DOI:10.13203/j.whugis2003.03.009

第28卷 第3期 武汉大学学报·信息科学版 Vol.28 No.3
2003年6月 Geomatics and Information Science of Wuhan University June 2003

文章编号:1671-8860(2003)03-0301-05 文献标识码:A

空间信息自然语言查询接口的研究与应用*

马林兵[1] 龚健雅[1]

(1 武汉大学测绘遥感信息工程国家重点实验室,武汉市珞喻路129,430079)

摘 要:提出了空间信息自然语言查询接口(SINLQI),并讨论了基于E-R语义词典的建立、中文分词、查询文法规则及其应用领域等主要问题。

关键词:自然语言;空间信息;空间查询

中图法分类号:P208

空间信息查询是地理信息系统中不可或缺的基本功能。目前,空间信息查询主要采用两种方式:一是基于GUI应用界面,往往由GIS系统定制;二是基于扩展SQL语言(GeoSQL)。这两种方式能满足一般的查询要求,但还有些不足。基于GUI应用界面方式受限于界面本身,不能完成某些空间关系的查询;基于扩展SQL语言的查询方式较灵活,能满足大多数空间查询请求,但是并没有形成类似SQL那样被所有GIS平台支持的标准。另外,扩展SQL语言对使用者有一定的要求。

目前,GIS的应用从特定的专业领域走向普通信息服务,GIS应用系统从原来的单层模式应用变为多层模式,从孤立的桌面应用系统变为基于Internet和无线互联网的应用系统,所使用的终端设备除了传统的台式机、笔记本电脑外,还有各种各样的掌上电脑、PDA、手机等,因此,采用自然语言方式进行空间信息查询能满足更广泛、更普通用户的需要,空间信息自然语言查询接口(spatial information natural language query interface,SINLQI)已开始越来越受到重视。

1 SINLQI的应用范围和可行性

自然语言查询接口属于人工智能的自然语言理解的研究范畴。所谓自然语言理解,实质上是把一种表达转换为另一种表达的过程,这种转换也可视为映射。建立自然语言理解系统就是寻求映射的算法,使机器能够得到与人在理解上相当的输出(刘开瑛,1991)。可以说,自然语言理解是一个相当复杂的过程。

作为一种查询方式,SINLQI并不适用于所有GIS应用。在如下几种情况下,自然语言接口难以实现或作用不大。

1)GIS应用以海量空间数据管理为目的;

2)GIS应用以专题图制作或地图输出为目的;

3)GIS应用是面向专业人员的。

在上述应用中,所处理的是海量数据,不便于建立用于SINLQI的数据词典及文法规则。对于专业人员而言,他们能自如地使用各种GUI查询界面或GIS平台提供的查询语言完成查询要求,在这种情况下,即使花一定代价实现SINLQI,其使用效率也不高。

文献[3]认为,作为通用工具的GIS,不可能在自然语言基础上建立一个统一的通用查询操作模型,但是,作为某一具体的GIS应用领域,基于自然语言的查询表达是很有吸引力和发展前景的,这种查询的实现需要数据库和专门的知识库。

目前,GIS应用越来越面向公众服务,如位置信息服务(location-based service)、房地产信息查询、交通信息查询、旅游景点介绍及小区介绍等。显然,在这些应用中,采用自然语言的查询方式,更能被普通用户所接受和使用。基于手机短消息(SMS)和WAP的城市位置信息服务是目前移动GIS的开发热点,如果能让用户以自然语言的方式结合语音技术查询地理空间信息,无疑增强了

* 收稿日期:2003-03-12。

项目来源:国家863计划资助项目(2002AA131030);国家973计划资助项目(G2000077904)。

系统的易用性和可扩充性。

在中文自然语言处理技术中，数据库自然语言接口是自然语言理解和数据库技术相结合的产物。国外有关数据库自然语言查询接口的研究可以追溯到 20 世纪 60 年代，比较著名的原型系统有 BASEBALL、LUNAR、LIFER 以及 Microsoft 商品软件 Englih Query。国内在这方面也取得了一些研究成果，如东南大学的 CQI 系统、中国人民大学和香港中文大学的 Chiql 系统等。

空间信息查询与数据库查询有一个共同特点，就是从已有的、特定的数据源中查找满足某种约束条件的数据子集。它们在查询语义操作上能采用几乎类似的形式语言来描述，这一点已在对 GeoSQL 语言研究中得到证实。因此，在自然语言接口方面，二者也是相通的。

在空间数据库中，保存了空间数据和相关的属性数据。属性数据与普通数据库中的数据几乎是一样的，而空间数据是对真实地理要素的抽象，例如，一条道路被抽象成一条线，一个行政区被抽象成一个面等。但是，在许多情况下，人们在对空间信息提出查询请求时，都不自觉地有一个语义“还原”过程。例如，人们会提出“107 国道穿越哪几个县”的查询请求，而不是“ID 号为 1100 的线会与哪几个面相交”。这种空间查询的自然语义使得可以采用与数据库自然语言查询接口类似的方法研究 SINLQI。

2 基于 E-R 模型的语义词典

自然语言理解中，词典是必不可少的，它是中文分词、语法分析、语义理解的基础。作为 SINLQI 中的词典，它与其他的中文自然语言处理系统的词典是不同的，按照词性（如名词、动词等）划分的词典在 SINLQI 中是没有意义的。SINLQI 作为受限语言理解系统，其词典的设计应该与语用环境紧密联系。因此，本文提出基于 E-R 模型的语义词典。

E-R 模型即实体-关系模型，是目前数据库研究中最成熟的语义模型。E-R 模型是描述整个事物的概念模式，能够比较充分地反映现实世界的客体及其联系。

在空间数据库中，实体（E）表示真实地理要素（F），地理要素由一个空间几何属性和若干非几何属性所组成：

$$F = (S;\ A_0, A_1, A_2, \cdots\cdots, A_n)$$

S 表示空间几何属性，A_i 表示非几何属性。

每一个地理要素不是孤立存在的，它必然同其他地理要素存在某种关系。实体之间的关系（R）是通过空间关系来表现的，即拓扑关系（相交、包含、叠加、相离、重合、穿越）、方向关系（东、南、西、北）、度量关系（距离、面积、周长）。图 1 是一个道路和行政区在空间穿越关系下的 E-R 图。

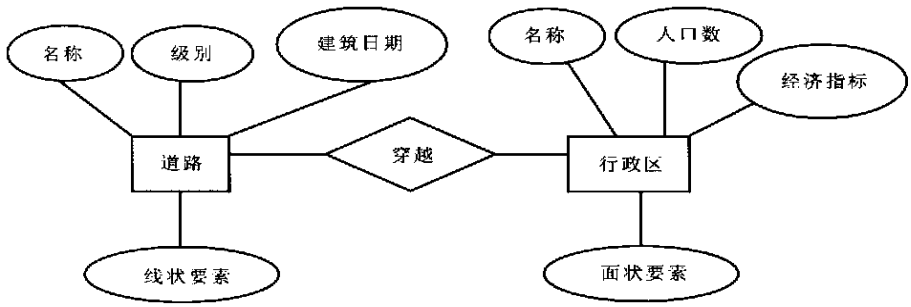

图 1 道路和行政区 E-R 图

Fig. 1 E-R for Road and District

因此，在 SINLQI 词典的词汇中，设计了以下语义信息。

1）TYPE，即词汇的类别，分为 ES（表示实体的集合）、EI（表示实体的个体）、EA（表示实体的属性）、SR（表示空间关系）及 O（其他）。O 类词是一些通用的与空间数据库查询关联的词，如大于、小于、等于以及介词、连词、数量词等。O 类词一般与具体 E-R 语义无关，可以作为通用词典单独处理。

2）SELECT，即选择分类语义，实体词在空间数据库中的分类选择形式。例如，在某个空间数据库中，有一个 School 图层，对于词条“学校”，有选择语义：

SELECT（layer（“School”））

如果 School 层有 SchoolType 属性，则对于词条“大学”有：

SELECT （ layer （“ School ”， field

("SchoolType")=3))

3)KEY, 即关键字, 它是 EI 类词的特有语义项, 如词条"武汉大学"对应 School 图层中的 geoID=1200 的空间对象。

4)CONNECT, 即属性关联, 它是 EA 类词的特有语义项。如词条"电话号码", 关联的是 School 图层中的 PhoneNumber 属性。

5)S-RELATION, 即空间关系, 它是 SR 类词特有的语义项。如词条"相交", 表示空间关系 Intersect, "南面"表示空间关系 South 等。

中文与西文的最大区别就是, 前者的语句是连续的, 词间无间隙, 因此, 中文信息处理的第一步就是分词。分词的形式有多种, 如基于词典的分词、基于语料库的分词、二者相结合的分词。在受限中文信息处理系统中, 一般采用基于词典的分词。分词的方向有正向和逆向, 分词的方法有最大匹配法和最小匹配法。最小匹配法已被证明不适于作为数据库查询语句的分词方法, 见文献[7], 因此我们采用正向最大匹配法。

最大匹配法的一般算法是, 取词条的最大长度, 依此长度截取待切分语句的一个子串, 与字典中的词条进行比较。若不相等, 则子串长度减 2, 继续比较, 直到找到相等的词条, 将它切分出来。再循环进行, 直到切出所有的词条。

在空间信息查询的词典中, 必然包含大量的地名、机构名(地名、机构名往往作为地理实体的标识, 因此将地名、机构名作为词条处理), 这些词条的长度一般在 5 个汉字以上, 因此, 我们设计了如下的词典数据结构。

1)以词条首字的区位码作为哈希值建立哈希表。哈希函数如下:

H(key) = (HiByte−176) * 94 + (LoByte−161)

2)采用链地址法处理地址冲突。在同一链中, 词条按汉字内码由小到大排序, 为了保证最大切分, 若某一词条比另一词条短, 且是它的子串, 则长词条排在前面。

3)排序词条之间有一定的相关性。我们在建立词典时, 为每个词条增加一个域, 记录该词条与前一词条的相关系数。如"武汉工业学院"相对于上一词条"武汉工业大学"的相关系数为 8, 即前 4 个汉字是相同的。该系数能有效地减少在词典查找中字符的比较次数。

3　文法规则及处理

3.1　短语结构文法

由于中文语言的复杂性, 对中文语义的理解是很困难的。目前, 大多数自然语言理解系统都建立在受限文法的基础上。文法分析的作用有两个: 一是确定输入语句的文法结构, 二是使文法结构规整化。基于短语结构的文法, 是目前自然语言处理中广泛采用的方法。

一个短语结构可以由四部分组成。

V_T: 终结符集合, 这是一些已被定义的词组组成的表, 属于不能再往下推导的符号集;

V_N: 非终结符集合, 这是一些用以说明文法的符号表, 可往下推导, $V = V_T \cup V_N$;

P: 产生式集合, 每个产生式可以描写成 $a \rightarrow b$;

S: 开始符, V_N中的一个符号。

上述 a 是 V 中一个或多个符号序列, 即 $a \in V^+$ (V 的正闭包); b 是 V 中零个或多个符号序列, 即 $b \in V^*$ (V 自反正闭包)。

短语结构文法可以用来描述任何可以递归枚举的语言, 它是一种非常强大的形式体系。上下文无关文法是短语结构文法的受限形式, 它很适合于用 BNF 来表示。我们采用 BNF 来表示查询文法规则。

3.2　查询文法规则定义

对中文查询语句制定了如下的限制规则。

1)查询语句必须是正向的, 即查询条件在前, 要查询的目标在后;

2)一个查询语句中查询目标限于一种类型;

3)不允许指示代词, 如"它"、"其"等;

4)暂时不考虑聚集查询。

在进一步研究后, 可以考虑放松上面的限制条件。事实上, 根据人们最直接的查询习惯, 大多数查询请求都能满足这些限制条件。下面仅列出了查询文法的概要部分:

〈SENTENCE〉::=〈QUERY〉

〈QUERY〉::=〈CP〉〈OP〉

〈OP〉::=〈 *ES〉|〈 *EA〉|〈M〉

〈M〉::=〈AREA〉|〈LENGTH〉|〈DISTANCE〉

CP 表示查询条件短语, OP 表示查询目标短语, *表示该类词的终结符。查询的目标分为空间实体(ES)、属性(EA)、量度(M)。

查询条件又分为空间约束条件短语〈SCP〉和属性约束条件短语〈ACP〉。

〈CP〉::=〈SCP〉|〈ACP〉|〈SCP〉〈ACP〉|〈ACP〉〈SCP〉

〈SCP〉::=〈 *SR〉〈 *EI〉|〈 *P〉〈 *EI〉〈 *SR〉|〈 *EI〉〈 *SR〉

P 代表介词，如“在”、“与”等。事实上，根据 EI 的空间类型（点、线、面）的不同，SR 与 EI 之间有更深的语义约束关系，在此不做讨论。

〈ACP〉::=〈 *EA〉〈 *RL〉〈EAV〉

RL 表示关系运算符，EAV 表示属性值。EAV 根据 EA 的类型由相应的终结符组成。

3.3 查询文法分析

作为面向最终用户的查询接口，对用户的输入不能有太严格的规则限制，即允许用户输入一些不太规范的句子，如缺少助词、连词等。例如，用户在输入复合查询条件时，可能没有输入“并且”、“或者”等关系连词，这种情况可统一作“与”关系处理。在 SINLQI 中，文法分析的主要目的是找到查询条件和查询目标，并在分析过程中，结合词典中的 E-R 语义，生成有利于以后语义分析的数据结构。下面是基于 LA（loose analysis）原则的算法思路。

1）汉字切分，将切分后的词放入队列中；

2）从队列中找 EA、EI、ES 类型的终结符，并作标志；

3）根据文法规则限制 1）、2），从队尾取已作标志的终结符，若查询目标有多个，且类型一致，可依此判断取出多个终结符；若类型不一致，则认为输入有错；

4）将取出的 OP 短语从队列中删除，单独存放，并从词典中找到其对应的 E-R 语义；

5）根据文法中 SPC 与 ACP 的短语规则，利用上下文无关的文法推导树，生成“实体·空间关系”、“属性·关系符·值”的形式，并找到相应的关系连词，若关系连词省略，一律作“与”关系处理；

6）根据 E-R 词典确定 CP 中终结符的语义。

文法分析后就是语义分析。在数据库自然语言查询接口中，语义分析过程实际上就是如何生成 SQL 语句的过程。一旦 SQL 语句生成，就可以很容易得到查询结果。但是，由于在地理信息系统平台间没有形成统一的空间查询语言（GeoSQL），各个平台之间的空间信息查询的实现方式是不同的。因此，SINLQI 语义分析模块的主要任务就是将查询条件转换成最基本的“原子”查询形式，再结合不同的 GIS 平台，得到查询结果。图 2 是 SINLQI 的总体框架。

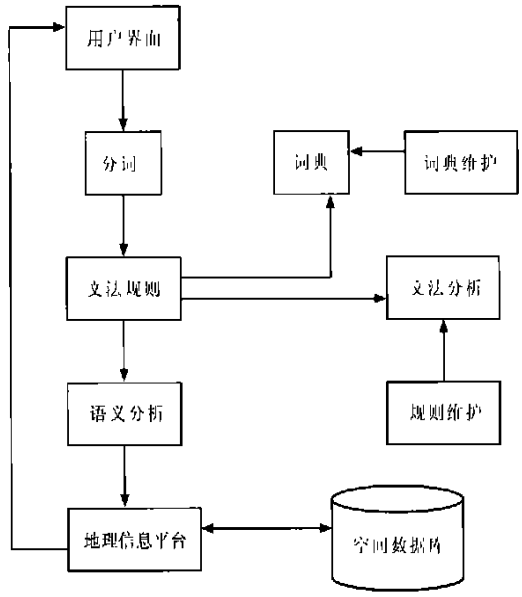

图 2 SINLQI 总体框架

Fig. 2 Main Frame of SINLQI

4 SINLQI 在 Mobile GIS 中的应用

数字城市是目前空间信息应用发展的重要方向，其核心就是利用集成的空间信息为行业和大众服务。Mobile GIS 是数字城市的一个重要服务方式，其直接的表现形式就是利用各种移动终端设备如 PDA、手机、车载系统，通过无线访问空间数据库服务器，得到地理信息服务。通讯方式主要有 SMS 方式、WAP 方式以及移动拨号上网方式（HTTP）。用户使用移动设备主要用于空间信息的查询，但是，移动设备普遍屏幕较小，CPU 功能弱，不可能运行比较复杂的查询界面程序，操作起来不方便，并且不同的终端，其屏幕大小和操作方式不相同。采用 SINLQI，只需要向用户提供简单的输入界面，用户以自然语言的方式同服务器交互，查询条件不受限于已有的界面。

WAP 服务的普及性和方便性为 SINLQI 提供了一个很好的应用平台。笔者开发了一个简单的基于 WAP 方式的城市位置信息服务的 SINLQI 原型系统。系统能提供多类兴趣点（学校、医院、商场、车站等）和城市道路的空间位置信息查询。在某些应用领域里，基于 SINLQI 的应用服务系统，具有很好的易用性和可扩展性。该原型系统的一个运行示意图见图 3。

5 结 语

SINLQI 是地理信息系统走向大众化服务的应用方向，是人工智能同地理信息系统相结合的产物。在当前 GIS 发展形势下，本文提供的基于自然语言的查询接口，能极大地扩展 GIS 的应用

图 3　SINIQL 在 WAP 下运行示意图

Fig. 3　SINLQL' s Running Demo Based WAP

范围和应用方式, 同时, 也可能为 GIS 的互操作提供一种新的手段。当然, 一个好的、实用的 SINLQI 系统, 需要经过多次的反馈实验, 不断丰富文法规则和词库的语义才能成熟。

参　考　文　献

1　刘开瑛, 郭丙炎. 自然语言理解. 北京: 科技出版社, 1991

2　姚天顺, 朱清波, 张 [illegible] 等. 自然语言理解. 北京: 清华大学出版社, 2002

3　李　霖. 地理信息系统空间目标查询模型的研究: [博士论文]. 武汉: 武汉测绘科技大学, 1997

4　黄　波, 林　珲. GeoSQL: 一种可视化空间扩展 SQL 查询语言. 武汉测绘科技大学学报, 1999, 24(3): 199～203

5　孟小峰, 王　珊. 数据库自然语言查询系统 Nchiq1 中语义依存树向 SQL 的转换. 中文信息学报, 2001, 15(5): 40～45

6　许龙飞, 杨晓昀, 唐世渭. 基于受限汉语的数据库自然语言接口技术研究. 软件学报 2002, 3(4): 537～543

7　徐九韵, 仝兆岐, 向逐聪, 等. 数据库汉语查询语言的分词研究与实现. 中文信息学报, 1998, 12(4): 53～59

第一作者简介: 马林兵, 博士生。主要研究方向是空间数据库、分布式地理信息系统、人工智能。

E-mail: mlb1999@yeah. net

Application of Spatial Information Natural Language Query Interface

MA Linbing[1]　　*GONG Jianya*[1]

(1　State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan, China, 430079)

Abstract: The paper puts forward spatial information natural language query interface (SINLQI), which can make a geographical information system easier to use. The authors discuss the establishment of dictionary based E-R semantic, Chinese word segmentation, query grammar rule, application field etc. The authors design a set of common interfaces for dictiorary organizing, word processing, grammar and semantic analysis. Using these interface, The authors implement a SINLQI demo based wireless application protocol (WAP). It establishes a base for applying SINLQI to GIS project.

Key words: natural language; spatial information; spatial query

About the first author: MA Linbing, Ph. D candidate. His main research interests are spatial database, distributed geographical Information system and artificial intelligence.

E-mail: mlb1999@yeah. net

(责任编辑:　涓涓)