基于汉语框架知识库的旅游信息问答系统设计

王文晶，李茹，宋小香

（山西大学 计算机科学与技术学院 山西 太原 030006）

Email: wwjlhk0624@yahoo.com.cn

摘要： 本文在汉语框架语义知识库的基础上，利用语义WEB语言，通过汉语框架语义知识库对问题进行语义分析，并利用旅游本体知识库对答案进行抽取并对答案处理。山西大学建立了1004条旅游问句库和旅游本体模型，同时用本体编辑工具Protégé编码。

关键字： 本体; 汉语框架语义知识库; 自动问答

Design of Tourism Question Answering System Based on the ChineseFrameNet knowledge database

WANG Wen-jing, LI Ru, SONG Xiao-xiang

(School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006)

Email: wwjlhk0624@yahoo.com.cn

Abstract: This article has constructed a Chinese FrameNet, and describes it with the semantic WEB language. Carries on the semantic analysis through Chinese FrameNet to the question, and carries on the answer-retrieval and to answer processing using the traveling Ontology system. Shanxi University has built 1004 traveling interrogative sentence storehouse and the traveling ontology model, simultaneously uses the main body editor tool Protégé code.

Key words: ontology; ChineseFrameNet; Question-answering Automation

1 引言

虽然搜索引擎如Google、Yahoo、百度等在飞跃发展，但目前对海量信息的能力却还很差。现有的旅游信息系统是由旅行社和旅游管理部门提供的信息网站并且由他们封闭管理，同时对用户问题的回答准确率都很低。目前因特网在信息表达和检索方面存在的缺陷，主要是没有提供给计算机可读的信息，所以限制了计算机在检索中的自动分析能力。

为了使旅游信息解决系统之间异构的问题，以及使旅游问答系统更具有智能化，同时为了解决可读的信息，我们引入了本体和语义WEB的思想。

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[1]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)[2]语言是W3C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。 所以首先

基 金 项 目： 国家863高技术研究发展计划资助项目(2006AA01Z142)

作 者 简 介： 王文晶 (1981—)，女，硕士生，主要研究方向为自然语言处理，李茹(1963-)，女，教授，研究方向为智能信息处理。宋小香(1984-)，女，硕士生，研究方向为自然语言处理;

从用户问题为问题出发点，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，研究构建一个汉语框架语义（Chinese FrameNet，简称 CFN）[3]。它由框架库、句子库和词元库三部分组成，使用 XML(可扩展的标记语言)、RDF（资源描述框架)、OWL(网络本体语言)语言对资源进行描述。并且用语义 Web 标记语言表示该语义知识库的各种资源，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。

目前，山西大学 CFN 课题组已就汉语 1760 个词元构建了 130 个框架，涉及动词词元 1428 个、形容词词元 140 个、事件名词（即有配价的名词）词元 192 个，标注了 8200 条句子。

该系统设计了基于本体的山西旅游信息知识库，并用本体语言OWL进行了描述。对用户的问题，利用汉语框架语义知识库（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 系统构架

本系统的构架如图1所示，它主要包括：预处理模块，问句匹配以及在旅游知识本体库中的答案抽取模块，答案处理模块。如图1所示，面向山西旅游信息的自动问答大致流程是：

1、 提交问题：用户先把查询请求包装在一个SOAP消息中，然后提交该信息给HTTP服务器。

2、 预处理：随后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都用到CFN以及专业领域库。专业领域库是旅游领域的知识条目，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。

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（框架元素），并进而描述该框架和其他框架之间的概念关系；句子库记录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，它可以作为训练数据供计算机处理语言使用；词元库是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自动生成的。

同时，使用 WEB 语言描述了 CFN 资源。用 RDF 描述 CFN 词汇语义资源，XML 标记 CFN 数据库的文档内容。同时还对标注的句子也进行了编码。CFN 提供的框架元素数量多、类型细化，突出了框架的个性，在语义表示深度上具有明显的优势。

3、问句匹配：由于问答系统中存在口语词汇较多，所以建立了旅游领域中的词汇对应的口语词汇词典，以便更好的语义理解。在进行简单的语义分析之后，通过关键字的粗略提取，利用旅游扩展词库的提取出用户的查询要求，同时结合cfn中框架的语义可以高效率的提取出有用的相关信息。从而确定了检索的类型以及检索的策略。看查询要求是否能和问句例库和问句模板库中的类型进行匹配。

4、语义知识推理:进行答案的查找。入口是转化生成的RDF三元组问句向量，然后利用本体知识库中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服务器把结果包装成SOAP消息发送给用户界面。

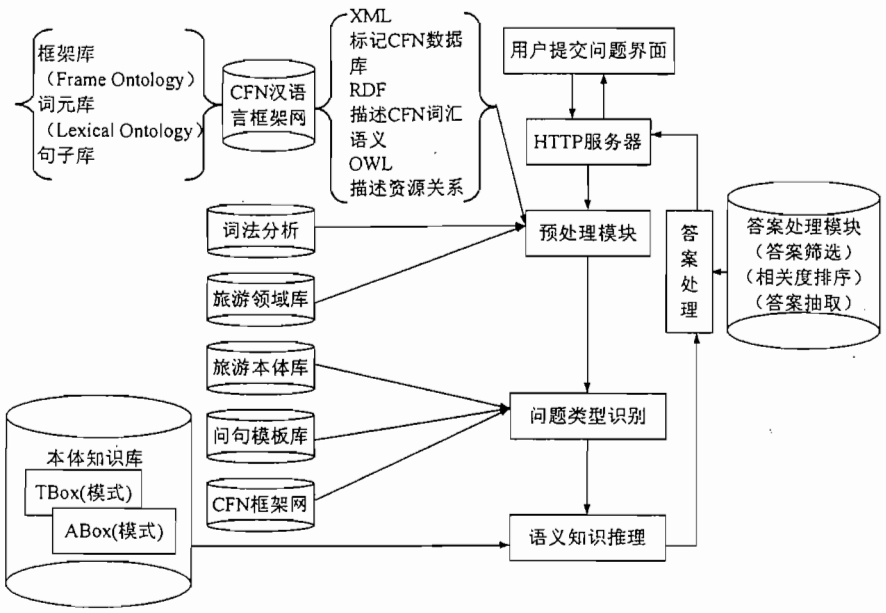

图 1 旅游信息问答系统构架

3 旅游信息本体库的构建

本文面向山西旅游 QA 信息系统，依据 Studer[4]提出的本体是“共享概念模型的明确的形式化规范说明”中定义的四层含义“概念模型、明确、形式化和共享。同时依据山西旅游景点网站收集的 1000 条旅客常问问句，构建了山西旅游领域的本体，并且将其作为自动问答系统的知识库。

要设计出一套完整的旅游本体图，使的不同的代理人能够利用本体库来互相沟通，协助完成使用者完成旅游行程的规划。所以在建立本体库之前，要下载省内所有的旅游线路（标出价格），并能进行实时更新。可以通过 http://www.5566.net/jt-.htm 来下载。以及下载省内区间内铁路、航运和客车情况，（标有价格）并能进行实时更新。还有各个城市的天气情况，可以通过点击 http://www.cma.gov.cn/查询。通过旅游网址 http://www.lywzz.com 得到省内各个旅行社的路线，以及详细的旅游信息。图二为旅游本体中的概念以及关系图，考虑到类的继承把住宿中的服务和房间分别定义为不同的类，门票和景点定义为不同的类。其中定义了关于景点的 6 个大类，分别为：特色小吃、住宿、娱乐、景点、购物、交通工具。图三为这 6 类（概念）之间的关系模型图。

本系统是采用采用了OWL Lite进行本体模型的编码和美国斯坦福大学的本体编辑工具 Protégé[5] [6]。图2为Protégé中建立的类，分为旅游地的风俗、导游服务、住宿、购物、交通、门票、景点，同时定义了类之间的对象属性ObjectProperty以及类自身的数据类型属性 DatatypeProperty。本体的建立严格要求定义了类之间的逆关系、传递关系、函数关系、对称关系、逆函数关系以及对属性的限制。如定义has_tickets的逆属性为be_sold，由于这对属性的定义域和值域相互可以调换。逆关系可以获取正反两方面的信息，如果知道五台山门票的情况，推理机可知符合条件的门票的景点是五台山。景点和酒店之间的关系为near，此属性为对称属性。

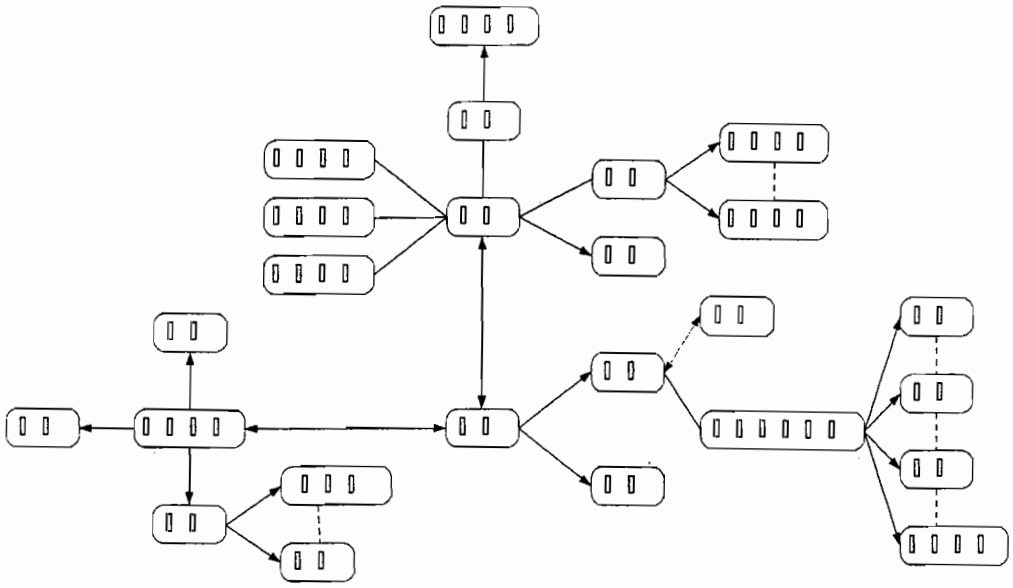

图 2　旅游领域本体模型

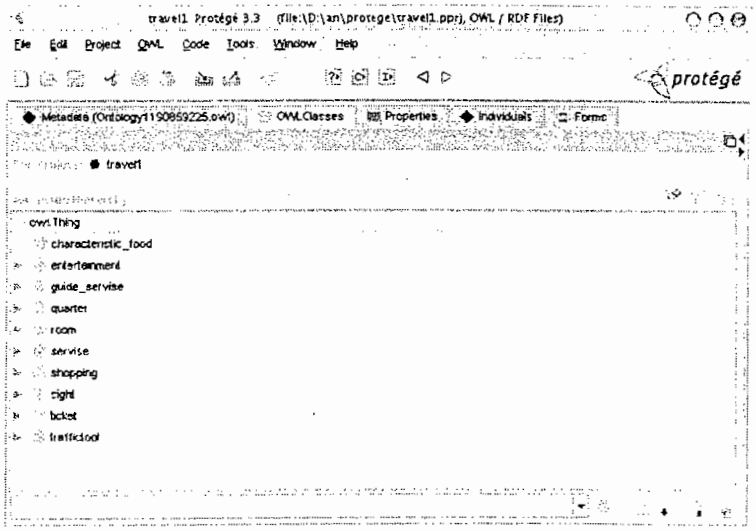

图 3 旅游本体库分类图

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。同时利用Protégé中的RacerPro推理机进行推论，RacerPro推理机在辅助建模阶段有很大的作用。可以用于检查一致性、推理出新的分类体系等。

旅游问题并不是一个单纯的问题，除了逐一检查具有高的复杂度的巨量空间路径以外，还需要考虑空间之间的路径与旅游者所提出的各种条件组合哪一个为最佳。为使旅游者提出的各种资

源都考虑到，就要对旅游资源的影响因素进行分析，比如说空间（即符合起点到终点条件的路径有好多条），资源（符合旅游者提出时间和预算的多个结果中，找出最符合的组合排列）。便会涉及到一些评估的方法[7]。

4 问题分类及答案抽取

4.1 问题分类

问题分类中，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。最重要的是从内容上分，可以直接利用top—lever ontology的概念分类体系，较全面多层次进行分类。从不同的角度分析问句，有利于更准确地把握问句的含义，从而保证最终抽取出正确答案。 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会议提出的7大类60小类[8]合的基础上，利用本体的思想，对问题分类，分类流程图如下：

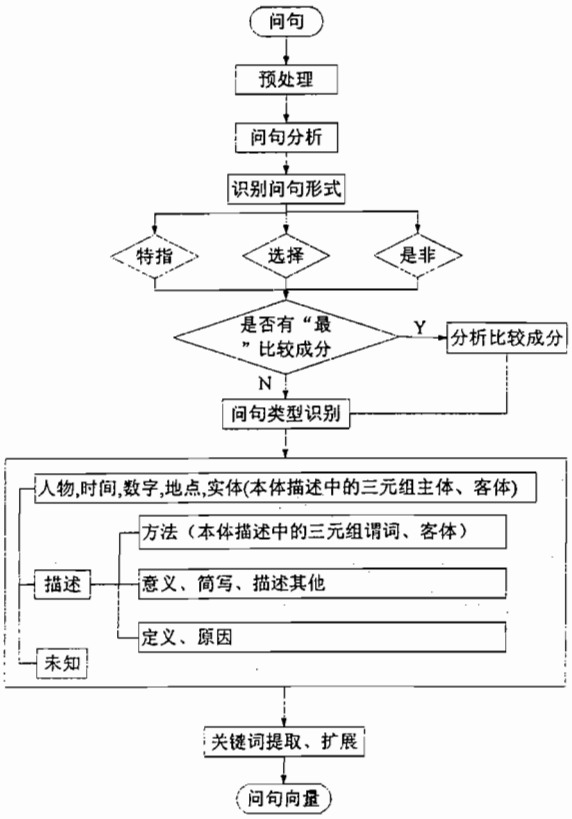

图 4 问句分析流程图

问句分类，建立在 7 大类 60 小类与本体相结合的基础上。问句先进行问句形式识别，然后识别问句中是否有比较成分，如果有，需要分析比较成分。然后再对问句进行类型识别。由于在本体库中，资源的描述用 rdf 描述，即三元组：主体-谓词-客体。所以问句类型分为三大类：第一大类：在本体库中，所以人物、地点、时间、数字、以及旅游领域中的实体，都属于对客体的提问，实体的提问也可能是主体可能是客体。第二大类：描述类，其中又分为三类：a 方法。此

类问题有时需要语义转化为主体的谓词。如：疑问词：“怎么走”需要转化为谓词“路线”。b 意义、简写、描述 c 定义、原因。在旅游本体库中，此类属于对三元组中谓词。第三大类：未知。

目前对收集到关于旅游景点五台山的 1004 条问句进行了类别统计。如表-1。

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	188	18．7%
	特指	754	75．1%
	选择	62	6．2%
7 大类 60 小类与本体结合	第一大类：人物、地点、时间、数字、（旅游领域）实体	562	66．0%
	第二大类：描述	202	20．1%
	第三大类：未知	89	8．9%

4．2 用户询问类型及其处理策略

目前此系统大致可以识别以下 3 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？

对于问题（1）可以识别出简单的三元组<五台山，气候，属性值>，然后提取出来属性值。对于第二个问题通过类型识别器识别是否存在次三元组<五台山，附近，旅馆>如果有的话，则回答有。

对于问题（2）通过问题器中对疑问词和疑问意向词的识别出问句的问点是询问路线，找出把三元组<汽车，出发点，北京> 和<汽车，目的地，五台山>属性值与系统中所有的汽车子类自驾车实例，RDFS:lable 进行匹配，如果匹配成功，则列出所有符合条件的 RDF 结点。如果没有则告知用户没有。最后把满足条件的实例的路线属性值返回。

对于问题（3）需要从领域文本中提取，所以问题库中专门收集了此类问题的文本回答。所以本系统对此类问题的回答有很大的局限性。

对于问句中有比较成分的问句，首先分析比较的成分，经过对此类问句的分析之后，总结出不同类型的问句模板。如特指疑问句：“自驾从北京 到五台山，走哪条路最好？”，类似此问句的模板为：疑问词+疑问意向词+比较成分（d+aq/d+a/nt），其中 d 为副词，aq 为性质形容词，a 为形容词，nt 为时间名词。d 一般为“最”“比较” “更”等副词，aq 一般为“方便”“合适”等性质形容词。nt 一般为“季节”“最近”等时间词。比较成分为 aq 最近的成分。

4．3 汉语语义框架与本体三元组提取

汉语语义框架 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，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。例如对于询问交通路线或者交通工具的问句中关于交通路线的句子大多有“到达”“去”等词语，同时框架中的词元对动词进行了同义扩展。

对用户提出的问题首先通过山西大学开发的“现代汉语自动分词系统”进行切词，切出来的字符串词组经过二次标注传到问题类型识别器进行识别，从而识别出问题类型。问题识别器中主要是依据问句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。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。例如在关于交通的问句中，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。由于问句与陈述句之间的差异，而 CFN 目前标注的主要是陈述句，所以问句的框架元素多数都缺省。表 2 简略展示了“到达”框架的内容。

表-2　“到达”框架

框架名	到达	
定义	指转移体朝目的地方向的移动。目的地可直接表达出来，或从上下文中得到理解，动词本身隐含目标之义。	
核心框架元素：	目的地 Goal [Goal]	目的地表现的是转移体运动终止之地，或行将终止之地。 ●在 9 点午夜前到了五台山。
	转移体 Theme [Thm]	转移体指移动者，它可以是一个能够自动的实体，虽并非往往出于必需。 ●警官朝房子靠近。
非核心框架元素：	并行转移体 Cotheme [Thm_c]	并行转移体指第二个移动的物体，常表现为直接宾语或间接宾语。 ●导游和旅客一起来到乔家大院街上
	形容 Depictive [Dep]	形容指用来描写转移体到达时状态的话语。
	目的地状态 Goal_conditions[G_c]	转移体到达目的地时目的地所呈现出的状态。 ●参议员走到长时间起立鼓掌的人群前。
	修饰 Manner [Manr]	表现修饰的话语用于对动作特性的描述。该动作并不直接与动作的轨道发生关联，用来描述运动的速度、恒性、姿态和方法以及其他情况的描述性用词都可以看作表现修饰的话语。
	方法 Means [Mns]	该框架元素用于表现转移体到达的方式。
	传送模式 Mode_of_transportation [MoT]	传送模式指作用于主体的运动模式，通过传送主体的主体身体或交通工具而实现。交通工具可以以任何方式运动，通常以 in 或 by 得以间接地表达。●自驾车到五台山怎么走？
	轨道 Path [Path]	轨道指运动的轨道，既非源点，也非目的地。在该框架中，表现轨道的用语差不多都有一个经过义。
	源点 Source [Src]	源点即明确表达运动的出发点，该框架中出现表达源点的用语是可能的，但出现的频率却相对不高。如果出现，常用来表现转移体发出运动的大致方向，而不会是远离该方向的一个地标。
	时间 Time [Time]	该框架元素表现到达这一动作出现的时间。
词元	到达 v，来到 v，进入 v，抵达 v，返回 v，走到 v，走进 v，赶到 v，回来 v，归来 v，到 v，回到 v	

例如问句 “驾车从太原到五台山路怎样走最近？”首先经过预处理，经过问题识别器识别之后属于第二大类：描述中的方法类，同时分析出比较成分：路线属性值。然后需要形成三元组从owl文件中读出所以需要满足的实例，并进行比较。即满足<自驾车?，出发点，西安>，<自驾车?，目的地，五台山> 的汽车子类自驾车的实例，然后对所有实例的路线属性值进行比较。而三元组<自驾车，出发点，西安>，<汽车，目的地，五台山>的抽取就利用了cfn的标记。

依据表 2 框架结构，以及 cfn 的标注规范。进行 cfn 标记如下：

<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? 通过标注可以得到传送模式：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。

目标词的选择直接影响到提取语义，例如问句：“太原到五台山乘什么交通工具便捷?”选择“到达”框架，标注结果为：

<src-sp-subj 太原> <tgt=到达 到> <goal-sp-obj 五台山> 乘什么交通工具便捷?

如选择“乘”为目标词，标注为：

<path-vp-va 太原到五台山 > <tgt= Ride_vehicle 乘> <Vehicle-np-obj 什么交通工具便捷?可见目标词的选择尤为重要。

通过结合多角度的问题分类和本体的思想，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，使问题类型识别器效率提高。

4．4 答案的提取

用户的问题经过分析之后，得到的是若干的实体和属性，或者是实体对，而与库中三元组的匹配，就要用到能够解析和查询RDF模型的工具包，所以我们用到了开发工具Jena。Jena是一个Java开发的工具包，Jena本体解析器可以对RDF的解析，对RDQL的查询支撑和对OWL的解析。Jena同时提供基于规则的推理机。

推理机的工作原理是：推理机注册机制根据基本RDF 三元组描述和Ontology 模型创建出推理机，由此推理机可以生成包含推理机制的模型对象(Inference Graph，InfGraph)，在Jena中，图(Graph)也被称为模型(Model)，而表现形式为模型界面(Model Interface)，然后可以使用Model AP I 和Ontology API 对此模型进行操作和处理，从而实现语义层面的信息检索。

通过 Jena 推理得出本系统的本体库中所需要的全部直接关系与间接关系，所有的关系都是三元组的形式（<subject , property , object>），然后把这些以三元组表示的关系存入的数据库中，这样避免了在检索时进行推理。将全部关系存放在数据库中，当需要进行本体推理时，使用 SQL 检索从数据库中获取相应的知识，这样可以使系统的运行效率提升。

使用Jena推理机制实现这种查询功能，必须做好两方面的工作。第一，把图2中所有与数据库相关的概念、关系和实例用OWL和RDF表示出来，存储成XML文档。这又分两个步骤：(1)把图2中本体的概念用OWL表示出来，存储成concept．owl文件。把图2中本体的实例用RDF存储instance．rdf文件。第二，根据搜索条件构造规则来实现搜索功能。

例如：“我在天津，如何驾车如何到达五台山的鑫隆贵宾楼？而本体库中没有到达宾馆的交通方式，但是由于鑫隆贵宾楼位于五台山风景名胜区，所以可以理解为如何到达五台山。根据查询条件构造一条推理规则：

Rule : (?x is near ?y), (?y arriver at Wutaishan by ?z) —>(?x arriver at nearby hotel by ?z)

把自定义的规则加入推理机中，然后把concept. Owl可以本体概念文件和描述信息资源的instance. rdf文件读入到推理机。同时利用Jena推理机中的printStatements函数输出所有与数据库有关的推理结果。

五 结论

旅游信息系统具有典型的信息分布分散和信息形式多变的特点，本系统解决了山西旅游信息系统的异构问题，并且利用我们自己构建的语义知识库CFN，解决了交通问句中的语义的问题。但是目前我们的问答系统的知识库是靠人工加工的，需要很大的人力和物力。同时知识库有限的容量，对旅游者提出的问题可能有局限性。所以如何能够把网上动态的旅游信息加入知识库是很有挑战性的。基于本体的全球旅游信息问答系统是计算机和旅游产业的热点，同时实现具有语义智能化的搜索技术也成为发展的方向。

参考文献

[1] Gruber T R. A Translation Approach to Portable Ontology Specifications. Kn-owledge Acquisition, 1993(5):199～220

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

[3] 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 [A], 中文信息学报，2007，21（5）：98～138.

[4] Studer R, Benjamins V R , Fensel D. Knowledge Engingeering , Principles and M-ethods. Data and Knowledge Engingeering, 1998, 25(1-2):161～197

[5] Informatics, S. M. Protégé [EB/OL]. http://protege.stanford.edu/, 2006

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

[7] 王治立，陈鸿文. 旅游语义网整体服务系统之建制[D] . 台湾：大叶大学，1993.

[8] DelZ hang, We eSu nL ee. Q uestioncl assificationu singsupportv ectorm achin-es[A] . In:the26 thA CMS IGIR[C]. 2003.