DOI:10.11992/tis.201507038
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.tp.20151111.1633.004.html

从用户需求语句建立问题可拓模型的研究

王定桥[1],李卫华[1],杨春燕[2]
(1.广东工业大学 计算机学院,广东 广州 510006; 2.广东工业大学 可拓学与创新方法研究所,广东 广州 510006)

摘 要:准确地建立待解决问题的可拓模型是可拓策略生成的关键步骤。目前的可拓策略生成系统在建立可拓模型时因自然语言理解的困难,未能充分理解用户需求,所以较难自动建立问题的可拓模型。提出了解析用户自然语言需求语句、并自动建立可拓模型的方法。该方法的核心包括 4 步:1)对用户需求语句进行组块分析得到短语序列;2)对短语序列进行分类;3)使用匹配规则抽取分类后的短语,得到便于计算机处理的需求信息;4)结合数据库技术进行可拓模型的建立。以租房问题为案例,实现了该方法。实验结果表明,该方法能较好地理解用户需求信息并成功建立租房问题可拓模型。

关键词:可拓学;可拓模型;可拓策略生成;信息抽取;分类
中图分类号: TP391 **文献标志码**:A **文章编号**:1673-4785(2015)06-0865-07

中文引用格式:王定桥,李卫华,杨春燕. 从用户需求语句建立问题可拓模型的研究[J]. 智能系统学报, 2015, 10(6): 865-871.
英文引用格式:WANG Dingqiao, LI Weihua, YANG Chunyan. Research on building an extension model from user requirements [J]. CAAI Transactions on Intelligent Systems, 2015, 10(6): 865-871.

Research on building an extension model from user requirements

WANG Dingqiao[1], LI Weihua[1], YANG Chunyan[2]
(1.School of Computer, Guangdong University of Technology, Guangzhou 510006, China; 2. Research Institute of Extenics and Innovation Methods, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Building an effective extension model to solve a problem is a key step in generating an extension strategy. Due to the complexity of natural language processing, the current extension strategy generation system is insufficiently clear with respect to user requirements, so it is hard to automatically build an extension model. In this paper, we propose a method for parsing the user requirement sentence in order to then automatically build the extension model. This method contains four core steps. First, chunk parsing is performed on the sentence containing the user requirements to obtain the phrase sequence. Secondly, the phrase sequence is classified with a classifier. Thirdly, based on the matching rule, information is extracted from the classified phrase to obtain the information required for computer processing. Next, database technology is used to build the extension model. Using a tenement building as an example, we implemented and tested our proposed method. Based on our experimental results, we proved that the proposed method is effective for understanding user requirements in order to build an extension model.

Keywords: extenics; extension model; extension strategy generation; information extraction; classification

矛盾问题是指在现有条件下无法实现人们要达到的目标的问题。矛盾问题智能化处理的研究对现代科学的发展具有重要意义[1]。可拓学研究的矛盾问题主要分为不相容问题和对立问题,本文主要讨论不相容问题。

收稿日期:2015-07-23. 网络出版日期:2015-11-11.
基金项目:国家自然科学基金资助项目 (61273306).
通信作者:王定桥. E-mail: wangdingqiao2012@qq.com.

解决不相容问题,一般包括 6 个步骤[1],其中第 1 个步骤就是建立问题的可拓模型。因此,要借助计算机智能化地处理不相容问题,首要的任务是准确地建立问题的可拓模型。

目前,建立可拓模型主要通过 2 种方式:1)在人充分理解问题的基础上,利用形式化符号手工建立。这种方式主要由少数专家和研究人员使用,对可拓学专业知识要求较高,不适合广大用户;2)通过可拓策略生成系统的界面输入问题相关的参数,来辅助系统建模。例如早期研究的自助游可拓策略生成系统[2]、租房可拓策略生成系统[3]、求职问题可拓策略生成系统[4]等都是采用这种方式。但使用这种方式时存在 2 个问题:1)当参数过多时,输入界面通用设计变得困难;2)如果输入文字稍长,系统难以快速理解用户问题,建模效率低。

1 关键技术及解决思路

1.1 问题可拓模型建立所涉及到的技术

建立不相容问题的可拓模型,实际上是一个收集与问题 P 相关的信息,然后界定问题的目标 G 和条件 L,形成可拓模型 $P = GL$ 的过程。其中主要涉及到以下技术:

1)信息抽取技术

信息抽取技术是指从一段文本中抽取指定的事件、事实等信息,形成结构化的数据并存入一个数据库,供用户查询和使用的过程[5]。从用户需求语句,抽取属性及量值,实际上就是一个信息抽取的过程。

2)领域本体

领域本体是用于描述特定领域知识的一种专门本体。它给出了领域实体概念、领域属性概念、领域属性值及相互关系,以及该领域所具有的特性和规律的一种形式化描述[6]。实际上在可拓策略生成系统整个过程中,都需要借助领域本体知识。在建立模型时领域本体能够为抽取属性的种类、量值范围、量值单位提供一致的指导。

3)数据库技术

可拓策略生成系统需要借助数据库技术,存储基础数据、知识库、规则库等内容。在建立模型时用户提供的需求语句可能只提供了目标或条件之中的一个,或者提供了不完整的目标和条件,这些情况下需要利用数据库中数据对可拓模型进行补充和完善。

1.2 用户需求语句信息抽取的主要内容

当前信息抽取还只是面向特定领域开展,能够真正实现大规模应用的信息抽取系统仍然未出现[7]。知网的中文信息语义处理技术[8]有一定的参考价值,但仍然不能直接用于建立可拓模型。在实际应用中,用户表达的语句通常会出现不完全合乎语法、信息省略、包含错别字、简写、歧义等情况,为信息抽取增加了难度。因此,结合实际问题需要,本文将从用户需求语句主要抽取的信息分为 4 类,如下:

1)可量化的量值

这类信息是指,用户表达的明确的属性和量值。例如:

例 1 一个人想在沙坪坝租房,只租 1 个月,有空调、卫生间,房租大概 350 元。

这个语句中用户给出的区域、租金、租期和配套设施都属于可量化的量值。

2)抽象的量值

自然语言表达中通常会不自觉地出现一些抽象描述,当这些描述与可拓策略生成系统期望的量值类型不一致时,仍然需要抽取,以便做出更合理的决策。例如:

例 2 我要在大连市内找工作,想租个房子,月租便宜点、交通方便点的。

这里用户提供的租金描述为便宜的、交通状况为方便的,都属于抽象量值,而可拓策略生成系统实际需要的为数量值。

3)优先级信息

用户语句中很可能通过“必须”、“一定要”、“最好”等关键字,来表达他的特殊需求,例如:

例 3 想在滨州市新北中附近租房。便宜点的,合租也可以。一定要有暖气。

用户表达的需求“一定要有暖气”可作为可拓策略生成的一个筛选条件。

4)逻辑关系信息

逻辑关系,主要包括用户表达的并列、或者、否定、反义等逻辑关系。例如:

例 4 我要在南宁市内租房,一室或者二室都可以,500 元以内,不要中介的,安全的。

第 1 类信息的抽取,是一个命名实体识别的过程。命名实体识别(named entity recognition, NER)的主要任务是识别出文本中的人名、地名等专有名称和有意义的时间、日期等数量短语并加以归类[9]。实际研究中,命名实体识别的对象根据不同应用而有所改变,例如在医学文本中识别生物命名实体[10]、中文旅游景点的识别[11]等。目前命名实体识别主要的方法包括:基于规则和词典的方法、基于统计的方法、二者混合的方法。文献[12]对比并指出了各个方法的优点和局限。

上述第 2 类信息的抽取是一个分类的过程。对于用户提供的不够具体的量值,首先确定其描述的内容属于什么属性,然后可以按 2 种方式处理。一种是为抽象描述提供预设值,例如为租金构造离散函数,根据值域分为便宜、一般、高价 3 个等级,这样用户提供的抽象值也可以量化。另一种是利用抽取的抽象值,指导后续的人机交互过程。

上述第 3 类和第 4 类信息,主要是在确定了属性和量值后,在这个量值所在的上下文环境中,通过有限状态机实现。构造一个包含表达优先级、反义这类信息的关键词的词典,通过有限状态机中状态之间转移来实现。例如量词短语“1 000 元”所在上下文为“租金超过 1 000 元的就不要了”,首先获取的量值 1 000 元,通过输入单词“超过”和“不要”,量值转换为最终的区间值[0,1 000]。

1.3　问题解决思路

在处理具体问题的用户需求语句时,时间、货币、日期等实体占据很大比例,其识别比较简单,可以在分类后采用模式匹配方式实现;而其他实体类数量比较少,识别比较困难。针对这一情况,本文决定采用混合的方法,即分类和规则匹配结合的方法来完成属性和量值的抽取。文献[13]中采用混合的方法提高了命名实体识别的准确率和召回率。受到此方法的启示,本文从用户需求语句中提取信息时,先对用户语句进行组块分析获取短语序列;然后对短语序列进行分类,通过对分类后的短语使用规则匹配获取属性和量值;最后,使用这些属性和量值并结合数据库技术建立问题的可拓模型。

2　建立可拓模型的步骤

建立可拓模型的处理流程如图 1 所示。

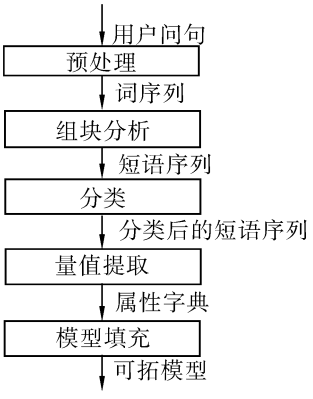


图 1　建立可拓模型的流程图

Fig.1　Steps to build extension model

2.1　预处理

预处理的主要目的是为了简化后续处理。这一阶段完成工作包括:过滤、替换、数据格式调整、分词。过滤主要是过滤客气词(例如“请问”)、语气词(例如“急求”)、询问相关词(例如“有没有”)。替换包括错别字替换(例如“500 一下”替换为“500 以下”)和同义词替换(例如“旁边”、“周围”等替换为“附近”)。数据格式调整,包括数值都使用数字表示,数值范围调整为统一格式。分词时保留原句中的逗号等分隔符,将长语句分割为短语句,得到多个短语句的分词序列。

2.2　组块分析

组块是一种语法结构,是符合一定语法功能的非递归短语[14]。组块分析包括组块的划分和识别,也就是识别出语句中像动词短语、形容词短语这类短语的过程。本文借助 Stanford Parser 来完成组块分析。Stanford Parser 中文解析器是基于 Chinese Treebank 的,具体的组块标记可参考文献[15]。

在实验的过程中,发现组块切分的粒度,对于抽取的信息数量有较大影响,尤其是当用户语句中量值信息密度较大时。

例 5　2 个 800 块以内的单间。

预处理后形成的语义树,如图 2 所示。在此片段中,需要抽取包括房间数量(两间),租金(800 块以内)以及房子样式(单间)在内的 3 个属性和量值。如果仅切分为一个 NP 短语,那么后续阶段处理时可能漏掉属性;而切分为 QP、DNP 和 NP,借助上下文信息,则能很好地捕获 3 个属性信息。

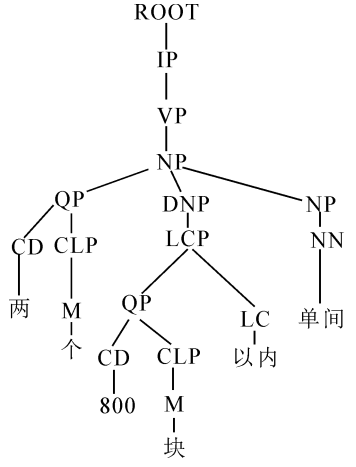


图 2　例 5 对应的语法树

Fig.2　The parse tree of the fifth example

Chinese Treebank 提供了 17 个短语标记,其中 CP、IP 和 UCP 粒度过大,需要处理其内部节点;PRN、LST 和 DP 一般不出现在用户需求语句中,不予处理;CLP 类型需要处理其上级 QP 短语,VP、

DNP、DVP 需要处理其内部节点;FRAG 是不能构建完整结构的片段元素,也需要处理其内部节点;主要处理的类型包括 PP、QP、NP、LCP、ADJP、ADVP 6 种短语。

6 种主要短语中,最复杂的是 NP。NP 分为简单名词短语和复合名词短语。简单名词短语由单个普通名词 NN、专有 NR、时间名词 NT 构成;复和名词短语的情况主要包括 5 种情况,QP-NN 复合(例如“一个月”)、NN-NN 复合(例如“个人房源”)、NN-CC-NN 复合(例如“空调和洗衣机”)、多个时间名词复合(例如“3 月 29 日”),以及 NR 与若干个 NN 复合(例如“北京海淀区附近”)。

根据上述分析,采用自底向上的搜索方法来获取短语序列,实现伪代码如下所示:

```
getPhraseList(Tree root,List<String> phList) {
root = pruneTree(root);leaves = root.leaves;
while(! leaves.isEmpty()) {
  curLeave, tNode = leaves[0], null
  p2 = curLeave.ancestor(2, root);
  switch( p2 .label) {
      case "QP":
tNode=handleQP(root, p2 ,phList);break;
 case "NP":
tNode=handleNP(root, p2 , phList);break;
 case "LCP":
tNode=handleLCP(root, p2 , phList);break;
 case "ADJP","PP","ADVP":
            phList.add(chToStr( p2.label, p2);
        tNode= p2 ;break;
 default:
 handleDefault();break;
 }
 if(tNode! =null)
      leaves.remove(tNode.getLeaves());
    else leaves.remove(curLeave);
```

其中 pruneTree 完成语法树的剪枝工作,移除 SP、PN、PU 等标记的节点,移除一些常见动词(例如“想”),副词 AD 和形容词 JJ 仅保留词典中存在的词;ancestor 为从当前节点向上获取父节点,参数为向上查找层数。handelDefault 处理的是默认情况,默认情况下仅处理包括动词 VV,形容词 VA,名词 NN 这些单词。对于这类词,不使用包含它们的父节点类型标记它们,而是直接使用它的词性作为标记,将他们作为其他短语的上下文环境保留起来,以便于后续的分类工作。handleQP、handleNP、handleLCP3 个函数分别处理 QP、NP、LCP 短语。给定例句:

例 6 一个人想在郑州中央商务区附近租个 350 块左右单间。

得到短语序列:[QP:一个/CD, NN:人/NN, PP:在/P 郑州/NR 中央/NN 商务区/NN, VV:租/VV, QP:个/M, LCP:350/CD 块/M 左右/LC, NN:单间/NN]。

2.3 分类

使用分类算法的关键是找到有效的特征向量。本文选取的特征包括:短语类型,包含测试特征,以及词或者词性特征。包含测试特征是对短语是否包含某类词,进行测试而得到的整型值。不同短语测试后的特征个数也不统一,因此把包含测试特征附加到短语类型上,作为一个特征。共选取了 6 个特征用于分类,如表 1 所示。

表 1 用于分类的特征向量

Table 1 Features used in classification

短语或词	包含测试特征	词或词性特征
QP	连词,序数词	量词,左边名词,左边动词,右边名词,右边动词
PP	地址,时间,数词,连词	量词,首词,末尾词或其词性,左边动词或名词,右边动词或形容词
LCP	地址,时间,数词,连词	量词,末尾词,最后一个名词,左边动词或名词,右边动词
NP	地址,时间,连词	名词 1,名词 2,左边动词,右边名词,右边动词
ADJP	无	形容词 JJ,左边动词,右边名词,其余置为空
ADVP	无	副词 AD,左边名词,右边动词或形容词,其余置为空
VA	无	形容词 VA,左边名词,右边名词,其余置为空
NN	无	名词 NN,左边动词,右边动词或形容词,其余置为空
VV	无	动词 VV,左边动词,右边动词,其余置为空

包含测试特征中,连词是指标记为 CC 的单词,序数词是标记为 OD 的单词,时间是指标记为 NT 的单词,数词是指 CD 或者 OD 的单词。包含地址测试需要借助分词系统完成,使用单词的词性测试其是否属于地址类词性。

需要注意,某些单个 NN(例如“单间”)、VA(例如“便宜”)、VV(例如“合租”)本身就能表达一个量值,用户很可能单独使用它们来表达需求,因此,需要将这类词记录在词典中。在遇到这类词时,将其添加到分类任务中,这类单词的特征列在表 1 的末尾 3 行。

PP 短语中,如果末尾词是普通名词则使用单词本身,否则使用其词性。包含单个 NN 的 NP,将以 NN 标记独立处理。对于其他 NP,如果包含地址或日期,名词 1 和名词 2 置为空。对于不包含地址或日期的复合名词短语,需要特别处理。2.2 节中提到的 NN-NN 和 NN-CC-NN 类短语,将其 2 个 NN 作为名词 1 和名词 2 填充;NR 与若干 NN 复合的情形,将 NR 与 NN 连成一个词,作为名词 1 填充,名词 2 置为空。

在有监督的分类器训练的过程中,根据问题和关注的属性,使用不同的标签。与问题无关的短语或词,统一标记为无关类,在后期过滤掉这些内容。使用训练后得到的分类器,对短语序列分类,并合并相邻的同类标签,得到最终分类后的短语序列。

2.4 量值提取

对分类后的短语,针对每一类别,建立一系列匹配规则来抽取量值。匹配时间和数字类表达式的规则比较通用;对于名词、动词、形容词等可以根据分类结果,借助词典来更准确地确定边界。

例如租房问题中,匹配区域的规则,用正则表达式书写并按照优先级列出如下:

rule 1:　(在?)(.*)(附近)

rule 2:　(在|靠近)? (.*)(租)

rule 3:　(离|靠|距)(.*)(近)

rule 4:　(在?)(.*)(环)

rule 5:　(在?)(地铁|公交)(.*)(线|路)

rule 6:　拼接词性表示地点的单词

除了匹配外,还需进行 3 项工作:

1)理解优先级、逻辑关系

在短语对应的原文中获取表达这类信息的关键词,通过有限状态机,即可获取用户真正要表达的量值。这种方法仅在用户将关键词混在多个量值之间,并且不加任何分隔符的情形下失效。在实际应用中这种情形出现的概率很小。

2)同类合并和歧义消解

对于集合类型的量值,需要对量值进行归并;对于单一类型的量值,需要根据量值特点,进行歧义消解。例如用户首先提供了一个范围比较大的地址,接着又补充了一个小范围地址,可以使用大地址后加上小地址的方式,准确定位地址。

3)量值标准化

同一属性的不同量值需要转换为单位统一的量值,以便于处理。例如租房问题中用户提供租期属性的量值,可能是“半个月”,“半年”,“一个星期”等可以统一调整到以月为单位的数量值。

经过这一阶段的处理,得到了最终的属性字典。例 6 最终得到属性字典如下:

{区域:郑州中央商务区, 租金:[0,350], 样式:单间, 住户人数:1, 租房数量:1}

2.5 模型填充

这一阶段,使用上一阶段获取的属性字典,并结合数据库技术,建立可拓模型。首先将属性字典中各个属性和量值填充到目标或者条件基元中去。对于目标或者条件基元中缺少的部分,则需要根据领域本体,借助数据库或者人机交互来补充。

经过上述流程的 5 个阶段,最终从用户语句建立了可拓模型。

3 实现案例

3.1 案例介绍

文献[3]给出了一个租房问题,下面以此问题为背景来展开实验。实际语料中用户表达的属性通常都有多个,本文一共关注了 16 个属性,表 2 给出了部分属性的示例。

表 2　租房问题中用户表达的属性示例

Table 2　User expressed attributes in tenement question

属性	量值类型	量值单位	量值示例
区域	字符串	无	番禺大学城
租金	整数	元	800 块
面积	整数	平方米	80 平米
样式	字符串	厅,室	两室一厅
楼层	整数	楼,层	10 楼
房源	字符串	无	个人

一般地,上述多个属性,可以根据实际应用情况,为每个属性分配不同的权重用于指导可拓策略的生成和评价过程。

在实验过程中使用的资源包括:

1)语料资源,在百度和好搜两大网络平台,使用爬虫程序抓取到与租房问题相关的语句;

2)分词系统,使用哈工大讯飞语言云服务;

3)组块分析,使用斯坦福中文解析器;

4)分类器,使用张乐博士 maxent 工具箱;

5)词表,手工编制了 2 个词表,预处理词表大小为 600,匹配使用的词表大小为 140;

6)数据库,修改了文献[3]中爬虫程序,获取了租房信息的数据并存贮在数据库中;

7)条件随机场,使用 CRF++工具箱。

3.2 实验结果及分析

按照惯例,使用信息抽取任务中的准确率 P、召回率 R 以及 F 值来评测系统性能。作为对比试验,选取文献[16]中用于识别微博命名实体的条件随机场方法,并使用了文中的特征模板。采用 4-tag(B、M、E、S)对每个属性进行标注,利用 CRF++工具进行了实际抽取工作。在处理的语句中,采用 10-cross validation 验证方法,得到的平均正确率 P、召回率 R、F 值,如下表 3 所示。

表 3 实验结果

Table 3 Experimental result

方法	样本容量	准确率 P/%	召回率 R/%	F 值/%
本文	550	96.00	91.99	93.95
CRF	550	94.58	87.22	90.75
本文	1 100	96.60	93.06	94.79
CRF	1 100	95.52	90.70	93.05

上述结果表明,本文方法同 CRF 方法相比,性能有所提高。其中,准确率的提高在于使用了匹配规则抽取分类后的短语;召回率的提高在于使用组块分析后,对短语进行分类。CRF 对样本依赖比较大,当样本容量较小时,本文方法更具优势。

文献[3]中策略系统只考虑了区域、租金、交通状况和面积 4 个属性。利用本文方法,不仅能获取更多的属性,还能理解抽象量值、优先级关系和逻辑关系,从而能更容易地为用户生成理想的策略。

对用户语句进行信息提取后,结合数据库检索技术,就能建立最终的可拓模型。例如对于语句:

例 7 广州大学旁求租房!不想通过中介,3 月 29 号左右可以入住,拟租时间 3 个月以上,希望有一室一卫的公寓,能连接宽带,月租不超过 600 都可以。通过上述方法,从用户需求语句,获取了目标物元 M;并从数据库中查找到一条最接近用户目标的房子,确定为条件物元 L。则最终确定了问题的可拓模型表示为

$$M=\begin{bmatrix} 目标房子\ w & 区域 & 广州大学 \\ & 租金 & [0,600] \\ & 样式 & 一室一卫 \\ & 类型 & 公寓 \\ & 配套设施 & 宽带 \\ & 租期 & 3\ 月 \\ & 入住时间 & 3\ 月\ 29\ 日 \\ & 房源 & 个人 \end{bmatrix}$$

$$P=GL=\begin{bmatrix} 租用 & 支配对象 & M \\ & 施动对象 & 租房者 \end{bmatrix}\cdot$$

$$\begin{bmatrix} 出租房\ A & 区域 & 北亭村云程大街\ 8\ 号 \\ & 租金 & 750 \\ & 样式 & 一室一卫 \\ & 类型 & 公寓 \\ & 配套设施 & 宽带\ \wedge\ 空调\ \wedge\ 阳台 \\ & 房源 & 个人 \end{bmatrix}$$

在完整的可拓策略生成系统中,下一步工作就是由可拓模型,求出核问题模型。当核问题模型中的条件满足目标的要求时,就不是不相容问题,不需要解决,说明系统帮用户找到了所需要的房子。当核问题模型中的条件不满足目标的要求时,就是不相容问题,需要利用可拓策略生成系统,首先判断问题不相容的程度,然后通过拓展、变换和评价,生成解决不相容问题的策略。策略生成的详细步骤参见文献[3]。

4 结束语

本文通过对用户需求语句进行组块分析后得到的短语序列进行分类,并结合匹配规则进行信息抽取,得到了计算机较容易识别的需求信息。这种方法有效实现了从用户需求语句到可拓模型的转换,减轻了人的劳动,提高了可拓模型建立的效率和质量,为可拓模型的建立提供了新的方法。

试验表明本文的方法已经得到比较满意的结果。今后还可以通过 2 种方式进一步完善:1)针对特定问题,在分词时使用用户字典,提高分词的准确率;2)使用实际语料训练 Stanford Parser,提高它词性标注和句法分析的准确率。

另外,限于目前本项目还没有建立通用的问题语料库,本文仅实现了租房问题案例。下一步工作是建立其他问题的语料库并进行相关测试,以利于开发较为通用的可拓策略生成系统。

参考文献:

[1]杨春燕, 蔡文. 可拓学[M]. 北京: 科学出版社, 2014: 1-250.

YANG Chunyan, CAI Wen. Extenics[M]. Beijing: Science

Press, 2014: 1-250.

[2]方卓君, 李卫华, 李承晓. 自助游可拓策略生成系统的研究与实现[J]. 广东工业大学学报, 2009, 26(2): 83-89.
FANG Zhuojun, LI Weihua, LI Chengxiao. Research and realization of extension strategy generating system for independent travel[J]. Journal of Guangdong University of Technology, 2009, 26(2): 83-89.

[3]李承晓, 李卫华. 租房可拓策略生成系统[J]. 智能系统学报, 2011, 6(3): 272-278.
LI Chengxiao, LI Weihua. Research on a tenement extension strategy generation system[J]. CAAI Transactions on Intelligent Systems, 2011, 6(3): 272-278.

[4]陈亚男, 李卫华. 求职问题可拓策略生成系统的研究与实现[J]. 广东工业大学学报, 2012, 29(1): 88-93.
CHEN Yanan, LI Weihua. Research on the extension strategy generating system for job-seeking problems[J]. Journal of Guangdong University of Technology, 2012, 29(1): 88-93.

[5]刘迁, 焦慧, 贾惠波. 信息抽取技术的发展现状及构建方法的研究[J]. 计算机应用研究, 2007, 24(7): 6-9.
LIU Qian, JIAO Hui, JIA Huibo. Research on approaches of information extraction system[J]. Application Research of Computers, 2007, 24(7): 6-9.

[6]于江德, 李学钰, 樊孝忠. 信息抽取中领域本体的设计和实现[J]. 电子科技大学学报, 2008, 37(5): 746-749.
YU Jiangde, LI Xueyu, FAN Xiaozhong. Design and implementation of domain ontology for information extraction[J]. Journal of University of Electronic Science and Technology of China, 2008, 37(5): 746-749.

[7]郭喜跃, 何婷婷. 信息抽取研究综述[J]. 计算机科学, 2015, 42(2): 14-17, 38.
GUO Xiyue, HE Tingting. Survey about research on information extraction[J]. Computer Science, 2015, 42(2): 14-17, 38.

[8]董振东, 董强, 郝长伶. 知网的理论发现[J]. 中文信息学报, 2007, 21(4): 3-9.
DONG Zhendong, DONG Qiang, HAO Changling. Theoretical findings of HowNet[J]. Journal of Chinese Information Processing, 2007, 21(4): 3-9.

[9]CHINCHOR N. MUC-7 Named entity task definition[C]// Proceedings of 7th Message Understanding Conference. Virginia, USA, 1998.

[10]张向喆, 王明辉, 赵洪波, 等. 生物医学文本中命名实体识别研究[J]. 上海交通大学学报:农业科学版, 2010, 28(2): 132-137.
ZHANG Xiangzhe, WANG Minghui, ZHAO Hongbo, et al. Research on named entity recognition from biomedical literature[J]. Journal of Shanghai Jiao Tong University: Agricultural Science, 2010, 28(2): 132-137.

[11]薛征山, 郭剑毅, 余正涛, 等. 基于 HMM 中文旅游景点的识别[J]. 昆明理工大学学报: 理工版, 2009, 34(6): 44-48.
XUE Zhengshan, GUO Jianyi, YU Zhenftao, et al. Recognition of HMM-based Chinese tourist attractions[J]. Journal of Kunming University of Science and Technology: Science and Technology, 2009, 34(6): 44-48.

[12]孙镇, 王惠临. 命名实体识别研究进展综述[J]. 现代图书情报技术, 2010, 26(6): 42-47.
SUN Zhen, WANG Huilin. Overview on the advance of the research on named entity recognition[J]. New Technology of Library and Information Service, 2010, 26 (6): 42-47.

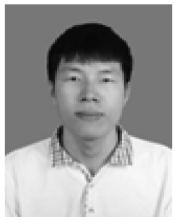
[13]LIN Yifeng, TSAI T H, CHOU Wenchi, et al. A maximum entropy approach to biomedical named entity recognition [C]//Proceedings of the 4th ACM SIGKDD Workshop on Data Mining in Bioinformatics. Seattle, Washington, USA, 2004.

[14]李素建, 刘群, 杨志峰. 基于最大熵模型的组块分析[J]. 计算机学报, 2003, 26(12): 1722-1727.
LI Sujian, LIU Qun, YANG Zhifeng. Chunk parsing with maximum entropy principle[J]. Chinese Journal of Computers, 2003, 26(12): 1722-1727.

[15]XUE Naiwen, XIA Fei, CHIOU Fudong, et al. The Penn Chinese Treebank: phrase structure annotation of a large corpus[J]. Natural Language Engineering, 11(2): 207-238.

[16]邱泉清, 苗夺谦, 张志飞. 中文微博命名实体识别[J]. 计算机科学, 2013, 40(6): 196-198.
QIU Quanqing, MIAO Duoqian, ZHANG Zhifei. Named entity recognition on Chinese microblog[J]. Computer Science, 2013, 40(6): 196-198.

作者简介:



王定桥,男,1988 年生,硕士研究生,主要研究方向为智能软件。

李卫华,女,1957 年生,教授,硕士生导师,主要研究方向为面向 Agent 计算、网络信息系统、智能软件,发表学术论文 40 余篇。