基于本体的空间搜索引擎研究

段 磊 李 琦 毛 曦
(北京大学数字地球工作室 北京 100871)

摘 要 提出了一种智能空间搜索引擎的解决方案。通过分析传统搜索引擎在处理空间语义方面的缺陷,将本体和自然语言处理技术引入搜索引擎中,解决基于自然语言查询的空间检索问题。初步构造了基于本体的空间搜索引擎的结构框架,分析了本体在空间搜索引擎中的应用范畴,并构建了相应的本体库以及解析自然语言查询的模式库,提出了自然语言式空间查询的解析方案。最后通过建立空间搜索引擎原型系统证明了该方案的可行性。

关键词 本体,智能,空间搜索引擎,自然语言

Research on Geographic Search Engine Based on Ontology

DUAN Lei LI Qi MAO Xi
(CyberGIS Studio, Peking University, Beijing 100871, China)

Abstract This paper proposed a solution for intelligent geographic search engine. By analyzing the disadvantages of general search engine on dealing with geographic semantic, the technology of ontology and natural language were used to solve the problem of geographic retrieval. The framework of intelligent geographic search engine was established, Ontology library and pattern library for analysing natural language were built, and geographic retrieval solution based on natural language was given. At last, one intelligent geographic search engine prototype system was designed and implemented.

Keywords Ontology, Intelligence, Geographic search engine, Natural language

1 引言

一方面,人类的生产和生活离不开地理空间,因而地理信息与人类活动密切联系在一起,如何有效地获取和利用地理信息就成为人类研究的焦点。地理信息系统在研究和实践方面都取得了长足的进步,但它仍然以处理地图数据为主,对如何处理和利用与地理位置相关的文本信息(包括网页信息)却很少考虑。

另一方面,网页中包含有丰富的空间信息,传统搜索引擎是通过关键词来检索网页,没有考虑到自然语言的理解,也不能很好地处理空间语义,以获取所需要的地理信息。例如用户希望查询"北大附近的肯德基店",那么搜索引擎会把包含所有关键词的网页查找出来,但是根本不了解关键词的语义。结果网页可能既不是关于肯德基店的,也不是北京大学附近的,甚至都可能与北京大学无关。而类似的请求随着移动通信和全球定位技术的发展变得越来越普遍[3]。同时,在结果排序方面,最先返回的结果可能没有考虑用户所在的地理位置,因此与用户的期望相差很大。例如北京的用户查询"肯德基店",网页可能返回很多关于上海、广州等城市的肯德基店信息。因此本文提出了基于本体的空间信息搜索引擎(Onto-GSE),强调在传统搜索引擎基础上增加地理偏好的理解[1]。

2 系统框架

基于本体的空间信息搜索引擎的结构框架如图1所示。自下而上,包括3个层次,即数据层、中间层和前台处理层。

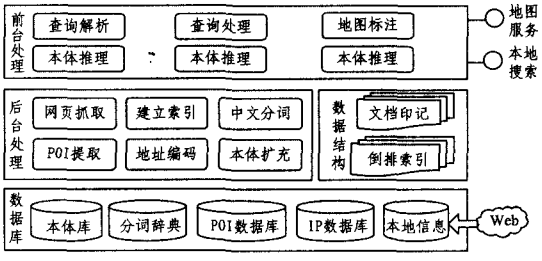

图1 Onto-GSE系统框架

最底层是数据层,表示搜索引擎中需要使用到的数据源,该数据源主要包括3个方面:一个方面是从Web中获取数据。通过垂直搜索,获取网页数据,进而通过半结构化网页信息提取的方式获取本地信息及相关的评价信息;另一方面是利用已有的地理信息数据库,包括POI(感兴趣点)、电子地图、IP数据库以及分词词典;第三方面是利用领域专家知识建立的本体库,包括地名本体、实体本体、位置本体和量词本体等。

中间层是后台处理模块,不直接和查询过程关联,主要是为完成空间查询必须做的准备工作:抓取网页数据、利用包装器提取本地信息、对提取的本地信息进行地理地址编码等。

前台处理模块是查询模块,主要负责界面交互功能,完成对用户查询的自然语言解析、查询的本体扩展及查询结果的显示。

到稿日期:2008-04-01 本文受国家973项目(编号:2006CB701306):"网格环境下空间信息智能服务及应用示范"项目资助。
段 磊 硕士研究生,E-mail:duanlei.pku@gmail.com;李 琦 教授,博士生导师;毛 曦 博士生。

Onto-GSE 的流程图如图 2 所示。

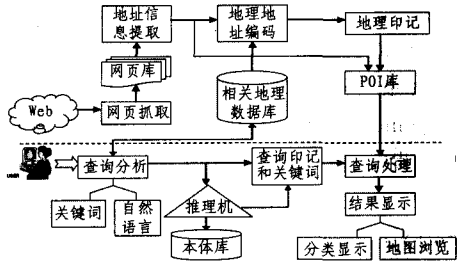

图 2　Onto-GSE 流程示意图

后台处理：

i)利用垂直搜索技术，对黄页网站和其他的半结构网站进行爬取，获取网页数据；

ii)提取网页的包装器，将网页中包含的实体名称、属性和地址等提取出来；

iii)利用地理编码服务器对地址进行地理编码，将地址映射成地理坐标并入库；

前台查询：

iv)用户提交查询请求后，获取用户 IP，利用 IP 数据库获取用户地理位置；然后对该位置进行地理编码，获取用户所在位置的坐标数据；

v)利用本体库和本题推理机对自然语言的查询进行解析，获取查询印记和关键词；

vi)根据空间叠置分析，获取查询结果；

vii)计算查询参考点和查询目标之间的欧几里德距离，根据远近排序显示及根据区域分类显示；

viii)通过地图标注，得到查询范围所在的电子地图，在浏览器中显示，并将查询结果显示在地图上。

3　关键技术实现

3.1　空间信息本体的构建

网络异构信息源中的地理信息除了存在数据结构上的异构，还存在语义上的异构[7]。语义异构一般分为以下两种：(1)异形同义词，即不同的词汇表达同一个含义，表现在地名上为“同地异名”现象，如“北京大学”也称为“北大”、“燕园”等。(2)同形异义词，即同一个词汇表达不同的含义。表现在地名上为“同名异地”现象。如“海淀区”可能是北京市的海淀区，也可能是其他城市的海淀区。

针对上述网络异构信息源中语义异构问题，Onto-GSE 通过 OWL 构建了一系列描述地理实体中具有普遍性的特征和语义关系的本体。

a)地名类本体(PC)：顾名思义，地名类本体中的所有类和实例都是地名，如山东省、北京市、海淀区、北京大学等，都属于地名类本体。地名之间存在等级层次关系，如海淀区是北京市的子类，北京大学是海淀区的子类，北京大学、北大、燕园属于等价类。

b)实体类本体(OC)：实体类本体包含了各种命名实体，比如餐厅、宾馆、医院、学校等实体类。个实体类下面有其子类和类的实例：如餐厅下面有子类肯德基、麦当劳、九头鸟等。肯德基和 KFC 属于等价类。

c)位置类本体(CC)：位置类本体是用来描述空间关系的本体，如 IsNear 类、Overlay 类、Isleft 类等。IsNear 类包含：附近、周边、周围、旁边等实例。Overlay 类描述的是包含关系，Isleft 类描述地名的左边。

d)量词类本体(NC)：量词类本体是用来描述各种量词和量词之间的转换关系的本体，比如公里、英里、千米、米等量词类。

在 Onto-GSE 中，应用上述建立的本体，通过 Jena 对地理本体实例进行语义推理，解决了“同地异名”和“同地异名”问题。

3.2　模式库的建立

当用户提出自然语言式查询时，Onto-GSE 需要对查询语言进行语义和词性标注，而标注后的结果需要通过模式匹配转化成计算机可理解的空间查询。为此，Onto-GSE 构建了一系列进行自然语言解析的模式。下面介绍其中的 4 种模式。

a)实体类

例如：查询“肯德基”

Query(OC) →Query(CC,OC) →Overlay(Point,Area)=1

b)地名类＋实体类

例如：查询“北京的肯德基”

Query(CC,OC)→Overlay(Point,Area)=1

c)地名类＋位置类＋实体类

例如：查询“北大附近的肯德基”

Query(PC,CC,OC) →

Overlay(Point,Buffer(PC,1000))=1

d)地名类＋位置类＋数词＋量词＋实体类

例如：查询“北大周围 2000 米范围内的肯德基”

Query(PC,CC,N,NC,OC) →

Overlay(Point,Buffer(PC,N))=1

3.3　查询处理

当用户提出查询时，Onto-GSE 一方面需要将查询转化成计算机可理解的空间查询，另一方面需要对查询语言进行本体推理。最后根据推理结果，系统将产生新的查询来代替原始的用户查询。处理流程如图 3 所示。

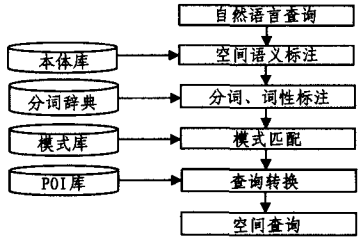

图 3　自然语言解析流程示意图

流程介绍：

i)构建本体库，构建模式库；

ii)利用本体库，采用正向最大匹配法算法识别出查询中的空间语义本体角色并进行标注，将不含标注的部分过滤掉[6]；

iii)利用分词辞典对语句中非标注片段进行分词处理，并进行词性标注；

iv)根据标注提取查询模式，与模式库中的标准模式进行模式匹配；

v)根据对应的模式进行空间查询解析；

vi)对地名类本体和实体类本体进行本体推理，完成查询

处理。

3.4 检索和排序

由于 POI 库的数量很庞大，如果在整个库中进行检索则会大大影响检索效率。因此在 Onto-GSE 的检索机制中引入了“父类映射”机制：通过本体推理，可以得知关键词的父类。检索时，可以直接定位到父类，然后再进行查询。这样便缩小了查询的范围，提高检索效率。比如搜索“肯德基”，将肯德基定位到父类“餐厅”上，直接到餐厅的 POI 库进行搜索，大大提高搜索效率。

在排序和结果显示方面，Onto-GSE 采用了分类和混合排序相结合的方式。当查询的地名本体有子类时，首先按子类进行结果分类，然后对每个子类中的结果进行混合排序。混合排序算法是综合考虑距离(d_i)和主题相关性(t_i)，然后对二者进行加权，并进行归一化处理：

$$Score_i = Sd_i + St_i$$

其中，Sd_i 是计算用户查询参考点和查询目标之间的欧几里德距离计算得出。St_i 是根据结果向量和查询向量的关联程度计算得出。

4 查询实例分析及结果演示

基于本文的思路和涉及到的关键技术，在一系列 Cyber-SIG Studio 服务器的支撑下，设计并实现了 Onto-GSE 原型系统。以下分别以查询“肯德基”和查询“北大附近的 KFC”为例，说明自然语言查询的解析过程、模式匹配过程和空间检索过程。

用户查询：肯德基

IP 地址获取并解析：北京市海淀区

标注结果：肯德基/OC

模式匹配：Query(OC)

查询转换：Query(北京，肯德基＋KFC)

→Overlay(肯德基＋KFC，Area(北京))＝1[4]

结果显示：浏览器左侧按各区分类显示查询结果，因查询用户位于海淀区，故海淀区排最前面。右侧为地图显示。如图 4 所示。

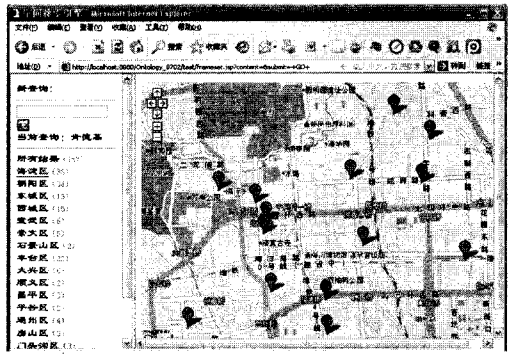

图 4 查询结果示意图 a

用户查询：北大附近的肯德基

IP 地址获取并解析：北京市海淀区

标注结果：北大/PC 附近/CC 的/U 肯德基/OC

模式匹配：Query(PC，CC，OC)

查询转换：Query(北大，附近，肯德基)→

Query(北京大学，附近，肯德基) →

Query(point(x，y)，IsNear(1000)，肯德基＋KFC) →

Overlay(肯德基＋KFC，Buffer(point(x，y)，1000))＝1

结果显示：位置本体“附近”初始值设定为 1000m 缓冲区，因此查询实质就为北京大学周围 1km 范围内的肯德基，结果如图 5 所示。

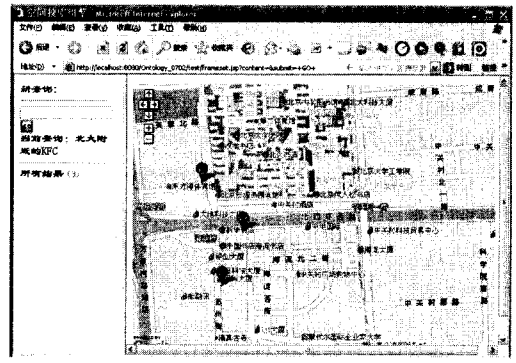

图 5 查询结果示意图 b

结束语 基于本体的空间搜索引擎同时涉及到本体、搜索引擎和空间信息处理 3 个方面的研究领域的内容，是近期学术界和企业界研究的热点问题。本文通过与传统搜索引擎进行比较，指出研究智能空间搜索引擎的重要意义。同时给出了基于本体的搜索引擎的体系架构，提出了利用本体解决“自然语言式空间查询”的方案，并通过领域专家知识建立了相应的本体库和模式库，接着建立了智能空间搜索引擎原型系统来对方案的可行性进行了验证。随着研究的进一步深入，我们将不断完善自然语言查询的解析方案，使其支持更复杂的自然语言式的空间查询。

参 考 文 献

[1] Markowetz A，Chen Y-Y，Suel T，et al. Design and implementation of a geographic search engine // 8th Int. Workshop on the Web and Databases(WebDB). June 2005

[2] Studer R，Benjamins V R，Fensel D，et al. Knowledge engineering，principles and methods[J]. Data and Knowledge Engineering，1998，25(1/2)：161-197

[3] Himmelstein M. Local search：internet is the yellow page. Long Hill Consulting，LLC Computer，IEEE Computer Society，2005

[4] Shekhar S，Chawla S. 空间数据库. 谢昆青，马修君，等译. 北京：机械工业出版社，2004

[5] http://www.w3.org/DesingIssues/Toolbox.html

[6] 乐小虬. 非结构化网络空间信息智能搜索与服务研究[D]. 博士学位论文. 中科院遥感应用研究所，2006

[7] 虞为，曹加恒，陈俊鹏. 基于本体的地理信息查询和排序[J]. 计算机工程，2007，33(21)：27-32