基于地名首字母模糊查询的关键技术研究

徐 齐，陈 杨

（92117部队，北京 100072）

摘 要：根据地名信息的特点，提出一种基于地名首字母的模糊查询方法。采用Access数据库，结合微软自带的字库实现了地名首字母的提取与入库，特别是解决了多音字在首字母提取及入库时的处理，并提供了一种文件批量输入的方法，实现了基于地名首字母的模糊查询。设计开发的地名查询系统，使用简单，效率也得到了极大的提高，可以获取对应的地理位置，为后继GIS系统开发提供方便。

关键词：模糊查询；地名库设计；首字母提取；地名查询

中图分类号：P281　　**文献标识码**：A　　**文章编号**：1672－5867(2016)03－0184－03

Research of Key Technology for Query Methods Based on Palcename Initial

XU Qi, CHEN Yang

(92117 Troops, Beijing 100072, China)

Abstract: This paper find a new fuzzy query method based on palcename initial according to the recent palcename query methods used in GIS, design the database according to the character of toponym geographic information. Realized the distill and input database for palcename initial, especially solved the difficulty of polyphone, and also gave a input way for file. The palcename query system that this paper designed used simply, it not only could upgrade efficiency, but also could give the other geographic information. In fact, it will bring much more convenient for GIS in the future.

Key words: fuzzy query; palcename database design; distill initial; palcename query

0 引 言

地名是客观存在的社会现象，是人类对地理环境具有特定位置、范围及形态特征的地方所共同约定的语言代号[1]。地名查询是地理信息系统及地名资料管理中必不可少的关键技术之一，是通过输入地名，在地名数据库中获取相应的地理信息，从而可以在地图上准确定位，以达到信息检索的目的。地名查询技术的应用，大大减轻了以往手工作业的工作量，而且更加准确，效率更高。由于地名查询在GIS系统中使用极其广泛，其查询方式也多种多样，主要归类为以下3种[2-3]：①汉字输入查询：用户在系统交互界面上输入想要了解的地名汉字全称或其包含的部分汉字，从而得到查询结果。这种查询方式设计起来比较简单，可以满足一般的查询应用。但是在遇到一些不常用的汉字时，查询效率较低；②同音地名查询：不需要直接输入地名汉字，仅需要提供地名的拼音字符串，就可以获取相应的地名信息。在设计上充分考虑了汉字输入所存在的问题，在应用上更加方便。但是，这种查询方式需要输入完整的地名拼音字符串，倘若遇到汉字较多的地名，也不太适合。如“鄂温克自治旗”，就需要输入“ewenkezizhiqi”，需要击键13次之多。拼音不同于汉字，人们往往对其敏感程度不够，这种输入法不仅效率很低，而且极易出错。一旦输入的拼音中含有错误字符，很难得到预期的结果；③基于地名首字母的模糊查询：通过输入地名的部分或全部首字母就可以得到地名，输入的内容在键盘上就可以全部找到，不需要对照字库，减少了击键次数，使上述问题得到了很好的解决。即使不会读不会写，只要大致知道声母发音就可以进行查询，更适用于一些大众化的查询，方便使用者并提高了查询效率。

很明显，基于首字母进行的查询方式明显要好于汉字输入查询和同音查询。本文通过比较分析，设计一种基于首字母的模糊查询方法。

收稿日期：2015－08－24

作者简介：徐 齐(1977－)，男，四川乐山人，工程师，硕士，2002年毕业于西南交通大学通信与信息系统专业，主要从事地理信息系统应用工作。

1　地名库的设计

在进行地名查询时,无论是用哪种方式,地名数据库的管理都是必不可少的。合理规范的地名库设计,是提高地名查询效率的前提。

1.1　单字拼音表的生成

地名是由单个汉字组成的,获取地名首字母,首先需要建立包含所有汉字的拼音表,同时,该表中还应包含多音汉字的各个拼音,多音字处理是地名数据库设计中的难点。

采用 Windows 系统提供的输入法生成器 IMEGEN.EXE,直接运行进入输入法生成器窗口。通过对全拼字典库进行逆转换,可以生成一个纯文本文件 WINPY.TXT,这个纯文本文件中包含了所有微软字符库中的汉字及词组的各种拼音形式,去除所有词组记录即可生成一个汉字拼音字典查询数据库[4]。

1.2　地名库的设计

进行地名查询,如果直接用单字表来查询,查询效率极低。为了提高效率,需要建立一个地名数据库,作为查询基础数据。使用者所输入的查询条件最终也是需要在该库中进行查找,地名库设计的好坏直接影响着查询的精确性与效率。

进行基于首字母的地名查询,地名数据库必不可少地要有"地名""首字母"的属性值。考虑到实际应用,作为 GIS 系统的一部分,还应包括其地理定位信息及经纬度属性,本文最终在 Access 表中设计了"地名""首字母""经度""纬度"等 4 个字段名。

对于一些地名,其对应的首字母仅有一个,如"北京",其首字母为"bj"。但大多数地名都存在多音字的问题,其首字母就不仅一个,则其所对应的地名首字母组合也就不可能唯一。如"广州",其中,"广"是多音字,有"an""guang""yan"三种读音,所以,其整个地名拼音首字母就有"az""gz""yz"三种。这些首字母如何在数据库中存放,是地名库设计需要重点考虑的因素。结合数据库连接方式,综合考虑数据查询效率等因素,将相同地名的不同首字母组合放入一个属性字段"首字母"中,不同组合之间用分隔符如"|" 隔开。如地名"广州",其"首字母"列中即为"az|gz|yz"。大大减少了数据表的数据量,使数据表的设计更加人性化,更加合理,其最主要的优点就是可以提高查询效率,是地名数据组织方式的首选。

2　地名信息的入库与查询

2.1　地名首字母提取及入库

基于地名首字母进行模糊查询需要有首字母地名库的支持,需要通过从地名汉字中提取出首字母进行构建。

2.1.1　地名首字母的提取

首先,获取地名中汉字的个数。将地名看成一条字符串,获得该字符串的长度。每个汉字字符的长度是 2,以地名字符串长度除以 2,获取汉字个数。

地名首字母的提取以拼音单字表作为基础数据,在建立数据源之后,首先需要打开其中的单字表,再在单字表中查找地名中各个汉字的汉语拼音,然后再提取出各个汉字的首字母并组合。需要注意的是,地名中各个汉字的判断查询,应当有一定的秩序性,按照从前至后或者从后至前的顺序提取首字母,结合上面提到的汉字字符单个长度为 2,当循环次数增加时,可依次提取地名字符串中最右边的两个字符(即最右边的一个汉字)的拼音首字母。如"乌鲁木齐",第一次提取的是其中"齐"的首字母,第二次提取的就是剩下三个汉字中最右边的"木"的首字母。借助 SQL 语句及 ADO 数据访问方式,可以轻松实现记录表的打开及查寻。

若没有多音字存在,仅需依次将单个汉字首字母加入地名首字母组合中即可。但几乎每一个汉字都存在着多音字的情况,这里采用动态输入的方法来获取地名的首字母:在获取地名中汉字的各个不同读音的首字母后,依次加入已经存在的不完全首字母组合中。同时,删除首字母中存在的重复组合,得到最终的首字母集合。以"李庄"为例,可以从图 1 中看到其首字母的组合方式。

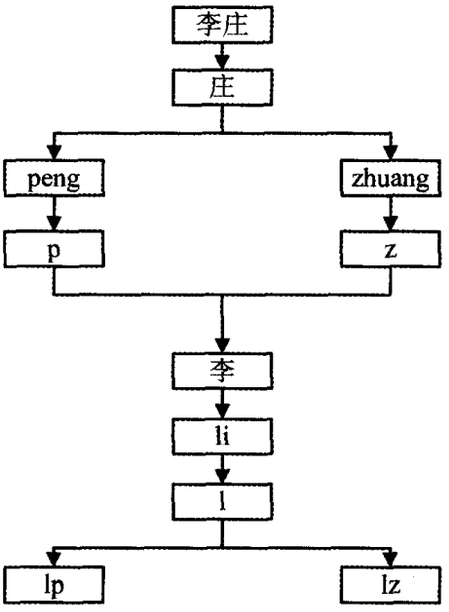

图 1　首字母提取顺序

Fig. 1　Extraction sequence

2.1.2　地名首字母的入库

地名首字母的入库也是基于首字母进行地名查询的主要技术之一。在地名首字母提取以后,打开地名库,首先需要判断地名库中是否已经含有相同的地名信息,假如没有相同的地名,则说明该地名是新建信息,我们只需要添加一条新记录,修改已有记录可采用相同的方式替换。

单个地名首字母的提取及入库,简单快捷。地名信息并非单独几个,全国县以上的地名就有几千个之多,手工录入费时费力,不切实际。这里提供了一种地名信息的批量录入方式——文件输入,更加符合面向对象的思想。根据单个地名信息输入的情况,文件信息包括地名、

经度、纬度三种属性,在入库时进行循环处理。

2.2 基于首字母的模糊查询

2.2.1 基于首字母的地名模糊查询

相对于精确查询,模糊查询得到是一类结果[5]。地名首字母的模糊查询是在地名首字母的提取与入库基础上进行的,也是使用者最终要应用的内容。它采用的就是一种人机交互的方式,通过使用者输入自己想要查询的首字母,计算机按照这样的条件在地名数据库中查找相似的地名信息,并最终返还给用户。

基于首字母的地名查询是在首字母地名库建成以后,按照地名首字母进行的模糊查询。如使用者想要获取"上海"的地理信息,则只需要在查询条件栏里输入"sh"即可,这是基于首字母的精确查询。在进行模糊查询时,使用者甚至不需要输入完整的首字母组合"sh",只输入"s"就可以显示首字母组合中包含该字符的所有地名。不过,当输入"sh"时,返回的查询也不会仅仅是"上海",还包括所有首字母中含有"sh"的地名,是一组结果。

2.2.2 查询结果的信息显示

在地理信息系统中,进行地名查询不仅是为了获取地名,而是想通过地名所对应的地理信息来实现定位或者获取其他信息。考虑到这种需要,在实现首字母地名查询的基础上,应该提供部分地名信息的获取方式。根据地名数据库的设计内容,主要获取的是地名的经纬度信息。原理与利用首字母进行地名查询是基本一致的,在获得一组满足首字母要求的地名后,选择目的地名,即可得到该地名的地理位置坐标。

在进行地名信息输入时,一般所定义的经纬度值是以小数的形式输入到地名数据库的,而在实际情况下,更为大家接受的是度分秒的形式。所以在信息显示之前,应当进行两者之间的转换,根据度分秒之间 60 进制的关系来实现这种功能,简单方便。

3 结束语

基于首字母进行模糊查询具有重要的使用价值,可以扩展到其他的查询系统中,操作简单,使用方便,充分体现了面向对象的理念,改变了以往查询系统对使用者文化程度特别是普通话水平要求较高的情况。使用首字母进行查询,无论用户是何种口音,只要大致了解查询条件的声母,就可以准确无误地实现查询,获取相应信息。

参考文献:

[1] 任家强,武文波. 基于 MapObject 和高分辨率遥感影像的地名查询系统建立[J]. 矿山测量,2005(1):11-14.

[2] 赵斌,顾彦慧. 采用 Java 实现的汉语拼音查询模块[J]. 计算机与现代化,2006(12):52-53.

[3] 王东,熊世桓. 基于拼音首字母查询的去重优化设计[J]. 贵州师范学院学报,2010,26(6):37-39.

[4] 张炜,唐慧强. 将汉字转化为拼音的研究与实现[J]. 计算机应用,2003,23(Z1):4-5.

[5] 阎红灿. 语音模糊查询在信息管理系统中的实现[J]. 计算机系统应用,2006(3):52-55.

[6] 於建峰,王光霞,万刚. 基于汉字模糊音的地名查询方法设计与实现[J]. 测绘科学技术学报,2008,25(2):120-123.

[7] 陈健,张斌,梁汝鹏. 基于语义类型本体的地名组配查询[J]. 地理空间信息,2013,11(3):126-128.

[编辑:栾丽杰]

(上接第 183 页)

三维激光扫描仪与数码相机结合采集地面数据,既获取了高精度的三维矢量数据,同时还获取了真实的光学影像数据,两种数据匹配后,为后期地形图编辑成图提供了可视化模型,提高了数据编辑的可靠性,缩短了成图周期。

三维激光扫描仪获取的海量点云数据在计算机处理过程中遇到瓶颈,必须抽稀数据后,才能顺利进行编辑工作,如何快速、高效地处理海量点云数据是未来研究的一个课题。

参考文献:

[1] 关云兰. 地面三维激光扫描数据处理中的若干问题研究[D]. 上海:同济大学,2008.

[2] 曹先革,杨金玲,司海燕,等. 地面三维激光扫描点云数据精度影响因素及控制措施[J]. 测绘工程,2014,23(12):5-7.

[3] 张毅. 地面三维激光扫描点云数据处理方法研究[D]. 武汉:武汉大学,2008.

[4] 施贵刚. 地面三维激光扫描数据处理技术及作业方法的研究[D]. 上海:同济大学,2009.

[5] 王旭,王昶. 三维激光扫描仪数据的建模研究[J]. 北京测绘,2013(2):55-57.

[6] 臧克. 基于 Riegl 三维激光扫描仪扫描数据的初步研究[J]. 首都师范大学学报:自然科学版,2007,2(28):79-81.

[7] 刘会云,李永强,刘文龙,等. 三维激光扫描仪平缓地形分区扫描方法研究[J]. 测绘工程,2015,24(4):6-10.

[8] 邓飞,张祖勋,张建清. 利用激光扫描和数码相机进行古建筑三维重建研究[J]. 测绘科学,2007,32(2):29-30.

[编辑:栾丽杰]