当限制检索范围，提高检准率。例如 PMCEM 能够识别检索式“中兴集团的 IPTV”中的机构实体“中兴集团”，避免将“中”、“兴”、“集团”等无关词作为检索关键字。②本体标引和本体检索是智能信息检索重要研究方向，PMCEM 作为本体自动构建的基础，能够在一定程度上促进本体构建技术的研究，从而促进检索技术从词语匹配到语义匹配过渡。

3）个性化推荐。个性化推荐是指信息所有者根据用户的个性化需求，应用现代信息技术以一定方式主动向用户推荐其感兴趣的信息。个性化推荐需要解决的问题有：如何获得用户个性化需求？采用何种方式（或推荐模式）主动向用户推荐信息？如何判断信息适合用户的个性化需求，即信息分类聚类？挖掘用户访问日志是获得用户个性化需求的主要方式。通过 PMCEM 可以从用户访问日志中

获得用户曾经访问的各种概念（如人物、企业等），再通过统计学方法确定用户感兴趣对象，从而推测用户的个性化需求；其次个性化推荐过程中涉及的信息分类聚类，也是以概念切分为基础的。

4）竞争情报系统。竞争情报系统是指实现从公开环境中获取信息，经过加工处理、推理论证，形成竞争情报分析报告提供给用户，支持用户决策这一整个过程的信息系统。竞争情报系统涉及人物情报、单位情报、产品情报等，需要对大量的人名、机构名、数据、时间等实体进行处理，而 PMCEM 能够对这些信息进行比较准确的识别，提高情报分析的自动化程度。例如在 PMCEM 对数字实体识别的基础上，结合数值单位，可以形成数值知识元等。

6 结束语

前面笔者对 PECEM 系统的结构、涉及的关键技术以及系统具体应用作了深入的分析。为了考察系统的实际应用价值，笔者以《人民日报》1998 年未经标注语料库中 600 余条数据作为封闭性测试对象，以①准确率（P）= 系统识别正确的概念总数/系统能够识别出来的概念总数；②召回率（R）= 系统识别正确的概念总数/测试集中出现的正确的概念总数；③F 值 $= 2 \cdot P \cdot R/(P+R)$ 作为测试指标，评价 PMCEM 系统对各类概念的识别效果（见表 4）。

表 4 PECEM 系统识别各类通用本体概念的封闭性测试结果

概念类型	时间	空间	自然人	物	性质	行为
识别概念数	516	1 059	824	6 705	3 758	6 330
准确率（P）	99.81%	88%	93.93%	96.96%	99.28%	100%
召回率（R）	100%	91.59%	95.44%	97.35%	94.72%	96.3%
F 值	99.90%	89.76%	94.68%	97.15%	96.95%	98.12%

并可获得以下结论：①在不考虑代词指代的情况下，各类概念识别的 F 值均达到或接近 90%，具有一定的实用价值。②空间概念即地名以及机构名的识别效果不太理想，需要进一步完善算法提高识别准确率。③概念之间相互关联，一类概念的误识别也会导致其他类型概念的识别错误，因此在提高 F 值的同时要把握整体识别效果。④PECEM 从文本中抽取概念并分类，对概念在文本或句子中所处的地位没有加以判断，无法对文本语义进行解析。PMCEM 系统的建立是本体自动构建研究的基础工作。在从中文文本中抽取出通用本体概念后，需要进一步解析文本语义，构建概念之间关系，并最终实现在本体层次上描述文本语义，实现基于本体的各种应用。因此今后研究的重点将是 PMCEM 概念抽取准确率的提高、指代概念的同指处理、概念关系自动构建、本体在构建学术资源网络中的应用、基于本体的智能信息检索等方面。□

参考文献

[1] Berners-Lee T, Fischetti M, Dertouzos T M. Weaving the Web: the original design and ultimate destiny of the World Wide Web by its inventor [M]. Harper, San Francisco: [s. n], 1999

[2] 董慧，余传明，杨宁，等. 基于本体的数字图书馆检索模型研究（Ⅲ）——历史领域资源本体构建 [J]. 情报学报，2006 (5)

[3] 唐爱民，真溱，樊静. 基于叙词表的领域本体构建研究 [J]. 现代图书情报技术，2005 (4)：1-5

[4] 张继东，余以胜. 利用叙词表构建本体的方法研究 [J]. 图书情报知识，2006 (4)：82-85

[5] Guarino N, Semantic matching: formal ontological distinctions for information organization, extraction, and integration [M]. [S. l.]: Springer Verlag, 1997

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

[7] 姜守旭，王晓龙，付国宏. 一种启发式的汉语词性标注算法 [J]. 计算机工程与设计，2000 (5)：61-64

[8] Chen H H, Ding Y W, Tsa S C, et al. Description of the NITU system used for MET2 [C] //Proc. of 7th Message Understanding Conference, 1998

[9] Black W J, Rinaldi F, Mowatt D. Facile: description of the NE system used for MUC-7 [C] //Proc. of 7th Message Understanding Conference, 1998

[10] Fukumoto J, Shimohata M, Masui F, et al. Electric industry: description of the Oki system as used for MET-2 [C] // Proc. of 7th Message Understanding Conference, 1998

[11] Bender O, Franz J O, Ney H. Maximum entropy models for named entity recognition [C] //Proceedings of the Conference on Computational Natural Language Learning. Edmonton, Canada, 2003: 148-151

[12] Isozaki H. Japanese named entity recognition based on a simple rule generator and decision tree learning [J]. IPSJ Journal, 2002, 43 (5): 1481-1491

[13] 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

[14] 王睿，张洁，张由仪，等. 基于混合模型的中文命名实体抽取系统 [J]. 清华大学学报：自然科学版，2005 (1)：1908-1914

[15] 庄明，老松杨，吴玲达. 一种统计和词性相结合的命名实体发现方法 [J]. 计算机应用，2004 (1)：22-24

[16] Wu Y, Zhao J, Xu B. Chinese named entity recognition combining a statistical model with human knowledge [C] //Proceedings

（下转第 291 页）

页链接都会通过主题蒸馏器的蒸馏处理以提取其核心链接源，随着系统中核心链接源被主题蒸馏器逐一予以识别，核心链接源和其相关链接的引用率和点击率会逐渐上升，而这些排序上升后的文档或网页链接将会显示在查询结果的显著位置。

2）信息内容过滤阶段。信息过滤阶段是指从大量的信息流中寻找满足特定用户需求的文档或信息过程。内容过滤器通过对文档或网页的分类决定该文档的可用性。文档或网页中的有代表性的主题特征被抽取出来，并输入到内容过滤器中[10]。系统通过用户的查询历史记录识别用户需求和偏好，并提取用户偏好的关键特征值，然后内容过滤器通过将系统中被标识的文档或网页的关键属性与用户偏好关键特征值的匹配，用来过滤与用户偏好不相关的信息内容。在信息查询中可以通过分析用户提交检索策略，建立用户的兴趣模块，对系统检索出的结果流进行过滤，提交用户最感兴趣的文本[11]。同时，系统对新增加的数字资源也可以利用过滤系统，主动推送符合用户要求的资源，然后判断信息流中的文本是否符合用户的需求，并将符合用户需求的文本提交给用户。

3）过滤结果保存和交互阶段。经过内容过滤器的过滤将过滤信息保存在个性化信息资源库中，这样当用户提交其查询请求时，系统就可以将与其需求最相关的查询结果推送给用户。系统的交互不仅包含用户与系统的交互，还包含分布式系统间的交互。整个查询系统是大的分布式计算机系统，系统要检索与各个节点相关的信息。系统中每个节点通过和另一个节点的交流实现分布查询任务、交流信息或完成查询任务。每个节点只搜索和存储用户感兴趣内容的网络模型和系统交互模式的交互内容，从而实现真正的分布式搜索和查询。为了避免用户查询结果中无关信息查询结果的呈现，系统不仅会根据用户输入的查询关键词予以判断，还会根据用户知识背景、偏好和用户对查询结果的满意度对查询结果做出相应的优化。

4 结束语

基于信息过滤的信息查询优化通过将信息内容和用户偏好或兴趣的匹配向用户提供与其需求最相关的信息和剔除无关信息，从而减轻了信息过载给用户造成的信息认识负担和大量无关信息的获取，有效地提高了用户获取相关信息的效率。□

参考文献

[1] 胡昌平，黄晓梅，贾君枝. 信息服务管理［M］. 北京：科学出版社，2003：273-274

[2] 安立华. 异构分布数据源中基于本体的个性化查询方法研究［D］. 大连：大连海事大学，2005

[3] Van Meteren R, van Maarten S. Using content-based filtering for recommendation1 ［EB/OL］. http://www.ics.forth.gr/~potamias/mlnia/paper_6.pdf

[4] Luo Si, Rong Jin. Flexible mixture model for collaborative filtering［C］// Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington: DC, 2003

[5] Luo Si, Rong Jin. Unified filtering by combining collaborative filtering and content-based filtering via mixture model and exponential model［C］// CIKM 04, Washington: DC, 2004

[6] Goetzinger L, Park J, Lee Y J, et al. Value-driven consumer e-health information search behavior［J］. International Journal of Pharmaceutical and HealthcareMarketing, 2007, 1 (2): 128-134

[7] 黄晓斌，邱明辉. 网络信息过滤系统研究［J］. 情报学报，2004，23 (3)：326-327

[8] Ferman A M, Errico J H, Van Beek P, et al. Content-based filtering and personalization using structured metadata［C］// Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries, 2002

[9] Zhuang Ziming, Cucerzan S. Re-ranking search results using query logs［C］. ACM CIKM'06, USA. Arlington: [s.n.], 2006

[10] Fu Xianghua, Feng Boqin. Towards an effective personalized information filter for P2P based focused Web crawling［J］. Journal of Computer Science, 2006, 2 (1): 97-103

[11] 张帆，杨炳儒. 基于文本过滤的数字图书馆个性化服务技术［J］. 计算机工程与应用，2006，42 (31)：206-208

作者简介：李枫林，男，1962年生，副教授。
贺娜，女，1984年生，硕士生。

收稿日期：2007-10-15

（上接第297页）

of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition, Sapporo, Japan, 2003: 65-72

[17] 俞鸿魁，张华平，刘群，等. 基于层叠隐马尔可夫模型的中文命名实体识别［J］. 通讯学报，2006 (2)：87-94

[18] http://ccl.pku.edu.cn/

作者简介：王昊，男，1981年生，博士生。
苏新宁，男，教授，博士生导师。

收稿日期：2007-09-27