

文章编号:1001-9081(2010)07-1953-03

基于分级地名库的中文地理编码

孙存群,周顺平,杨林

(中国地质大学(武汉)信息工程学院,武汉 430074)

(suncunqun@126.com)

摘要:地理编码在城市空间定位和分析领域内具有非常广泛的应用,但由于中文地址没有统一的规范和固定的模式,目前中文地址编码尚无很完善的解决方案。针对这一问题,采用基于分级地名数据库的中文地理编码方法,并详细阐述了实现该方法的关键技术:地名数据库的数据模型、地址串的拆分和地址匹配技术。最后通过实际数据进行验证,实验结果表明,该方案能较好地解决绝大多数地址数据的匹配问题。

关键词:地名数据库;地址模型;地理编码;地址拆分;地址匹配

中图分类号: TP311 **文献标志码:** A

Chinese geo-coding based on classification database of geographical names

SUN Cun-qun, ZHOU Shun-ping, YANG Lin

(College of Information Engineering, China University of Geosciences (Wuhan), Wuhan Hubei 430074, China)

Abstract: Geo-coding is widely used in urban spatial location and analysis, but there is no perfect solution on Chinese geo-coding for the present because there is no uniform specification and fixed model for Chinese geographical names. To solve this problem, the authors used Chinese geo-coding based on the classification database of geographical names in this paper, and detailed the key technologies to realize the program: database of geographical names data model, address split, and address matching. Finally, the authors verified the program through the actual data. The experimental results show that the program can solve most of the problems about address matching.

Key words: database of geographical names; address model; geo-coding; address split; address matching

0 引言

地理编码的概念有别于一般的编码定义,它不是用数字或字母来代表某一地物,而是将地址数据映射成地理坐标的过程,即通过把表现空间位置的地址描述数据与空间坐标相关联,从而将地址数据转换成可以被用于地理信息系统的地理坐标^[1]。

通过地理编码,能够将工商、税务、信用、规划、建设等经济社会部门的资料和数据库中的地址描述转换为真实的地理坐标(或经纬度),并映射到地图、遥感影像上,实现空间数据与非空间数据共享整合,进而可以完成对经济社会信息的分析、统计、管理、制图和可视化表示,为相关部门提供实时、准确和权威的集成与融合工具,以支持政府的管理和决策。对公众而言,地理编码技术可以提供便利的地址查找工具,帮助用户实现对未知地区、地点的快速查询和情况了解,减少出行时的盲目性和找不到目的地的困扰,节约时间和金钱。

因此研究并实现一个适应国内地址现状并能满足各级政府部门和普通大众对中文地址的建库、地址管理和维护、地名定位和查找需求的地理编码系统是非常有意义的。目前国外已有针对欧美地址模式的英文地理编码解决方案,例如, ArcInfo 的 Geocoding, MapInfo 的 MapMarker。但是由于中文地址、地名的特殊国情,中文地址没有统一的规范和固定的模式,如果套用现有的欧美地理编码的解决方案,则国内大部分

城市现行的地名地址体系就要作大的调整和规范,这是不现实的。国内也有一些针对中文地址解析和定位的软件,比如,北京长地计算机公司的“寻址神”、北大方正的“小红帽物流管理系统”等,但它们仅局限于特定领域和特定范围内^[2]。针对这一问题,本文采用基于分级地名数据库(简称地名库)的中文地理编码方法,它通过 TRIE 字典树的中文分词技术来自构造分级地名库,并对地名库的地址要素字段建立索引,从而实现地址的快速匹配。

1 基本概念

本文涉及到地址要素、地址要素级别、地址级别和标准地址四个概念。

地址要素是指在某一限定区域内,可以指定某一具体范围的地址。一个通信地址是由一个或多个地址要素组成,每个地址要素为地址串中的一个相对独立的部分。例如“武汉市鲁磨路 368 号”由三个地址要素组成,分别是“武汉市”、“鲁磨路”、“368 号”。

地址要素级别是为了说明地址要素的从属关系,用来标志其级别的数字。大多数地址都是按照地址的层级关系来描述的,比如上例中“368 号”从属于“鲁磨路”,而“鲁磨路”又从属于“武汉市”,因此可以构建基于层级地址的地址模型(Level Based Model, LBM)^[3]。参考建设部《中华人民共和国城市建设标准》——《城市市政综合监管信息系统地理编码》

收稿日期:2009-12-10;修回日期:2010-02-26。 基金项目:国家科技支撑计划项目(2006BAB10B02-B)。

作者简介:孙存群(1983-),女,湖北枣阳人,硕士研究生,主要研究方向:地理信息系统网络分析、地址编码;周顺平(1967-),男,云南楚雄人,教授,博士,主要研究方向:地理信息系统基础软件、空间数据库;杨林(1982-),女,山西永济人,讲师,博士,主要研究方向:地理信息系统、空间数据库。

(CJ/T215—2005),可以将地址要素划分为11个级别,级别依次从高到低,如表1所示。

表1 地址要素级别

级别	含义	级别	含义
1级	中国	7级	地片、区片
2级	省、直辖市	8级	道路、街巷
3级	省会、地级市	9级	门牌号
4级	区、县、市	10级	楼牌号
5级	街道、乡镇	11级	POI、标志物
6级	社区、村		

地址级别是指该地址的最低级别地址要素所处的级别。上例的地址处在第9级。

标准地址是对某个地址按照一定的标准进行规范化后的地址。某个级别的标准地址是不低于当前级别的所有地址要素的连接。一个地址可以根据需要自定义所需的级别,不必将11个级别全部指定。例如,可以定义一个标准,规定一个地址必须包括3、4、8、9级,那么上例的标准地址应该为“武汉市洪山区鲁磨路368号”,第4级的标准地址为“武汉市洪山区”。

2 地理编码的原理和方法

地理编码系统的实现原理为:将用来创建地名库的地址数据中的标准地址记录拆分成标准的地址要素,并根据标准地址要素的级别来创建分级地名库,地名库创建完成后就可以利用该库实现地址匹配。进行地址匹配时,需要创建地址索引,然后将待匹配数据中的地址数据在地址索引中进行地址匹配,如果匹配成功,就将地名库中的地理坐标赋给待匹配数据,从而实现对此记录的地理编码。由此可见实现地理编码的关键技术主要有地名库的创建、地址串的拆分和地址匹配。

2.1 地名库的数据模型

地名库是地理编码系统的核心基础,地理编码可以利用地名库来建立地址与空间信息的对应关系^[4]。地名库通常是国家或地区普查结果,包括路街巷或门牌号码的坐标位置、标准名称。目前中国各地区的地址信息并没有按照统一的格式入库,这导致了没有一个统一的全国范围内的地名库。为了统一各个地区各个领域地名库,这就需要设计出一套合理的地址数据模型,它不仅要对地址数据进行描述,还要能明确地址数据之间的联系。

地名库根据地图数据来创建,当数据库创建完成后就与地图数据脱离关系。地名库对基础地址数据进行管理,可以实现对地址数据添加、删除、修改等完善地址要素及其相关信息的功能,随着地名库的不断完善,匹配结果也将更准确。由于地址数据的运用范围及详细程度不一样,地名库不一定包含11个级别的地址数据,用户可以按照需求指定地名库所包含的地址要素级别,这种分级表示的方法使地址要素之间的从属关系更为清晰。地名库包括3种表:基本表、门牌号索引表和别名表。其中地名库的基本表主要记录了每个地址要素的空间及非空间信息,其字段定义如表2所示;门牌号索引表记录了道路的类型,以及道路的左右起始和终止门牌号,它是为了便于根据所描述地址的道路门牌号信息,利用插值原理将匹配结果定位到街道的两边^[5];别名表记录了地址要素的别名及标准名称的相关信息,一个标准名称可以对应多个别名,它主要是为了处理某些地名有不同称呼的情况。

表2 地名库的基本表结构

字段名称	说明	类型
CodeID	地址要素的编号	long
AddrFtName	地址要素名称	char[64]
AddrLevel	地址要素的级别	short
PrtAddrFtName	父地址要素名称	char[64]
PrtTbID	父地址要素所在表	long
PrtCodeID	父地址要素的编号	long
CoorX	X坐标	double
CoorY	Y坐标	double
RectWidth	范围宽度	double
RectHigh	范围高度	double
IsValid	标识地址是否有效	char

2.2 地址串的拆分

地名库的创建过程就是将拆分出来的地址要素及其空间信息添加到数据库的过程。地址串的拆分是将一个标准地址记录拆分成多个标准地址要素的过程,是地理编码的第一步。国外通常采用文字片段分割法来拆分地址串,但是由于中国各地区的用语习惯不同,区分不同级别的文字片段很复杂,有的甚至没有明显的文字片段,因此采用文字片段来拆分地址串有一定的困难。

通过分析发现,在地址要素级别为1至4级上的地址要素基本是固定的,变化不会特别大,因此拆分的重点放在5至11级的地址上,而某一级别的地址要素为剔除上一级别的所有地址要素后余下的部分,由此可见地址拆分类似于中文自动分词^[6]。所谓中文分词,即根据分词词典,在字符串中查找从某一指定位置开始的最长词(对应正向最大匹配分词法)。此时,分词词典即为利用地址要素所创建的分词词典,字符串即为地址串,最长词即为地址要素。

目前对于中文分词大多采用分词词典机制,主要有三种:整词二分法、TRIE索引树法和逐字二分法。基于词典的分词方法其精度依赖于词典的精度和对歧义的有效切分,速度则取决于所设计的词典结构,详细分析请参考文献[7]。地址分词又有别于中文自动分词,由于地址分词所基于的词库结构不一样,在地址分词中,每个地址要素(相当于词)含有多个字段属性,如:地址要素名称、经纬度、父地址等信息,这些字段的加入,使得拆分地址变得更加准确^[6]。在本系统中,笔者采用TRIE树词典的中文分词技术来实现地址串的拆分。

用来创建地名库的数据需满足如下条件:1)地名库指定的每一个级别的地址数据为该级别的标准地址;2)待创建的数据中包含地理坐标数据,或者可以根据图形数据得到地理坐标数据。

给定某个级别的标准地址,拆分其地址要素的过程描述如下。

1)初始化:当前处理级别为地名库指定级别中的最高级别,当前地址串为给定的标准地址。

2)若当前处理级别小于给定标准地址的级别,则根据当前处理级别的地址词典查找地址串中所包含的最大地址要素,并继续下一步;否则当前地址串即为给定级别的地址要素,转到步骤5)。

3)若在地址词典中查找最大地址要素成功,则需判断查找到的地址要素的父地址与前次查找到的地址要素是否相同,并继续下一步;否则拆分结束。

4)若查找到的地址要素的父地址与前次查找到的地址

要素相同,则当前处理级别为地名库指定级别中的下一个级别,当前地址串为剔除查找到的最大地址要素后余下的部分,转到步骤2);否则拆分结束。

5)根据地址词典判断得到的地址要素是否已经存在,若存在,则需判断该地址要素与已经存在的地址要素是否具有相同的父地址,并继续下一步;否则,该地址要素即为当前级别的地址要素,并将其相关信息添加入地名库中,拆分成功。

6)若得到的地址要素与已经存在的地址要素具有相同的父地址,则拆分结束;否则,该地址要素即为当前级别的地址要素,并将其相关信息添加入地名库中,拆分成功。

注意,将地址要素及其相关信息添加入地名库中的同时,需要将地址要素同时添加入 TRIE 树地址词典中,这样处理当前级别的地址串时,其高级别的地址词典已经生成,可以直接利用。处理到门牌号级别时需要将道路的类型,以及道路的左右起始和终止门牌号添加到门牌号索引表中。

2.3 地址匹配

地址匹配是将待匹配的地址串通过一定的匹配策略在地名库中查找出对应的地理坐标及标准地址的过程,是地理编码的核心部分。地址匹配的关键是确定匹配策略和对地名库的查询比较。一般来讲,用户所给定的待匹配地址串为某一级别的地址要素或者为某几个级别的地址要素的连接,因此在对非标准地址进行匹配时,可以采用正向最大匹配原则,按照从高级别到低级别的顺序,依次在地名库中找出所给定地址串中所包含的最大地址要素,从而实现模糊匹配。为了减少查询和比较次数,同时保证匹配的成功率和准确率,较好的方法就是为地名库中的地址字段建立索引^[8]。笔者同样通过 TRIE 树词典对地址要素字段创建索引,地址匹配的过程就是在每个级别的 TRIE 索引树中查询最大地址要素的过程。

给定一个待匹配的地址串,查找其标准地址及地理坐标的过程描述如下。

1)初始化:当前处理级别为地名库指定级别中的最高级别,当前地址串为给定的待匹配地址串。

2)若当前处理级别小于等于地名库指定级别中的最低级别,则根据当前处理级别的地址词典对待匹配的地址串进行正向最大匹配,从而找出地址串中所包含的最大地址要素;否则,转到步骤7)。

3)若查找的最大匹配地址为部分匹配,则转到步骤4);若查找的最大匹配地址为完全匹配,则转到步骤5);若查找失败,则在别名表中查找最大匹配地址,成功则转到步骤5),失败则将当前处理级别设置为地名库指定级别的下一个级别,并转到步骤2)。

4)定义临时处理级别,并将临时处理级别的初始值设置为当前处理级别降低一级,根据该临时处理级别的地址词典查找地址串中所包含的最大地址要素,若新查找的最大匹配地址比前一次查找的地址长,则当前处理级别设置为临时处理级别;否则将临时处理级别继续降低一级,并重复步骤4)的操作,当临时处理级别大于地名库指定的最低级别时,步骤4)结束,并继续下一步。

5)当前地址串为剔除最大匹配地址后余下的部分,若当前地址串结束,则将匹配信息添加到最终匹配结果中,若此时处理级别小于地名库指定的最低级别则将当前地址串重置为未剔除最大匹配地址之前的地址串。

6)将当前处理级别设置为地名库指定级别的下一个级别,并转到步骤2)。

7)得到最终匹配结果:若给定地址串完全匹配,则构建该地址串的标准地址,并从地名库中获取相关信息;若给定地址串不完全匹配,则给出最大匹配地址,及最大匹配地址对应的标准地址并获取相关信息,匹配结束。

注意,给定某地址串,根据该地址串查找最大地址要素,若最大匹配地址是查找出的地址要素的一部分,则为部分匹配;若最大匹配地址与查找出的地址要素完全相同则为完全匹配。最大匹配地址是指从某一位置开始的最长地址词。匹配过程中注意验证当前地址要素的父地址是否与前一次查找的地址要素相同,父地址与当前地址要素有可能相隔多个级别,最终匹配结果可能有多组。处理门牌号级别时需要根据用户给定门牌号信息,及门牌号索引表中记录的相关信息,利用插值原理将匹配结果定位到街道的两边。

3 实验结果及分析

基于以上理论分析,笔者在 Visual Studio 2005 环境下,采用大型关系数据库 Oracle10g 提供地名库的存储支持,在 MAPGIS7x 中实现了地理编码模块,该模块按照 MAPGIS7x 的结构分为三层,分别是:数据管理层、活动对象层和工具层。

数据管理层处在整个架构的最底层,它将提供的地址数据按照一定的规则进行处理,并创建地名库及索引,从而实现地址匹配及定位功能,所有功能接口都由该层完成。活动对象层是对数据管理层进行抽象封装,对外接口由该层提供。工具层主要通过接口调用,实现完成地理编码最基本的工具,主要有:地名库创建工具、地名库浏览编辑工具和地址匹配定位工具。

为了验证该方法的正确性及可行性,笔者选取了重庆市和杭州市,并分别建立地名库,进行地址匹配实验。表3列出了这两个城市建库及匹配的性能参数,图1是截取杭州市50条记录的匹配结果的一部分。系统硬件配置为:处理器 AMD AthlonTM 1.25 GHz,内存为 1.00 GB。

表3 地理编码实验数据

城市	级别数	记录数	建库耗时/min	读取索引耗时/s	匹配平均耗时/s
重庆	4	37 477	10	12	0.2
杭州	5	69 785	20	30	0.3

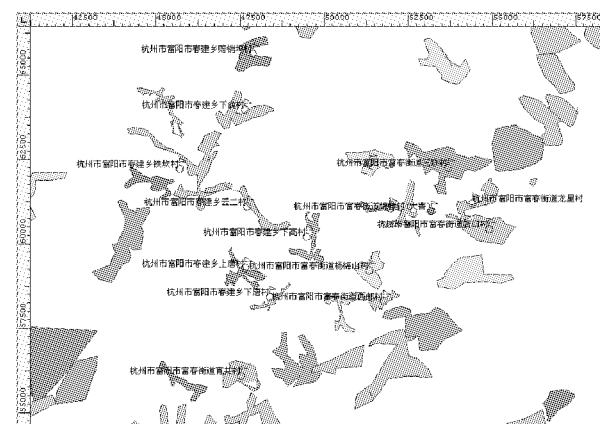


图1 杭州市匹配结果

实验数据显示,地名库创建耗时较多,且与数据量成正比,但是地名库只要创建好就可以直接使用,因此该时间可以接受。地址匹配平均耗时小于 0.5 s,时间性能参数说明本文提出的基于分级地名库的地理编码方法具有可行性,能较好

(下转第 1958 页)

小补集,从而提高了系统的总体效率。

可见,算法1利用OLAP查询之间的 $>$ 关系与 \Rightarrow 关系,充分利用已完成的查询来得到整个查询集的结果,并且中间计算最小补集与查询都可以并行在不同的处理机上运行,达到并行的效果。

3 实验结果与性能分析

为了测试优化算法的有效性,对算法进行了测试。测试所采用的数据集为TPC-R^[4],实验所采用的是若干台Intel Pentium IV 2.6 GHz, 512 MB 内存,运行Windows 2003 Server的PC机所搭建的并行处理网络。

实验1 随机选取了若干个OLAP查询,并比较随机调度与优化算法之间的总执行时间,其结果如图1所示。由图可以看出,当OLAP查询数量较少时,查询之间的 $>$ 关系与语义分解关系并不多,因而查询的执行顺序对于总的执行时间影响不大,优化算法由于要进行OLAP查询的预处理(计算OLAP查询之间的 $>$ 关系与语义分解关系)需要一定的CPU时间,因此可能优化后的总执行时间还比未优化时的查询执行时间要长;但在OLAP查询数量增加时,优化算法能识别出各个查询之间的公共查询并先执行得到结果,后续查询可以利用已有的结果得到,因此优化后的执行时间就少于未优化时的查询执行时间,并且这种效果随着OLAP查询数量的增多而愈加显著。

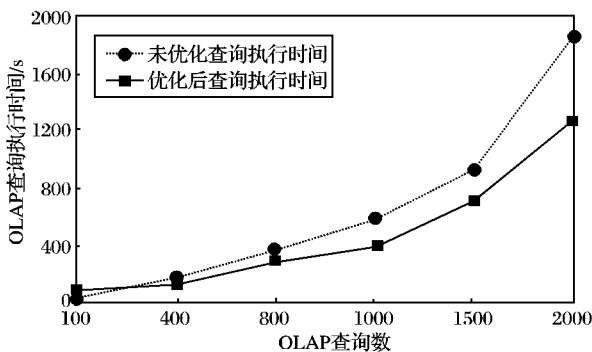


图1 基于语义分解的并行调度与随机调度执行时间比较

实验2 考查对于特定数量的OLAP查询,随机选取了1000个OLAP查询,考查处理机数目及补集阈值对于优化效果的影响,同时与未经优化的查询时间相比较,其结果如图2所示。由图可以看出,由于优化算法采用并行的方式来执行,因此处理机数目越多,优化效果越明显。同时,补集阈值的选取对于优化效果也有所影响,补集阈值越大,则能够选取更多的最小补集来确定语义分解关系与聚集关系,但同时最小补

集的计算时间也相应增加;补集阈值越小,最小补集的计算时间越少,但也越难于找到最优的最小补集,因此往往采用实验来确定一个经验值。就本实验中选取的1000个OLAP查询而言,如果处理机的数据为 n ,则补集阈值取 $n/4$ 到 $n/3$ 之间的整数,效果较好。

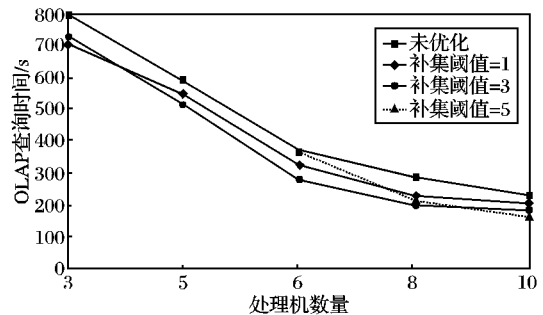


图2 用于并行的处理机数目与补集阈值对于查询优化的影响

4 结语

本文对OLAP查询进行形式化描述,定义了OLAP查询之间的聚集关系与分解关系,并采用并行计算OLAP查询补集的方式对OLAP查询进行整体优化,从而减少查询耗费的代价,改善了系统的性能。以后的研究方向将继续对模糊OLAP查询的匹配度进行研究,并对查询优化算法进行改进,以进一步提高模糊OLAP查询的效率。

参考文献:

- [1] SHIM J, SCHEUERMANN P, VINGRALEK R. Dynamic caching of query results for decision support system[C]// Proceedings of the 11th International Conference on Scientific and Statistical Database Management. Washington, DC: IEEE Computer Society, 1999: 254-263.
- [2] YANG J, KARLAPLEM K, LI Q. Algorithms for materialized view design in data warehousing environment[C]// Proceedings of the 23rd International Conference on Very Large Data Bases. San Francisco: Morgan Kaufmann Publishers, 1997: 136-145.
- [3] 薛永生, 黄震华, 段江娇, 等. 一种并行处理多维连接和聚集操作的有效方法[J]. 计算机研究与发展, 2004, 41(10): 1661-1669.
- [4] 杨科华. 提高联机分析处理(OLAP)性能若干关键技术的研究[D]. 南京: 东南大学, 2006.
- [5] Transaction Processing Performance Council TPC. TPC benchmarks H and R (decision support) [EB/OL]. [2009-08-06]. <http://www.tpc.org/>.

(上接第1955页)

地解决绝大多数非空间坐标地址的匹配问题。但是由于我国的门牌号分布并非完全有规律,使用地理编码技术定位出的点位的几何精度不高,因此它只适合使用在对空间数据精度要求较低的领域中。

参考文献:

- [1] 张铁燕, 翁敬农, 黄坚. 城市地理编码方法的探索与实践[C]// 中国地理信息系统协会第九届年会论文集. 杭州: 中国地理信息系统协会, 2005: 731-736.
- [2] 李军, 李琦, 毛东军, 等. 北京市地理编码数据库的研究[J]. 计算机工程与应用, 2004, 40(2): 1-3.
- [3] ZHANG XUEHU, MA HAOMING, LI QI. An address geocoding solution for Chinese cities[C]// Proceedings of Geoinformatics. [s.

l.]: International Society for Optical Engineering, 2006: 1-9.

- [4] 陈细谦, 迟忠先, 金妮. 城市地理编码系统应用与研究[J]. 计算机工程, 2004, 30(23): 50-52.
- [5] 朱建伟, 王泽民. 地理编码原理及其本地化解决方案[J]. 北京测绘, 2004(2): 24-27.
- [6] 孙亚夫, 陈文斌. 基于分词的地址匹配技术[C]// 中国地理信息系统协会第四次会员代表大会暨第十一届年会论文汇编. 北京: 中国地理信息系统协会, 2007: 114-125.
- [7] 孙茂松, 左正平, 黄昌宁. 汉语自动分词词典机制的实验研究[J]. 中文信息学报, 2000, 14(1): 1-6.
- [8] 郭会, 宋关福, 马柳青, 等. 地理编码系统设计与实现[J]. 计算机工程, 2009, 35(1): 295-299.