问答式移动空间信息服务中的本体实例搜索①

乐小虬② 杨崇俊* 刘冬林* 于文洋*

(中国科学院文献情报中心 北京 100080)

(*中国科学院遥感应用研究所遥感信息科学国家重点实验室 北京 100101)

摘 要 从地理本体入手,构造了空间语义角色和实例模式,通过空间语义角色标注、短语识别以及模式匹配等手段,从Web中提取出与空间位置相关的本体实例,为问答式移动空间信息服务提供了可动态更新且具有明确语义的数据源,并利用语义Web技术,为移动用户提供问答式本体实例查询服务。初步实验表明,该方法具有较好的准确率,召回率有待进一步提高。

关键词 移动空间信息服务,搜索引擎,本体实例

0 引 言

为移动用户提供随时随地、快速准确的信息服务一直是近年来研究和运用的热点。问答式移动空间信息服务(以下简称问答式服务)利用无线网络平台,以问答的形式为移动用户提供与地理位置相关的诸如地图、图像、相关文本等多种信息服务。同菜单式服务相比,问答式服务能够较好地解决因无线网络带宽窄,移动终端数据处理能力弱,屏幕显示受限等瓶颈造成的人机交互能力差,运用平台受限以及结果查找烦琐等问题。问答式服务适用于当前常用的SMS/MMS和WAP两种无线网络应用模式,以准确答案的形式返回结果信息不仅能降低数据传输量,而且能减少用户在检索结果中进行二次查找的过程。因而从理论上说它是移动空间信息服务的理想模式。但由于自然语言理解(NLP)等技术的复杂性,目前真正实用的应用系统还不多见,本文是对其实用化方法的一种探索。

问答式移动空间信息服务涉及问答系统(QA)、移动地理信息服务(MGIS)、地理信息检索(GIR)等方面内容。这些内容在其各自领域内均已有大量研究。在将这些内容相融合为一个整体并提供问答式服务方面,文献[1-3]及一些短信服务企业开展了相关的研究,但服务的内容基本局限在GIS库中的空间数据和属性数据上,对于互连网中大量以文本方式存在的空间信息及其关联信息则没有充分利用。如何将这部分信息从Web中自动挖掘出来并重新组织成有效数据源,对于问答式移动空间信息服务而言,不仅能动态扩充和更新服务内容,而且能使关联信息向不同的领域拓展。这种思路在QA系统中较为普遍[4],但目前专注于空间信息领域的这类研究还不多见。虽然在一些信息提取(IE)系统中有些关于WEB中空间信息/知识提取的研究[5],但由于研究的目标不同,所提取结果语义贫乏,很难满足问答式移动空间信息服务系统的要求。

本体通常被定义为某一领域内的共享概念集,它用结构化语言(XML/RDF/OWL)进行描述[6]。Femke等在文献[7]中创建了地理本体,它是OWL形式,其中蕴涵丰富的空间信息领域基本概念、关系和属性。这对于着眼于回答与地理位置有关的问答式服务系统而言意味着用户的大部分查询请求应与地理本体有关,以地理本体为基础搜索到的本体实例具有确定的语义结构,能为机器所理解,所以可以直接作为用户查询请求的备选结果,而且利用语义Web(SW)中的推理器还可使系统达到智能化服务的目的。基于这一思路,以下将具体讨论其实现原理、方法及试验结果。

1 本体实例搜索原理

1.1 地理本体描述

地理本体是地理空间概念的集合。本文参照文献[7]中的概念集,采用四元素描述方法,对地理本体作如下定义:

① 863计划(2003AA135119)和973计划(G2000077906)资助项目。

② 男,1968年生,博士生,工程师;研究方向:智能搜索引擎,WebGis;E-mail:xiaoqiule@yahoo.com.cn

(收稿日期:2005-09-14)

定义1 地理本体 $GeoOnto = \{C_S, A_c, R_S, H\}$，其中，$C_S$ 代表空间概念集合，C_S = {事物，空间实体，地名，建筑物，河流，湖泊，道路，桥梁，城市，村镇，居民地，宾馆，饭店，医院，机关驻地，学校，商店，公路，铁路...}；A_c 代表每一个空间概念的属性集合，A_c = {中心点坐标，长度，面积，人口，级别，规模，形状，事件，...}；R_S 为空间关系集合，R_S = {空间关系，拓扑关系，方向关系，度量关系，相等，相离，相交，相接，穿越，内部，包含，重叠...}；H 代表空间概念的层次体系，H = {(事物：(空间实体：(道路：公路)(地名：城市)(建筑物：宾馆)...))}。

1.2 实例定位语言

WEB中的正文信息多以自然语言的形式表达，它与结构化的本体语言存在很大差异。地理本体中表达的一个概念在自然语言中存在多种表达形式。如果直接以地理本体元素为基础采用模式匹配的手段提取实例，不仅执行效率低下而且会因自然语言的复杂性而使模式构建难以进行。所以本文采用地理本体角色标注的手段定位文本中的空间实例，并对定位语言做如下定义：

定义2 *GORTL*（地理本体角色标注语言）：$GORTL = \{SE_h, A_h, R_h, LA\}$，其中，$SE_h, A_h, R_h$ 分别对应 *GeoOnto* 中的 C_S, A_c 和 R_S，表示空间实体、实体属性和空间关系三种语义角色的标记符集合，下标 h 标记各角色中的类型和层次关系，用编码表示；*LA* 为词性标记符的集合。

如，C_S 中的“城市”被标注为 $/SE_{0101}$，“村镇”标注为 $/SE_{0102}$，“公路”标注为 $/SE_{0201}$ 等等，*LA* 选用北大语言所制定的现代汉语语料库加工规范。

定义3 *Reg*（提取规则）：由 *GORTL* 和非终结符（见定义5）构建的用于识别本体实例的正则表达式的集合。

定义4 *SSD*（空间语义字典）：$SSD = \{L^E, L^A, L^R, F, G, J\}$，其中 L^E，L^A，L^R 分别表示空间实体、属性、空间关系的关键词项；F，G，J 为关键词项到 *GORTL* 中 SE_h, A_h, R_h 的映射函数，其中 $F \subseteq L^E \times SE_h$，$G \subseteq L^A \times A_h$，$J \subseteq L^R \times R_h$。

1.3 实例识别语法

自然语言中的概念通常是通过词、短语和句子三个层次表达出来的，Web中的文本经地理本体角色标注后基本锁定了实例出现的位置，但这只是实例元素级处理，还无法得到更深层的关系实例，因而需要在短语和句子级别上进行更深入的考察。为此，定义如下识别语法：

定义5 G_s（实例识别语法），$G_s = (V_N, V_T, S, P)$，其中 V_N 为非终结符，V_N = {S, *Time _P*, *Quan _P*, *SE _P*, *Topo _P*, *Dire _P*, *Meas _P*}，S 代表实例符号，$S \in V_N$，*Time _P* 表示时间短语，*Quan _P* 表示数量短语，*SE _P* 表示空间实体短语，*Topo _P* 为拓扑关系短语，*Dire _P* 为方向关系短语，*Meas _P* 度量关系短语；V_T 为终结符，V_T = *GORTL*；P 为产生式规则集合，形式如：$\alpha \rightarrow \beta$，$\alpha \in V_N$，$\beta \in Reg$，例如：

(1) *SE _P*→(*A* u?)* *SE*+(c *SE*)*

(2) *Time _P*→p? t+

(3) *Quan _P*→m q

(4) *Topo _P*→(p? *SE _P*? u? R_{01}+(c R_{01})*)|(p *SE _P*)

(5) *Dire _P*→ p? *SE _P*? u? R_{02}+(c R_{02})*

(6) *S*→*SE _P* V *Topo _P*|*Dire _P*

(7) *S*→*Topo _P* | *Dire _P* V *SE _P*

(8) *S*→*SE _P* u? *A* v *Quan _P*

...

式中小写字母 u、p、m、q、c、t 为词性，分别代表助词、介词、数词、量词、连词、时间词。β 中的“*”表示前面的表达式可出现零次或多次，“+”表示出现一次或多次，“|”表示逻辑或，“?”表示出现零次或一次。

2 实现方法

2.1 结构框图

本体实例搜索是问答式移动空间信息服务系统的子模块，主要搜集Web中与空间实体相关联的实例文本。它同MGIS、GIR等模块一道共同完成为移动用户提供空间信息服务的任务。图1是系统的结构框图，其中实线部分为本文涉及的本体实例搜索内容，虚线部分属于其它模块的内容。工作流程可简单描述为：用户的查询请求经无线网络进入检索器，检索器对查询请求进行语义分析，如果能够得到正确的查询语义，则分别按照SW推理器和移动地图服务所要求的格式构造查询语句并发送请求；否则，构造关键词逻辑操作向GIR发送请求。所得结果根据用户的配置情况进行排序、取舍、合并和格式转换，最后返回给用户。限于篇幅，本文只重点讨论实例搜索部分，它主要包括网页搜集与分析、实例提取、语义查询三部分内容。以下分别进行讨论。

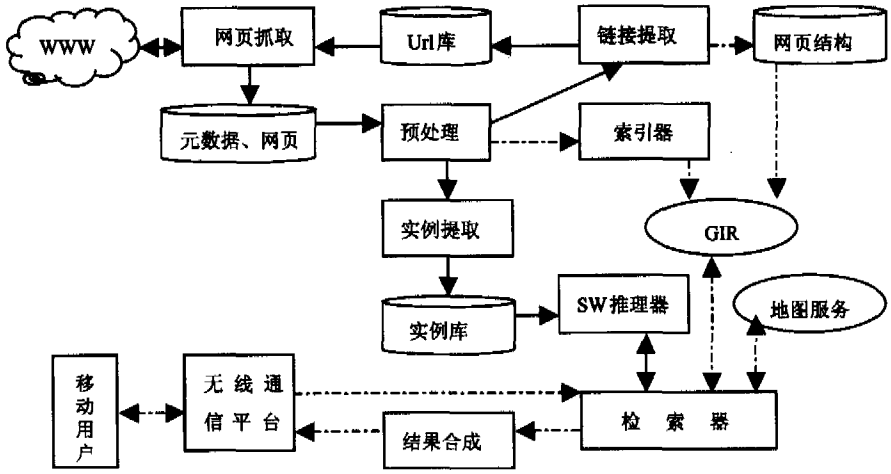

图1 移动空间信息服务系统中的本体实例搜索结构框图

2.2 网页搜集与分析

网页搜集过程与通用搜索引擎中的搜索器功能相同,目的是要高效获取 WEB 中的网页。它首先从 URL 库中获得未搜索过的 URL(或起始种子 URL),从 DNS 缓存中查找解析地址,如不存在则进行 DNS 解析,解析结果放入 DNS 缓存。然后建立连接,发送 HTTP 请求并接受数据。所得网页数据及其元数据(如版本信息、URL、IP、下载时间、文件大小、类型、网页头信息等)按照一定的格式被顺序存储在原始网页库中。

网页分析主要包括预处理、链接提取、正文提取三方面内容。预处理就是消除网页内的噪声内容(如 comments, scripts, css, 版本信息,广告等),尽量减少与正文不相关的信息给内容分析造成干扰。链接提取是根据 HTML 的规定,从网页体中提取出 URL 和相应的链接描述信息,从而扩充 URL 库,使搜集进程持续进行,并形成网页结构库。入库时需要对 URL 的类型和规范性作判断,因为一些对实例搜索意义不大的链接(如图片链接、网页格式链接等)如果放入 URL 库,会增加搜集进程的运行负担和存储空间,而且网页体中还有许多链接采用省略写法,如果不重新构造则不能形成有效 URL,这些不规范的 URL 同样会造成系统资源的浪费。另外,为了减少重复搜索问题,URL 入库时还要区分是否已抓取过,已抓取过和未抓取的 URL 分别存储于不同的列表中。在未抓取过的 URL 列表中,用因子 δ 表示每条 URL 的网页重要程度,其数值由 URL 对应的锚文本(anchor text)中空间命名实体频率及 URL 的目录深度来衡量,计算方法见(1)式。δ 值大者优先进搜索队列。

$$\delta = \frac{k \times N_{se} + N}{N} + \frac{m - D(url)}{m} \quad (1)$$

其中,N_{se}为锚文本中空间命名实体的个数;N 为锚文本中分词的总数;$D(url)$为 url 目录深度;m 为最大目录深度初值常数;k 为重要度调节系数。这是一种启发式宽度优先搜索模式。

正文提取主要是获取网页体的核心内容,它为地理本体的实例提取提供语句单元,如果单元中存在空间实体,则进行实例提取(见 2.3 节);否则进行切分,提取关键词及其权重信息、位置信息以及统计信息等,为 GIR 中的索引器提供基本数据单元,详细过程本文不予赘述。

2.3 本体实例提取

从网页正文语句中提取地理本体实例的过程分角色标注、短语识别、模式匹配三阶段进行。图 2 是流程框图,提取算法如下:

(1) 考察地理本体 *GeoOnto*,定义 *GORTL* 并构建 SSD。分析本体中实例结构,构造用自然语言表达相应实例的各种模式。

(2) 利用 *SSD*,采用正向最大匹配法识别文本串中的空间语义角色,并进行标注。不含标注的语句被过滤掉。

(3) 对语句中未被标注片段进行分词处理,并进行词性标注。

(4) 利用 G_s 构造空间本体实例的识别文法,用 Flex/Bison 分析器对该语句的标注语言结果进行模式匹配。

(5) 识别结果添加到相应的 RDF 文件中存储。

空间语义词典是与地理本体有关的关键词的集合,每个关键词被赋予一种空间语义角色类型。其

构建过程主要参考GIS各主题层对应的属性库。例如标注层中的所有地名对应空间实体角色 SE_h；图形库、属性库部分字段的中文名和别名(如长度、面积、人口等)在语义词典中对应属性角色 A_h；汉语表达8种拓扑关系(相等、相离、相交、相接、穿越、内部、包含、重叠)、八方向关系(东、南、西、北、东南、东北、西南、西北)、相对方向关系以及度量关系所用词汇对应空间关系角色 R_h。为了提高召回率，语义词典中各关键词对应的同义词、近义词、缩写等也被赋予相同的语义角色。

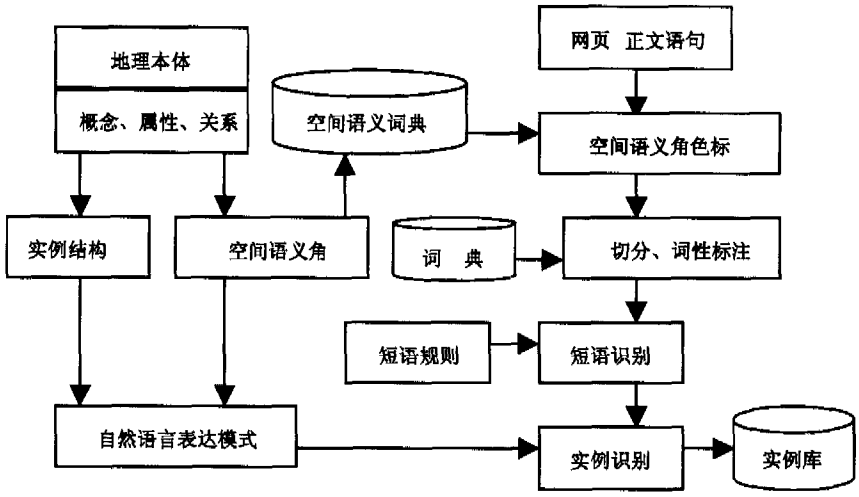

图2 空间本体实例提取流程

角色标注采用先进行语义角色标注，后切分词和词性标注的模式。具体做法是先利用空间语义词典采用正向最大匹配法进行空间语义角色的识别，含语义角色的语句被存入缓存并标注相应的语义角色，否则被丢弃；然后依次取出缓存中的语句，对其中未标注部分采用基于语料库的隐马尔可夫模型(HMM)进行分词和词性标注。这种处理模式针对性强，因为本文的实例提取仅涉及空间领域，多数网页正文语句与此领域无关，在进行空间语义角色识别时即被过滤掉，从而减少了后续处理的数据量，同时还能克服因分词结果不正确而丢弃的有效语句的问题。

正文语句的标注结果是一系列包含空间语义角色的 *GORTL* 串。实例提取进程以这些 *GORTL* 串为处理对象，通过模式匹配识别 G_s 中定义的短语和二元关系实例。做法是利用 G_s 构造巴科斯-诺尔范式(BNF)识别文法，构造过程作者在其它文献有相关描述。然后利用Flex/Bison分析器执行该文法便可得到实例识别结果。选用Bison分析器主要出于性能上的考虑。因为 G_s 中的模式有很多种，如果采用常规的匹配方法(如KMP，B-M)，则需为每个模式n构造一个有限状态自动机，然后逐个地匹配原文m，最好的线性复杂度为O(m + n)；而Bison是一种LALR(1)句法分析器，只要能构造出无冲突的分析程序，则无论有多少模式，其线性复杂度都是O(m)。

2.4 实例查询

分析器提取的实例结果以三元组的形式存放在临时文件中，将这些实例归并，以一个大的RDF文件的形式进行存储，供查询使用。在查询过程中，检索器对用户的查询请求进行语义分析，以三元组的形式构建RDF查询语句，然后向查询器发出请求。查询器获取请求后从RDF库中找出满足条件的实例结果返回给用户。

图3是一个具体的实例查询示意图。在实例库中存储了所有地理本体实例的搜索结果，用户的查询语义“[北京饭店 isGrade ?]”以RDF查询语句的形式向查询器发出请求，查询器从实例“[北京饭店 isGrade 五星级]”获取所需值“五星级”。

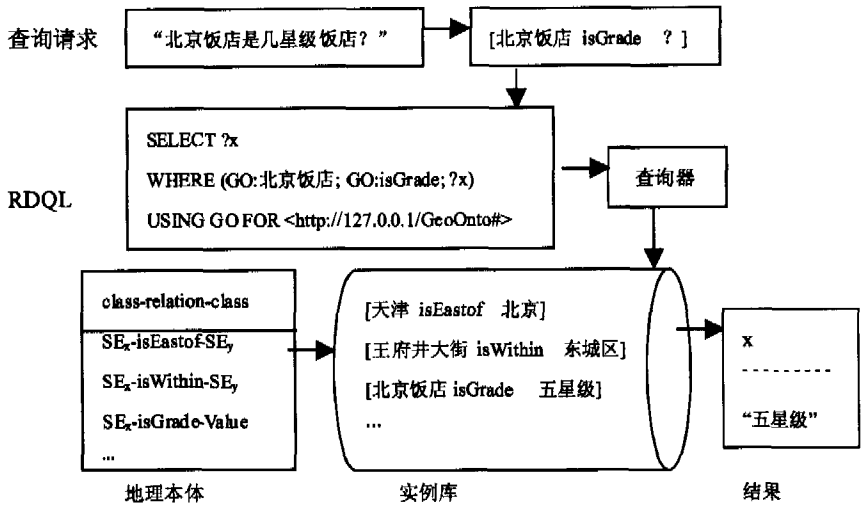

图 3 本体实例查询过程(注:示例中实例省略了命名空间)

3 实验结果与讨论

为了评估地理本体实例的搜索效果,本文选用了有效网页率(E)、实例召回率(R)、准确率(P)作为评价指标。有效网页率是指搜集到网页集中含地理本体实例的网页数占总数的百分比,用于评估网页搜集质量;召回率和准确率是信息提取中常用的评价指标,前者是正确提取的结果数与所有结果数量的百分比,后者是正确提取的结果数与所有正确结果数量的百分比。由于 Web 中所有地理本体实例数难于得到,召回率的计算采用样本估算的方式进行,即抽取一定量的网页样本,先以人工的方式将其中的实例提取出来,计算出现频率,然后与系统提取的结果相比较得出百分比。系统的测试环境为 PC (2.4GHz CPU, 520M RAM)。

E 值评估方法

在网页抓取模块中,启动相同数量的线程,以同一种子 url 为入口进行搜索。在计算 url 入搜索队列的优先级时,采用两种方式进行:

①用 2.2 节中的(1)式计算 δ 值;

②仅用(1)式中的($m - D(url)$)计算 δ 值,这种方式类似宽度优先策略,是通用搜索器常用的策略之一。图 4 为实验结果(参数取值:$m = 10$, $k = 5$)。

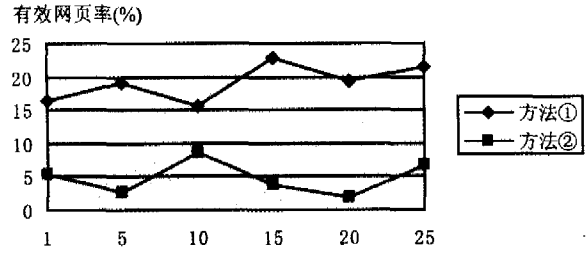

图 4 本体实例搜索有效网页率

R 值和 P 值评估方法

从有效网页集中随机抽取 30 篇网页,先将这些样本进行经预处理,去除网页中的标记符后取出正文部分。然后先以人工的方式将其中的空间实体(SE)和二元关系地理本体实例(GOS)提取出来,分别统计文中出现的次数。然后再将这些样本以文件的形式输入到系统中,记录系统识别 SE 和 GOS 的统计结果。然后与人工提取的结果相比较即可得出 R 和 P 的计算结果以及综合评价指数 F 的大小(见表 1),F 的计算方法参考文献[8],其中相对权重参数取值为 1。

表 1 本体实例的召回率与准确率

	召回率(%)	准确率(%)	F-Score
SE	92.3	86.1	89.1
GOS	53.6	81.9	64.8

从以上实验结果可以看出,采用启发式宽度优先搜索策略所得到的有效网页率要比直接用宽度搜索效果好。这主要是因为在搜索过程中对 URL 进行了重要程度评估,使得与地理本体相关的 URL 被优先搜索。在本体实例的提取过程中,空间实体的识别效果好于实例的识别,二元关系地理本体实例的提取与通常的信息提取系统性能(二元关系准确率和召回率约 60% ~ 70 %)相比,准确率有所提高,这主要是因为采取了空间语义角色标注、短语识别、模式匹配三层过滤机制所致;但召回率还有待提高,该值受人工搜集的实例表达模式完备性影响,需要在实验中不断地进行修正和补充。

4 结 论

利用地理本体,通过空间语义角色标注、短语识别以及模式匹配的手段,可以从 Web 中获取大量与空间位置相关的本体实例。这些实例作为问答式空间信息服务系统的可动态更新数据源,能为移动用户提供准确、简洁的问答结果。目前系统的实例准确率基本达到应用需求,但有效网页率和召回率还有待进一步提高。该系统是语义 Web 技术在空间领域应用的一种探索,可以作为问答式移动空间信息服务系统的重要补充。

参考文献

[1] Greenwood M A. Using pertainyms to improve passage retrieval for questions requesting information about a location. http://nlp.shef.ac.uk/ir4qa04/Greenwood-IR4QA.pdf, 2004

[2] Wahlster W. Multimodal interfaces to mobile webservices. http://www.dfki.de/~wahlster/ICT-Kennriscongress_2002/Multimodal_Interfaces_to_Mobile_Webservices.ppt, 2002

[3] Christian K. Situated interaction on spatial topics. http://www.comp.lancs.ac.uk/~kray/pub/2003_sisto.pdf,2003

[4] Kim S. Question answering towards automatic augmentations of ontology instances. http://eprints.ecs.soton.ac.uk/8911/01/sangheekimesws2004-prepress.doc,2004

[5] Morimoto Y. Extracting spatial knowledge from the web. http://www.mccurley.org/papers/SAINT03.pdf, 2003

[6] Bozsak E, Ehrig M, Handschuh S, et al. KAON-towards a large scale semantic web. http://www.aifb.uni-karlsruhe.de/WBS/dob/pubs/ecweb2002.pdf, 2002

[7] Hiramatsu K, Reitsma F. GeoReferencing the semantic web: ontology based markup of geographically referenced information. http://www.mindswap.org/2004/geo/geoStuff_files/HiramatsuReitsma04 GeoRef.PDF,2004

[8] 俞士汶主编.计算语言学概论.北京:商务印书馆,2003.322-323

Geo-ontology instances search for QA-based mobile spatial information service

Le Xiaoqiu, Yang Chongjun*, Liu Donglin*, Yu Wenyang*

(Library of Chinese Academy of Sciences, Beijing 100080)

(*The State Key Laboratory of Remote Sensing Information Sciences, Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing 100101)

Abstract

Based on the technology of common Search Engine, this paper defined several spatial semantic roles and instance expression patterns according to geographic ontology, and searched large quantities of geo-ontology instances by means of spatial semantic annotation, semantic phrases recognition and pattern match. It also provided QA-based instance service for mobile user with the help of Semantic Web technology. The primarcy experiment shows a good pricision and an ordinany recall.

Key words: mobile spatial information service, search engine, ontology instance