文章编号：1003-0077(2009)02-0034-07

基于汉语框架网的旅游信息问答系统设计

李 茹[1,2]，王文晶[1]，梁吉业[1,2]，宋小香[1]，刘海静[1]，由丽萍[2,3]

(1. 山西大学 计算机与信息技术学院 山西 太原 030006；
2. 山西大学 计算智能与中文信息处理教育部重点实验室 山西 太原 030006；
3. 山西大学 管理学院 山西 太原 030006)

摘 要：该文借助汉语框架网(Chinese FrameNet，简称 CFN)在语义表达方面的独特优势，探讨用本体描述语言建立面向特定领域的汉语框架语义知识库，并且以旅游交通领域中问答系统设计为例分析方法的有效性。方法中首先利用 TREC 分类与本体分类相结合的方式为查询问句分类，然后提出基于 CFN 的问句分析策略，通过 CFN 语义分析得到问句中三元组：语义谓词、语义主体和语义客体，在问句分析的基础上从旅游本体知识库中对答案进行抽取并对答案处理，同时用本体编辑工具 Protégé 编码，实验证实方法是有效的。

关键词：计算机应用；中文信息处理；汉语框架网；本体；问答系统

中图分类号：TP391　　**文献标识码**：A

Design of Tourism Question Answering System Based on the Chinese FrameNet

LI Ru[1,2], WANG Wen-jing[1], LIANG Ji-ye[1,2], SONG Xiao-xiang[1], LIU Hai-jing[1], YOU Li-ping[2,3]

(1. School of Computer& Information Technology, Shanxi University, Taiyuan Shanxi 030006, China;
2. Computer Intelligent and Chinese Information Processing of The Ministry Education Key Laboratory Built Together by Province and Department, Shanxi University, Taiyuan Shanxi 030006, China;
3. School of Management , Shanxi University, Taiyuan Shanxi 030006, China)

Abstract: Taking advantage of the semantic expression in Chinese FrameNet (CFN), this paper discusses the construction of the domain specific Chinese FrameNet semantic base using owl, and validate and analyze its effectiveness by the design of Question Answering System in the transportation domain. In the proposed QA system, the query questions are first classified by a combination of the TREC categories and the ontology categories. Then we propose a question analysis strategy based on the CFN, aiming at the triple of the question: Semantic predicate, semantic subject and semantic object. On the basis of the CFN semantics analysis, the answer is extracted from the tourism ontology base. This approach is implemented by the ontology editor Protégé, and the experiment proves the validness of this method.

Key words: computer application; Chinese information processing; Chinese framenet; ontology; question answering system

1 引言

自动问答系统是自然语言处理领域中的一个热点问题，它既能够让用户用自然语言句子提问，又能够为用户返回一个简洁、准确的答案，而不是一些相关的网页。目前对旅游信息的获取是通过旅游信息网站的搜索引擎，它是基于关键字匹配或主题分类[1-2]，对用户问题的回答准确率都很低，其主要原因是因特网在信息表达和检索方面存在的缺陷，没

收稿日期：2008-08-30　**定稿日期**：2008-10-28

基金项目：国家 863 高技术研究发展计划资助项目(2006AA01Z142)；国家社会科学基金青年项目(07CYY022)；山西省高等学校拔尖人才基金项目；太原市科技局项目；山西省大学生创新项目

作者简介：李茹(1963—)，女，教授，研究方向为智能信息处理；王文晶 (1981—)，女，硕士生，研究方向为自然语言处理；梁吉业(1962—)，男，教授，研究方向为计算智能。

有提供给计算机可读的信息，限制了计算机在检索中的自动分析能力。

本体(Ontology)原先是一个哲学概念，被哲学家用来描述事物的本质。1993 年，Gruber[3] 给出定义，即“本体是概念模型的形式化规范说明”，目标是获取、描述和表示相关领域的知识。本体是解决语义层次上的万维信息共享和交换的基础。本体的描述语言 OWL(Ontology Web Language)[4] 是 W3C(The World Wide Web Consortium)力推的本体描述语言，它以描述逻辑为基础，具有良好的语义表示和逻辑推理能力。

W3C 提出了语义网，其目标是人和机器都可使用、识别、解析 Web 上的信息。山西大学汉语框架语义网课题组在刘开瑛教授的指导下，以框架语义学为基础，以真实语料为支持，以伯克利 FrameNet 提供的数据为参照，研究构建了汉语框架网(Chinese FrameNet，简称 CFN)[5]，它由框架库、句子库和词元库三部分组成，使用 XML(Extensible Markup Language，可扩展的标记语言)、RDF(Resource Description Framework，资源描述框架)、OWL(Web Ontology Language，Web 本体标记语言)对资源进行描述，以期为语义 Web 等的应用提供一部计算机可读、可理解的语义词典，为实现语义 Web 中的语义知识共享以及智能化、个性化的 Web 服务提供基础资源[6]。

本文面向山西旅游 QA 信息系统，对用户的问题，利用汉语框架网(CFN)对问题进行语义分析，形成问句向量，利用本体知识库对答案进行抽取，最后通过答案处理模块对答案进行优化。

2　汉语框架网(CFN)

汉语框架网是一个以 Fillmore 的框架语义学[7-8] 为理论基础，以加州大学伯克利的 FrameNet[9] 为参照，以汉语真实语料为依据的供计算机使用的汉语词汇语义知识库。

CFN 数据库由框架库、句子库和词元库三部分组成。框架库以框架为单位，对词语进行分类描述，明确给出框架的定义和这些词语共有的语义角色即框架元素，并描述该框架和其他框架之间的概念关系；句子库包含带有框架语义标注信息的句子，即按照框架库所提供的框架和框架元素类型，标注句子的框架语义信息和句法信息；词元库记录词元的语义搭配模式和框架元素的句法实现方式，它们是从句子库提供的标注结果中生成。

3　旅游信息本体的构建

面向山西旅游信息，选取了 10 个有特色的山西旅游景点(重点是山西五台山)。针对每个景点都建立了语料库，并构建了汉语框架知识库。在景点语料库的基础上，根据旅游六要素即游、购、娱、食、住、行，对文档进行了术语的抽取，并进行了旅游本体模型的初步构建。构建本体模型过程中，参照了《中国分类主题词表》、《旅游服务基础术语》(gb/t 16766-1997)、《旅游规划通则》(gb/t 18971-2003)、旅游业各学科在中国图书馆分类法中所属类别、《旅游业标准体系表》、《旅行社国内旅游服务质量要求》(lb/t004-1997)、《导游服务质量》(gb/15971-1995)、中国国家标准网(www.chinabg.org)、旅游规划通则(gb/t 18971-2003)、旅游资源分类调查与评价(gb/t 18972-2003)及旅游服务基础术语(gb/t 16766-1997)等有关标准。

实验系统采用 OWL Lite 进行本体模型的编码，并使用了美国斯坦福大学的本体编辑工具 Protégé[10]。本体的建立严格定义了类之间的逆关系(Inverse Of)、传递关系(Transtive Property)、函数关系(Functional Property)、对称关系(Symmetric Property)、逆函数关系(Inverse Functional Property)以及对属性的限制。通过 Protégé，把与数据库相关的概念，关系和实例用 OWL 和 RDF 表示出来，存储为 OWL 文档。RacerPro 推理机[11] 在辅助建模阶段有很大的作用，它可以用于检查一致性、推理出新的分类体系等。

4　系统构架

实验系统的主要模块包括：预处理模块，问句匹配，答案抽取及答案处理模块。系统构架如图 1 所示。

1. 提交问题：用户提出查询请求。

2. 预处理：对用户提交的问句进行预处理，即识别有用的实体的信息，如命名实体识别，以及分词和词性标注。其中均用到 CFN 以及专业领域库，专业领域库中存储旅游领域的知识条目，并以 RDF 的形式命名了一个空间，以便系统对领域专有名词切分正确。

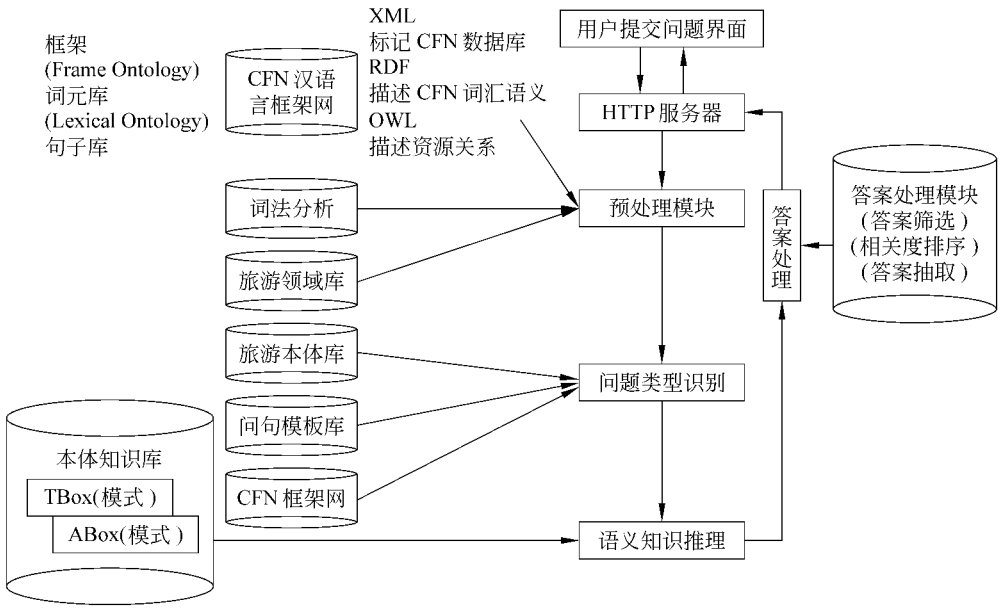

图 1 旅游信息问答系统构架

3. 问句匹配：由于问答系统中存在口语词汇较多，所以建立旅游领域中的词汇对应的口语词汇词典，以便更好地语义理解。在进行简单的语义分析之后，通过关键字的粗略提取，利用旅游扩展词库提取出用户的查询要求，同时结合 CFN 中框架的语义信息可以高效率地提取出相关信息，从而确定了检索的类型以及检索的策略。

4. 语义知识推理：入口是转化生成的 RDF 三元组问句向量，并利用本体知识库中 T—Box 和 A—Box 中进行语义知识的推理，即进行答案抽取。T—Box 中包括 ALC 概念间的蕴含和等同关系，A—Box 包括领域个体和概念以及个体对和关系间的隶属关系。

5. 答案的处理：即过滤掉与答案无关的内容，并进行相关度排序和答案的抽取。然后把查询结果递交给 HTTP 服务器，再由 HTTP 服务器把结果消息发送给用户界面。

5 问题分类

问题分类是问答系统中重要的组成部分。不同的角度可以有不同的问题分类，如形式上分疑问、设问、反问，或特指问、选择问、是非问；目的上分查找信息、验证事实、收集资料；从性质上分开放型、封闭型等等；从内容上分，可以直接利用 top-lever ontology 的概念分类体系[12]，较全面多层次进行分类。针对旅游领域，为了更好的分析和回答问句，本文采取了多角度分类形式，在 TREC[18] 分类的基础上，利用本体的思想，对问题分类。

目前对收集到关于旅游景点五台山的 1 566 条问句进行了类别统计，如表 1 所示。

表 1 问句分类统计

分类角度	类别	数量	百分比
问句形式	是非	268	17.3%
	特指	1 104	70.2%
	选择	98	6.3%
	其他	96	6.2%
TREC 分类与本体分类结合	人物	51	3.26%
	地点	90	5.75%
	时间	95	6.07%
	数字	185	11.81%
	(旅游领域)实体	410	26.18%
	描述	506	32.31%
	未知	229	14.62%

6 汉语框架语义角色与本体三元组

汉语框架语义角色的标注，实现了词一级的语义描述，直接提高了问句中提取本体三元组的效率。

6.1　汉语框架语义角色标注

CFN 标注有三层，第一层为框架元素，框架元素分为核心框架元素和非核心框架元素。核心框架元素是一个框架在概念理解上的必有成分，它们在不同的框架中类型和数量不同，显示出框架的个性。非核心框架元素并不显示框架的个性，而是表达时间、空间、环境条件、原因、目的等外围语义成分。第二层为短语类型标注，第三层为句法功能标注。

6.2　基于 CFN 的问句分析

在旅游交通领域中，目前针对交通工具这一类问句进行了研究与实验。由于问句识别主要是依据句子中的疑问词以及疑问意向词来识别，同时考虑到动词，因为本体知识条目的关系必然是动词。而动词在 CFN 标注中有相应的框架，从而可以找到具有语义的重要信息。

本文利用“到达”、“穿越”、“出发”、“位移”、“存在”5 个框架，对询问交通路线或者交通工具的问句进行分析，同时利用框架中的词元对动词进行了同义扩展。其中前四个框架主要用于特指疑问句的分析。由于“存在”框架下的词元（一个义项下的一个词）主要有：存在 v，真实的 a，存在 n，在 v，有 v，没有 v，无 v，生存 v，所以“存在”框架主要用于对是非疑问句的分析。

CFN 第一层可以把交通工具以及出发点和目的地很快的识别出。表 2 为旅游交通领域的部分问句标注示例。

表 2　部分问句标注示例

问句类型			问　句	疑问词	CFN 标注
特指疑问句	第一大类	LOC（地址）	从太原到五台山沿途经过哪里？	经过＋哪些＋地方	〈Area-vp-va 从太原到五台山〉沿途〈tgt＝穿越 经过〉〈path-np-obj〉哪些地方？ ：[穿越/Traversing 框架]
		OBJ（实体）	开车从北京出发去五台山，大概路线是什么？	路线＋是＋什么	〈car-vp-va 开车〉〈src-vp-va 从北京出发〉〈tgt＝位移 去〉〈goal-sp-obj 五台山〉，大概路线是什么？ ：[位移/Motion 框架]
					〈mot-vp-va 开车〉〈src-pp-adva 从北京〉〈tgt＝出发 出发〉〈goal-vp-va 去五台山〉，大概路线是什么？ ：[出发/Getting_underway]
		TIME（时间）	开车从北京出发到五台山，多长时间？	多长＋时间	〈car-vp-va 开车〉〈src-vp-va 从北京出发〉〈tgt＝到达 到〉〈goal-sp-obj 五台山〉，多长时间？ ：[到达/Arriving 框架]
	第二大类	DES（描述）：方法	我在西安，要去山西五台山怎么走？	怎么＋走	〈thm-np-ext 我〉在西安，〈goal-vp-va 要去山西五台山〉〈manr-dp-adva 怎么〉〈tgt＝位移 走〉？ ：[位移/Motion 框架]
			驾车从太原到五台山路怎样走最近？	怎样＋走＋最近	〈mot-vp-va 驾车〉〈src-pp-adva 从太原〉〈tgt＝到达 到〉〈goal-sp-obj 五台山〉路怎样走最近？ ：[到达/Arriving 框架]
是非疑问句	第三大类	Unknown（未知）	去五台山旅游是否有比较省钱的线路？		〈contain_even（通用非核心元素“所属事件”）-vp-va 去五台山旅游〉〈manr-dp-adva 是否〉〈tgt＝存在 有〉〈ent-np-obj 比较省钱的线路〉？ ：[存在/Existence 框架]

6.3 本体三元组的抽取

首先从问句的动词中获取语义谓词，并进一步与本体库中的条目关系进行比对。通过语义指数[14]来衡量语义谓词的重要性，基于规则评分[14]后提取语义谓词的主体和客体。

例句："驾车从太原到五台山怎样走最近?"

例句的CFN标注为：

〈mot-vp-va 驾车〉〈src-pp-adva 从太原〉〈tgt=到达 到〉〈goal-sp-obj 五台山〉怎样走最近? 通过标注得到传送模式：驾车(即交通方式：自驾车)。同时得到出行的出发点：太原和目的地：五台山。

通过多角度的问题分类和本体思想的结合，问题分类更加合理化，从而可以准确识别问句类型。同时通过CFN中的标注，提供了具有语义的重要信息，从而减少抽取三元组的时间，使问题类型识别效率提高。

7 用户询问类型及其处理策略

对大量不同问句分析后，归纳出具有同一语义的问句中短语之间的搭配关系和次序的不同情况，建立了问句模版(QM)。例如类型为"方法"的问句中交通工具"到达"的模板如下：

a QM(arrive)=(sp)+(txt=到)+(sp)+(qw=怎么)+(vp=去)?

b QM(arrive)=(sp)+(txt=到)+(sp)+(qw=如何)+(vp=去)?

c QM(arrive)=(vp=从…)+(qw=怎么)+(txt=去)+(sp)?

d QM(arrive)=(qw=如何)+(vp=从…)+(txt=到)+(sp)?

e QM(arrive)=(qw=如何)+(vp=从…)+(txt=去)+(ns)?

其中sp为处所短语，qw为疑问词。此模板都是选取"到达"框架的词元作为目标词，由于五个问句属于同一问题的不同询问方式，定义了同一个答案抽取的规则。答案都是返回本体库中的车类型以及路线。但是由于库中的问句库数量有限导致模板中句子数量有限，所以在当用户的问句在模板库中不存在时，涉及到计算问句和库中的此类模板问句之间的语义相似度计算。

根据问句库的统计，目前用户的问句类型分为以下三类：

(1) 简单的问本体的主体，客体。包括特指疑问句和是非疑问句中询问人物、时间、数字、实体。

如：五台山的气候怎么样? 五台山附近有没有旅馆?

(2) 询问方法，属于大类：描述。

如：开车从北京出发去五台山，怎么去?

(3) 原因、定义类的问题

如：为什么五台山是我国四大佛教名山之首?

8 答案的提取

本文采用了SPARQL语言和Jena推理机来进行答案的查找[16]，SPARQL[17]语言提供本体查询功能。具体的查询流程如图2所示：

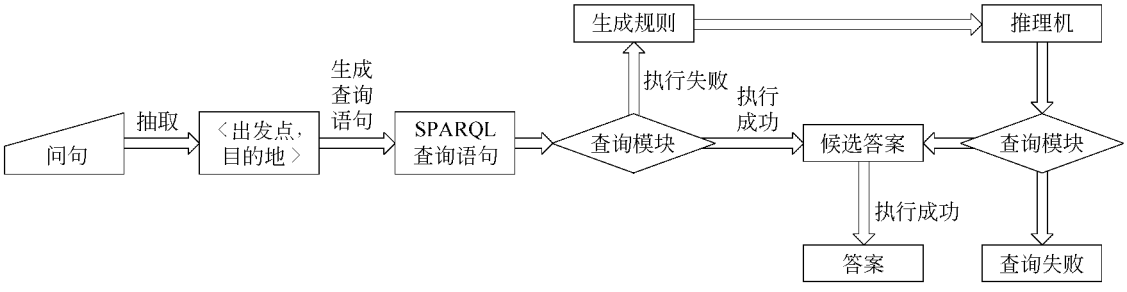

图2 答案查询流程

例如：当用户输入一个查询"从包头怎么去五台山?"，生成SPARQL查询语句，如下：

SELECT ? car_kind ?car_path WHERE{?x Travel:traffictool_end ?traffictool_end. ?x Travel: traffictool_start ?traffictool_start. ?x Travel: car_path ?car_path. FILTER(?traffictool_end="包头" && ? traffictool_start="五台山";

系统调用查询模块执行SPARQL，如果查找

成功，直接生成候选答案集。如果查找失败，则生成相应的查询规则（如：Rule：(?y traffictool_end ?x)，(?y arriver at ?z) —>(?y traffictool_start ?x)），并创建推理机，进行推理，同时生成相应的数据模型，再次调用查询模块进行查询，并生成相应的候选答案集（如果查询再次失败，则返回空答案）。最后，调用排序模块对候选答案集进行排序，并返回给用户。例子的返回答案为：

1. 火车 1674/1675：包头——忻州 火车 2462/2463：包头——忻州 大巴：忻州——五台山

2. 飞机 MU5690：包头机场——太原武宿机场 大巴：太原——五台山

3. 大巴：包头——太原 大巴：太原——五台山。

9　实验结果与分析

目前，课题组已构建了 130 个框架，涉及动词词元 1 428 个、形容词词元 140 个、事件名词（即有配价的名词）词元 192 个，8 200 条句子。

依据山西旅游景点网站，本系统目前收集了 1 566 条旅游常问问句，同时用本体语言 OWL 进行了描述。问句包括六个方面：特色小吃、住宿、娱乐、景点、购物、交通工具。

在 eclipse3.1.2 平台上进行了实验系统的实现，并使用了 jena2.3 工具包。本系统中主要涉及到了表 3 给出的旅游交通方面的四类型问句，并对加有 CFN 标注的问句进行了测试。目前本系统主要针对特指疑问句中的地点、时间、方法及是非疑问句部分问句进行了测试。实验结果如表 3，其中分类 1 为 TREC 分类，分类 2 为 TREC 分类与本体结合。采用召回率对系统进行评价，定义如下：

召回率＝（回答正确的句子/实际存在的正确样例）×100%

由表 3 可以看出，系统回答问题的召回率还较低，误差分析如下：

表 3　实验结果

问句类型	数量	召回率			
		分类 1	分类 2	分类 1 CFN 标注	分类 2 CFN 标注
特指疑问句：LOC	76	68%	68%	70%	72%
特指疑问句：TIME	65	67%	68%	71%	73%
特指疑问句：MEANS	183	65%	67%	68%	70%
是非疑问句	268	58%	59%	60%	62%

第一，分词以及词性标注对问句的分类有影响。由于旅游中景点名比较多，导致在山西大学现代汉语自动分词系统进行切词和 CFN 框架标注以后，有可能会产生分析错误。例如：中台顶的日出真的很奇特吗？分词结果为：中台 /ns 顶 /v 的 /u 日出 /v 真的 /aq 很 /d 奇 /r 特 /n 吗 /u ？其中中台顶是一个地名，但分词的错误导致本体库中找不到此地名。

第二，有些问句分类的语义类别很难确定是哪一类。例如："开车从北京出发去五台山，大概路线是什么？"应该分为"地址"，还是"实体"类别？类似这样的类别难于确定的问题给分类也会带来了精度的损失。

第三，本体知识库的模型的建立不完善，以至于机器对问句的理解和推理有误差。例如："从西藏怎么去五台山？需要推理得到中转站，系统是根据库中的信息来进行推理，库中路线资源如果很少，则可能提取不出答案。

第四，本体库中定义的规则的合理性。例如："从包头怎么去五台山？本系统抽取答案目前还没有考虑最优路线的问题。

第五，原因类问题无法做出回答。例如："五台山住宿为何那么贵？"原因类问题由很多因素导致，此类问题现阶段无法解决。

10　总结

本文针对旅游特定领域，对基于汉语框架网的旅游信息问答系统进行了探索。设计了基于汉语框架网的旅游信息问答系统构架，以旅游景点五台山

为例，初步构建了旅游本体模型，采取了多角度分类形式，在TREC分类的基础上，利用本体的思想，对问题分类。并对交通工具一类问句进行了基于CFN分析，提出了问题类型为“方法”的“到达”等框架模板，有效提取本体三元组及答案抽取。

旅游问答系统涉及面广，工程量大，需要做的基础工作很多，我们还面临很多问题：(1)针对旅游领域CFN的框架研究与补充；(2)旅游本体模型进一步完善；(3)框架所对应模板进一步补充和完善；(4)就交通工具的其他方面进一步探索。

参考文献：

[1] J Pierre, Z Pierre. Towards a medial question-answering system: a feasibility study [C]//Pierre Le Bux and Robert Baud eds. : Proceeding of the medical Information 2003: 463-468.

[2] R Fagin, P G Kolaitis, R J Miller et al. data exchange: Semantics and query answer[C]//Proc of the 9th Int Conf on Database Theory(ICDT2003), 2003.

[3] Gruber T R. A Translation . Approach to Portable Ontology Specifications[J]. Knowledge Acquisition, 1993(5): 199-220

[4] Smith M K , Welty C, McGuinness D. OWL Web Ontology Guide Language[EB/OL]. http://w-ww. w3. org/tr/2003/WD-owl-guide-20030331.

[5] 郝晓燕，刘伟，李茹，刘开瑛. 汉语框架语义知识库及软件描述体系[J]. 中文信息学报，2007，21(5)：98-138.

[6] 刘开瑛，由丽萍. 汉语框架语义知识库构建工程[C]//中国中文信息学会二十五周年学术会议，2006：64-71.

[7] Charles J. Fillmore. Frame semantics and the nature of language[C]//Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech. 1976, 280: 20-32.

[8] Charles J. Fillmore, Charles Wooters, and Collin F. Baker. 2001. Building a large lexical data bank which provides deep semantics[C]//Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation. HongKong, 2001: 3-26.

[9] Charles J. Fillmore, Collin F. Baker et al. The Berkeley FrameNet project[C]//Proceedings of COLING/ACL, Montreal, Canada, 1998: 86-90.

[10] Noy, N. F. , M. Sintek, S. Decker. Creating Semantic Web contents with Protege-2000[J]. IEEE Intelligent Systems and Their Applications. 2001, 16(2): 60-71.

[11] Guohua Shen, Zhiqiu Huang, Xiaodong Zhu, Lei Wang, Gaoyou Xiang . Using Description Logics Reasoner for Ontology Matching[C]//IITA Proceedings of the Workshop on Intelligent Information Technology Application, 2007: 30-33.

[12] 文勖，张宇，刘挺，马金山. 基于句法结构分析的中文问题分类[J]. 中文信息学报，20(2)：33-39.

[13] Del Zhang, Wee Sun Lee. Question classification using support vector machines [C]//26th ACMS IGIR. 2003.

[14] 董慧，余传明，姜赢，杨宁，等. 基于本体的数字图书馆检索模型研究(Ⅱ)[J]. 情报学报，25(4)：451-461.

[15] 丁晟春，顾德访. Jena在实现基于Ontology的语义检索中的应用研究[J]. 现代图书情报技术，2005，3(10)：129-134.

[16] SPARQL Query Language for RDF W3C Candidate Recommendation 6 April 2006 [EB/OL]. http://www. w3. org/TR/2006/CR-rdf-sparql-query-20060406/.

[17] Seaborne A. 2004. Jena tutorial a programmer's introduction to RDQL[EB/OL]. http://Jena. sourceforge. net/tutorial/ RDQL/.