文章编号: 1671-8860(2006)04-0348-04 文献标志码: A

GIS中文查询系统的词典设计与分词研究

徐爱萍[1,2] 边馥苓[1]

(1 武汉大学空间信息与数字工程研究中心, 武汉市珞喻路129号, 430079)
(2 武汉大学计算机学院, 武汉市珞喻路129号, 430079)

摘 要: 在分析系统应用领域的基础上设计了系统词典, 提出了基于扩展ER空间数据库环境的全匹配分词算法, 分析了算法的复杂度, 解决了切分歧义和未登录词的问题, 并通过一个实验原型对设计进行了验证, 为GIS中文查询语句的正确理解提供了有效的语义信息。

关键词: GIS; 中文查询; 系统词典; 分词; 全匹配
中图法分类号: P208

国内外学者在数据库的自然语言查询中已经进行了多年的研究和探索[1], 取得了很大的进步, 但基于中文语句的数据库查询离实际应用仍然有距离, 其主要原因是汉语不同于西方语言, 存在切分歧义和未登录词问题[2]。受限语言[3]的基本思想是在系统应用领域的基础上, 对自然语言适当加以限制, 以显著降低复杂性和减少机器处理的困难。因此, 研究基于系统应用领域的中文数据库查询接口是可行的, 因为数据库查询句相对简单, 表达的语义和查询的内容比较明确, 歧义大大减少, 因此, 对添加的限制是可以接受的, 相关研究可参见文献[4-6]。但现有文献中, 对空间数据库进行中文查询的研究成果还不多见。本文在分析系统应用领域的基础上设计了系统词典, 提出了基于扩展ER空间数据库的全匹配分词算法。

1 系统词典设计

自然语言理解中, 词典是中文分词、语法分析、语义理解的基础, 基于受限汉字的词典设计必须对应用领域进行分析和研究, 本系统的应用领域是基于扩展ER空间数据库实体关系模型(如图1所示)的一系列中文查询语句。

为便于实现通用、可靠的分词系统, 把要提取的词条分为三大类: 通用词、空间对象专用词、空间关系词, 分别存放在相应的词典中。

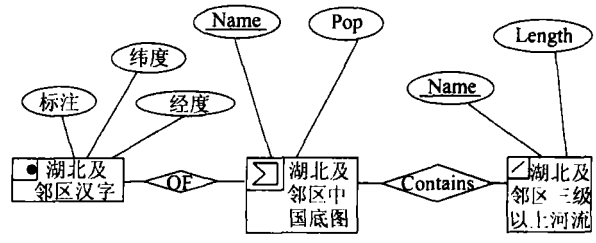

图1 扩展ER空间数据库实体关系模型
Fig.1 Relationship Model of Extended ER Spatial Database Entity

1.1 通用词典

属于领域无关词类的词存储于系统的通用词库中[7], 在系统移植时, 这些词一般不需要修改。其分类如下: ① 连词、介词、量词、助词、数词和限定词, 在词典里没有形式描述; ② 查询动词放在查询语句的最前面; ③ 疑问词是判断查询语句结构的关键词; ④ 关系词用于形成关系表达式, 如"等于/为、以上/大于、不小于、小于/以下、不大于/不超过、不等于"等, 它们在词典中的语义描述分别为"=、>、≥、<、≤、<>"等; ⑤ 逻辑词指"是/真、不/假/否、或/或者、异或、并/并且/和"之类的词汇, 它们在词典中的语义描述分别为"TRUE, FALSE, NOT, OR, XOR, AND"等; ⑥ 函数词对应着一个函数, 如"总数、平均数、计数、距离、面积、长度"等, 它们在词典中的语义描述分别为"SUM, AVG, COUNT, Distance

收稿日期: 2006-01-17。
项目来源: 武汉市青年科技晨光计划资助项目(20045006076-16)。

(Shape1, Shape2), Area (Shape), Length (Shape)"等; ⑦ 排序词主要用在排序短语中, 如"从大到小、从高到低"等, 这类词在词典里的形式描述为"ORDER BY"。

1.2　空间对象专用词典

所谓空间对象词, 是空间对象自然语义的标识[7]。在空间数据库中, 同一层内的空间对象都有一个 FID 作为标识符, 在相应的属性字段中, 一般有一个字段存放该空间对象的自然名称, 则该字段就可以作为空间对象的自然语义标识。

空间对象词分为两类, 一类表示空间对象集合, 一类表示空间对象个体。如在空间数据库中, 有一个"河流"层, 则该层所表示的空间对象的集合为河流(名称、经度、纬度、长度, 视具体情况而定), 在该层中, 有若干线空间对象, 每一个线空间对象对应有一个自然名称标识它, 如长江、黄河等。

作为受限汉字理解系统, 其词典的设计与应用领域紧密联系, 设三个基本表(表 1、表 2、表 3)中有部分空间对象。为了便于程序搜索, 本词典的结构与通用词典相同, 只是定义不同, 在此设计了以下语义信息: ① 词(word); ② 词类(wordtype); ③ 描述(describe)。

表 1　湖北及邻区中国底图.TAB

Tab.1　Chinese Base Map of Hubei and Nabe.TAB

Name	Pop.	Obj.
湖北	847 600.000	Polygonid=1
安康	84 300.000	Polygonid=2
九江	43 100.000	Polygonid=3
…	…	…

表 2　湖北及邻区三级以上河流.TAB

Tab.2　River of Above Third Class of Hubei and Nabe.TAB

Name	Length	Obj.
长江	6 300	LineStringid=1
黄河	5 500	LineStringid=2
…	…	…

表 3　湖北及邻区汉字.TAB

Tab.3　Chinese Characters of Hubei and Nabe.TAB

纬度/(°)	经度/(°)	标注
30.53	114.35	武汉
30.18	115.07	黄石
30.68	111.19	宜昌
…	…	…

本系统基于扩展 ER 模型的部分专用词典如表 4 所示。这种定义方法可以使后面的处理变得容易。另外, 还要考虑分词的二义性, 如"名称"可以对应"城市名称", 也可以对应"河流名称", 所以在定义时, 要避免这种二义性, 在词典中不要出现"名称"的词, 而要用"城市名称"和"河流名称", 并且还要有"城市的名称"和"河流的名称", 其中, "城市名称"和"城市的名称"对应的词类和描述完全相同。

表 4　部分专用词典

Tab.4　Part Special Dictionary

Word	Wordtype	Describe
城市	实体	湖北及邻区中国底图
城市名称	属性	湖北及邻区中国底图. Name
湖北	属性值	湖北及邻区中国底图. Name="湖北"

1.3　空间关系词典

空间关系词典要根据几何对象之间的空间关系而建立, 本测试系统是在 MapInfo MapXtreme 平台下完成的。在 MapInfo MapXtreme 中, 若两个表有图形对象, MapInfo MapXtreme 能根据对象之间的空间关系来连接表。这样, 即使几个表并没有共同列, 也能将其连接。MapInfo MapXtreme 有以下地理运算符: Contains 表示对象 *A* 包含对象 *B*, *B* 的中心在 *A* 的边界内任一点; Contains Entire 表示对象 *A* 完全包含对象 *B*, *B* 的边界完全位于 *A* 的边界之内。Within 表示对象 *A* 位于对象 *B* 之内, *A* 的中心在 *B* 的边界之内; Entirely Within 表示对象 *A* 完全位于对象 *B* 之内, 对象 *A* 的边界完全位于 *B* 的边界之内; Touchs 表示对象 *A* 与对象 *B* 相邻。因此, 线与线相邻、线与面相邻映射为 Touchs; 面的叠加、重合映射为 Object *A* contains entire object *B* 或 Object *B* entirely within object *A*; *A* 包含 *B* 映射为 Object *A* contains object *B* 或 Object *B* within object *A*; 距离映射为 Distance, 面积映射为 Area, 周长映射为 length, "在…内"映射为 Object *A* within object *B*, "在…外"映射为 not (Object *A* within object *B*)。

方位关系可分为三类: 绝对的、相对目标的和基于观察者的。绝对方位关系是在全球参照系统的背景下定义的, 如东、西、南、北、东北等; 相对目标的方位关系根据所给目标的方向来定义, 如前、后、左、右、上、下等; 基于观察者的方位关系是按照专门指定的, 根据观察者的参照对象来定义。绝对方位关系东、南、西、北分别映射为 Object *A* East Object *B*, Object *A* South Object *B*, Object *A* West Object *B*, Object *A* North Object *B*。

2　词的切分

分词的目的就是要得到包含正确切分结果并

符合词典意义的所有切分串的集合,即把查询句划分为若干个有意义的词条。同时,由于数据库中文查询句中属性值的未登陆词大大多于一般中文文本,所以要解决未登陆词的问题。在此,本文设计并实现了一种新的切分算法——基于扩展 ER 空间数据库应用领域环境的全匹配分词算法[2,4],根据它具体的空间数据库,建立空间关系词库、空间对象词库和相关普通词库,对句子进行多次扫描、消歧,并建立每个词的分类信息。

2.1 算法描述

设自然语句串为 input_String = $S_1 S_2 \cdots S_n$,切分算法描述如下。

1) 若语句串的长度 input_string_len= 0,则结束;

2) 初始化 search_str= "",m= 0,匹配结果的记录号字符串 SIDlist= "";

3) 若 S_1 的 ASC Ⅱ码值大于 127 (即不是 ASC Ⅱ字符,是一个汉字的一半),则取 $S_1 S_2 \cdots S_n$ 的前 2 个字符→search_str(即 w 为一个汉字),m + = 2;若 S_1 的 ASC Ⅱ码值小于 127 (即是 ASC Ⅱ字符),则取 $S_1 S_2 \cdots S_n$ 的前 1 个字符→search_str(即 w 为一个字符),m+ + ;

4) 查机器词典,匹配左端字段是 search_str 的词,并按词的长度由大到小排列;

5) 若匹配成功,则在 SIDlist 中记录下本次匹配结果的记录号,input_string_len 减 1,匹配的字符串中去掉 S_1,转第 3)步(此时 S_1 位置上的字符是 S_2);

6) 若 SIDlist= "",则匹配不成功,将干扰字存入字库,转用户交互或修改查询句或学习新词;

7) 若 LEN(SIDlist) < > 0,则通过 $S_i S_{i+1} \cdots S_{i+m-1}$ 查找其词典,取出描述信息,压入句子栈,结束本次切分;

8) 转第 1) 步,直到 input_string_len= 0。

2.2 算法的时间复杂度

分词方法的时间复杂度是衡量分词方法优劣的一个非常重要的标准。在其他条件相同的情况下,如果时间复杂度越低,说明这种分词方法的性能越好,分词速度越快[2]。计算分词方法的复杂度时,一般选择匹配次数作为基本衡量单位。

设句子的长度为 n,词典中词的总数为 m,对于待分析的字串 $S = s_{i1}, s_{i2}, \cdots, s_{in}$,取 s_{i1},查词典中是否有以 s_{i1} 起始的词,若有,则将它们按长度排序放在列 SIDlist 中;否则,s_{i1} 不是一个词。再取 s_{ij} ($2 \leqslant j \leqslant n$),查 SIDlist 中的词,是否有词的前 j 个字与 $s_{i1} \cdots s_{ij}$ 匹配,若有,则更新 SIDlist,j+ + ;若没有,则从 SIDlist 中取与 $s_{i1} \cdots s_{i(j-1)}$ 完全匹配的词作为匹配结果,同时令 $S = s_{ij} \cdots s_{in}$,重复上述操作至 $j > n$ 为止。所以该算法的时间复杂度 $T = O(mn)$,这是在最坏情况下的结果,实际上达不到这个数,因为词的第二个字以后的不需要查遍整个词典了,最好的情况下可以一次查到所需要的词。

2.3 与已有算法的比较

该算法的特点从以下几个方面来体现。

1) 查询速度。前面已经分析,该算法的查询速度是线性级的,而文献[4]中切分算法的查询速度是指数级别的,当查询语句较长时,其速度显然低于该算法。

2) 切分的完备性。文献[4]中的切分算法具备切分的完备性,即把所有的可以切分出的情况都考虑进来了;本文的算法虽不具备该性能,但该系统追求的正是分词的惟一性。对于分词的歧义性问题,在设计词典时,要尽量避开这样的情况,如果避免不了,将会影响分词的正确性,这时需要根据人机交互完善词典库。不过,出现这种情况的可能性非常小。

对于未登录词,分词子系统中提供交互模块来学习新词语并指定解释方法,必要时,提供切分模式,系统会自动记录,并在以后的切分中使用这些知识。这样,系统会随着使用者的训练而增加词语的切分处理能力,由此可以解决未登录词的问题。

3 算法测试

根据上述思想,在设计的系统词典的基础上,进行了分词算法的测试。

分词后,将词典中的形式描述转为句子涉及到的实体、查询目标、查询条件、关系运算符、空间关系运算符等。通过空间查询文法转换规则,建立短语结构,对查询语句进行规格化处理,生成中间形式化查询语言。然后将中间形式化查询语言转为标准的 SQL 查询语句,通过空间运算和查询,返回查询结果。GIS 中文查询界面如图 2 所示。

系统分词结果是"列出/所有/与/湖北/相邻/

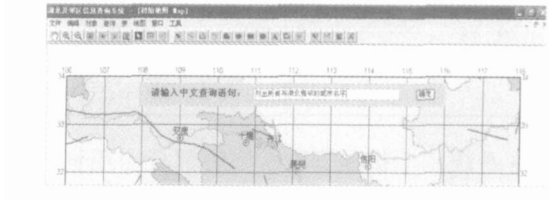

图 2 GIS 中文查询界面

Fig. 2 Interface of GIS Chinese Query

的/城市/名字", 其中, "湖北"是属性值, 在查询语句中对应于"湖北及邻区中国底图. Name='湖北'";"相邻"是空间关系词, 在词典中对应"Object *A* Touchs object *B*", 在转为 SQL 语句时, *A*、*B* 将用表名代替;"城市"是表名, 对应词典中的"湖北及邻区中国底图", 名字是属性, 对应词典中的"Name", 分词结果如图 3 所示。系统按以上方式对 45 条查询语句进行了测试, 有 38 条查询语句能正确分词。分析出错原因, 主要是出现错别字, 如"发源地"写成了"发原地", 这时只好当未登录词处理了。如果能采取模糊查询解决类似的问题, 将是分词中的一个新的突破。本课题将在这方面作进一步的研究, 以达到高智能化的分词效果。

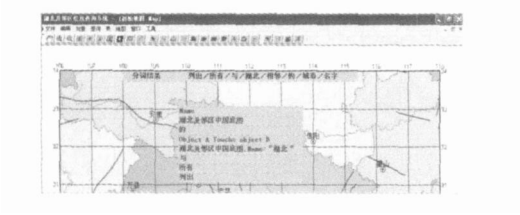

图 3　分词结果

Fig. 3　Result of Segment Word

测试结果表明, 本文所设计的能解决词的二义性问题和未登录词问题的词典与能有效解决切分歧义性问题的分词算法结合起来对切分词、提取词的语义信息效果良好, 但基于中文的 GIS 查询系统要求分词、文法分析、语义分析和语句分析在单个阶段提供惟一正确的结果, 从而完成自己的任务是不实际的, 只有这几个处理阶段互相协作、互相补充才能更好地完成查询句的理解任务。

参 考 文 献

[1] Rashid A, Shariff B M, Max J. Egenhofer, David M. Mark. Natural-Language Spatial Relations Between Linear and Areal Objects[J]. Interbation Journal of Geographical Information Science, 1998, 12(3): 225-246

[2] 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[J]. 中文信息学报, 1998, 12(4): 53-59

[3] 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[J]. 软件学报, 2002, 3(4): 537-543

[4] 崔宗军, 唐世渭, 杨冬青. 基于 ER 模型和受限汉语的数据库中文查询语言研究[J]. 中文信息学报, 2000, 15(4): 7-13

[5] 孟小峰, 王珊. 中文数据库自然语言查询系统 Nchiql 设计与实现[J]. 计算机研究与发展, 2001, 38(9): 1 081-1 082

[6] 马林兵, 龚健雅. 面向自然语言的空间数据库查询研究[J]. 计算机工程与应用, 2003, 22: 16-18

[7] 周炎坤, 李满春. 基于中文的 GIS 查询界面的初步研究[J]. 科技通报 2001, 17(1): 35-39

[8] 吴胜远. 一种汉语分词方法[J]. 计算机研究与发展, 1996, 33(4): 310-311

第一作者简介: 徐爱萍, 博士生, 副教授。现从事网上信息管理、WebGIS、自然语言理解等研究。

E-mail: xap1464@126.com

Dictionary Design and Word Segmentation Research in Chinese Query of GIS

XU Aiping[1,2]　*BIAN Fuling*[1]

(1 Research Center of Spatial Information & Digital Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

(2 School of Computer Science, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

Abstract: The dictionary based on the condition of use is designed, and the algorithm of word segmentation for full matching based on the condition of extended entity ration spatial database is put forword. The complex grade is analyzed. The problem of different meanings and without dictionalry is solved effectively and the test is done. It offers vaild semantic information for computer to understand the sentence of Chinese query accurately.

Key words: GIS; Chinese query; system dictionary; word segmentation; full matching

About the first author: XU Aiping, Ph. D candidate, associate professor. She is engaged in management information system on Web, WebGIS and natural language understanding.

E-mail: xap1464@126.com