引文格式：应 申，李威阳，贺 彪，等. 基于城市地址树的地址文本匹配方法[J].地理信息世界,2017,24(6):81-86.

基于城市地址树的地址文本匹配方法

应 申[1,3]，李威阳[1,3]，贺 彪[2,4]，王 维[1,3]，赵朝彬[1,3]

(1. 武汉大学 资源与环境科学学院，湖北 武汉 430079；2. 深圳市规划和国土资源委员会，广东 深圳 518034；
3. 武汉大学 地理信息系统教育部重点实验室，湖北 武汉 430079；4. 深圳市数字城市工程研究中心，广东 深圳 518034)

基金项目：
国际自然科学基金项目(41671381)资助

作者简介：
应申（1979-），男，安徽界首人，教授，博士，主要从事三维地籍及三维地籍关键技术、CityGML建模、地址分词与匹配研究工作。
E-mail:
shy@whu.edu.cn

收稿日期：2017-07-03

【摘要】为了适应中文地址数据的复杂性，本文依据其中地址要素的层级关系，建立城市地址要素的树形模型，并提出基于地址树的文本自适应匹配方法，该方法根据地址数据中各部分地址信息匹配的节点评价选择最优匹配结果，通过单元最大长度匹配法获得地址树中与地址信息相匹配的节点，参照节点的层级关系构建相对独立的地址节，根据地址节中的地址信息计算权重因子，回溯评价返回最优匹配结果；本文采用深圳市518 948条建筑物地址数据构建城市地址树，在此基础上进行地址匹配试验，达到85.6%的匹配准确率，可应用于地址标准化和地址匹配流程中。

【关键词】城市地址树；地址匹配；地址数据；地址节
【中图分类号】P208 【文献标识码】A 【文章编号】1672-1586（2017）06-0081-06

Address Text Matching Method Based on City Address Tree

YING Shen[1,3], LI Weiyang[1,3], HE Biao[2,4], WANG Wei[1,3], ZHAO Chaobin[1,3]
(1. School of Resource and Environment Sciences, Wuhan University, Wuhan 430079, China; 2. Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Land and Resources, Shenzhen 518034, China; 3. Urban Planning, Land and Resources Commission of Shenzhen Municipality, Shenzhen 518034, China; 4. Shenzhen Research Center of Digital City Engineering, Shenzhen 518034, China)

Abstract: Based on the hierarchical relationship of Chinese address element in the tree model city address elements, text matching method is proposed in this paper based on the adaptive address tree, according to the evaluation data of node address information to achieve the optimal matching result. Through the matching method to obtain the maximum length of unit node address tree and address information matching the hierarchy reference node, this method establishes a relatively independent address section, the weighting factor is calculated according to the address information address section, backtracking returns the optimal matching evaluation results. Shenzhen City address tree is constructed with 518948 address records, based on this ,the address matching test achieved a good matching accuracy.

Key words: city address tree; address matching; address data; address section

0 引 言

随着地理信息技术的快速发展和广泛应用，不同行业类型的地址不断产生并录入数据库中，城市部门机构都存有海量与地址有关的信息，基本上是非结构化、非空间化的地址数据[1]，以字符串的形式进行存储，这些地址数据大多是非标准化并且缺乏空间坐标信息，无法有效服务于城市基础地理数据的平台和多源数据的集成，地址匹配是将待匹配的地址数据通过一定的匹配策略查找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[2]。

地址数据表达了城市的行政区划、各级街道以及门牌号的信息，不同于国外地址的规则化表达，收录和存储在基本地理数据库中的地址数据缺乏统一的组织结构，在没有明确权威规则的情况下往往依据当地习惯进行地址描述，经常出现冗余或缺省[3]，地址信息表达地址数据中各部分的信息，其中单个字符的差异也会导致表达地理实体的不同，是中文地址匹配研究中不可回避的问题。中文地址匹配的方法主要可以概括为模糊匹配、层级地址精确匹配和空间推理匹配[2]：模糊匹配查找地址数据库中与待匹配数据相近的信息，通过语义分析解决错别字、地址歧义和模糊匹配问题[3-4]，精度相对低并且匹配结果不准确；层级地址精确匹配是查找分词结果与地址数据库的精确匹配方法[5]，在标准化地址数据中匹配良好，复杂地址中精度较差；空间推理是在

地址解析和语义分析的基础上，采用地址中隐含的空间关系进行定位[2]，但是同一个位置在中文地址表达中存在多种形式，需要模型解析并理解推理。另外有许多学者从构建地址匹配模型识别要素特征类型[6-7]、地名地址词典辅助搜索匹配[8]、优化存储结构改进速率低和空间开销大的不足[9]等角度进行地址匹配研究。

在中文地址数据中行政区划涵盖各级街道，街道涵盖社区、门牌号，因此以树形结构模拟地址数据是可行的，其中有研究人员提出基于地址树模型的地址数据标准化方法[10-11]，优化地址树结构提高地址匹配准确率[12]等，上述方法是以涵盖空间信息的地址树模型为基础，需要大量已知地理坐标的标准化地址数据进行建立，而地址数据大多缺乏坐标范围而难以构建城市范围内的空间信息地址树。本文基于不含空间信息的城市地址树提出文本自适应的匹配方法，旨在根据地址数据的各部分地址信息匹配最优结果来适应地址数据的复杂性，引入单元最大长度匹配方法进行地址树的全局搜索或节点搜索，合并地址块的匹配结果和计算地址节的权重，通过地址树中节点的回溯过程评价选择最优的匹配节点，即地址树中与原地址数据匹配最优的地址数据。匹配过程中不依赖绝对阈值，根据原始地址数据中的信息进行查找、匹配、合并和评价流程，依照文本信息适应性实现地址匹配过程，匹配方法具有一定的通用性和适应性。

1 技术路线

1.1 城市地址树

根据国家地理信息公共服务平台地理实体与地名地址数据规范[13]（CH/Z9010-2011），地名地址的描述规则具有不同粒度，采用分段组合的方式进行描述，主要包括三大类要素，分别是行政区域、基本区域限定物和局部点位置。地址数据是按照地址要素的组合排列进行描述，依照地名地址数据规范地址要素所表达的空间范围是逐渐缩小的，可用多叉树的数据结构建设城市地址树，地址树中的节点表达地址要素，地址要素能表示节点在城市内一定范围的空间信息，如图1中的“宝安区”“大浪街道”等，随着地址树的深度不断递进，地址要素从城市行政区、街道（路）、小区（村）到门牌号逐渐变化，地址要素表达的空间信息的范围逐渐缩减和精细，树中父节点的空间范围包含子节点的空间范围，树中的叶子节点表示具体的建筑物单元或门牌号。图1为深圳市的部分地址树。

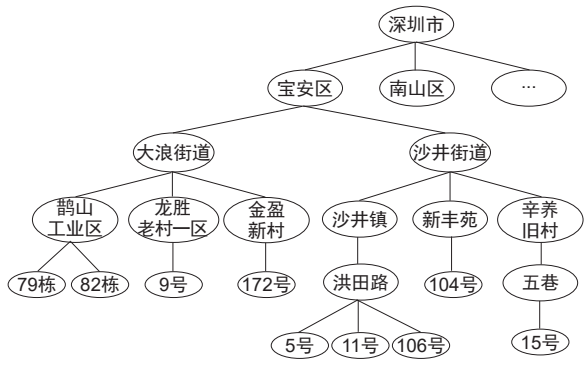

图1 深圳市地址树信息图

Fig. 1 Address tree of Shenzhen

1.2 全局搜索和节点搜索

在地址匹配过程中，认为连续的数字、符号，单个的中文字符为单元字符，可将地址数据看成是离散化字符单元的顺序排列。地址匹配过程中常常涉及到信息的查找，当从城市地址树中检索某条地址信息时，深度遍历和广度遍历查找的效率很慢，为提高检索的效率，建立树中地址要素的哈希表结构，以地址要素的第一个单元字符为索引，将相同索引值的地址要素存入集合当中。当检索未知地址数据时，获取该数据的第一个单元字符，定位到哈希表中单元字符映射的集合进行查找，极大地缩小遍历范围，即本文匹配过程中的全局搜索方法，通过构建地址树的单元字符哈希表结构，提高全局查找的效率。

当地址数据中一部分已经检索到树中匹配的地址要素，进而查找后方的地址信息时，优先选择启发式的广度优先遍历，即本文的节点搜索方法，是以已知的地址要素节点为父节点，搜索其子节点、孙节点中是否存在待检索的地址信息。由于已知的地址要素和未知的地址信息在地址数据中是相邻关系，可以映射到地址树中进行启发式的搜索，一般而言地址数据在描述时地址要素不会连续缺省，因此在广度优先遍历时只搜索其子、孙节点的范围。

2 地址匹配流程

《地名地址数据规范》规定地址要素按照一定的命名规范和习惯进行组织[13]，中文地址中实际上隐含着分隔符，如“街道”“区”“路”“大道”“村”“号”等地址要素中常见的分隔符，地址分隔符表示在地址

要素中普遍出现的、能够在地址数据中表示地址分隔的词。如地址数据“深圳市宝安区沙井镇洪田路106号602”简单分词之后，得到“深圳市宝安区”“沙井镇洪田路”“106号”“602”地址块的集合，以地址块的简短形式进行匹配，避免全局匹配时搜索效率低下，另外地址数据中经常出现遗漏或错字，地址块检索能减轻错误信息的干扰。

2.1 单元最大长度匹配法

本文在城市地址树的基础上提出了单元最大长度匹配法，通过地址块的字符单元数组多重精确匹配树中的地址要素，以地址块“深圳市宝安区”为例阐述单元最大长度匹配法。

1）地址块根据单元字符的定义（1.1中）离散化为单元字符的集合，如图2所示。

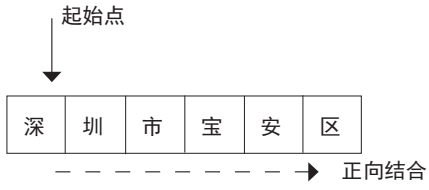

图2　地址块的离散化顺序排列

Fig. 2 Discretization order of address block

2）以单元字符集合的首字“深”为起始点，由于地址树中地址要素节点的单元字符长度大于1，单个单元字符不足以构成地址要素，因此取最小长度的两个单元字符“深圳”为当前搜索样本，进行全局搜索或者节点搜索（1.2中），转到3)或者4)。

3）若存在匹配的节点，表示在地址树中存在与搜索样本一致的地址要素，则从单元字符集合中取出下一个单元字符，追加到样本的尾部，即“深圳市”作为搜索样本，寻找最大长度的匹配结果。若仍存在匹配的节点，则继续增加样本的长度，直到不存在匹配的节点或者样本的长度达到最大值；找到该起始点下最大长度的搜索样本和匹配节点。

4）若不存在匹配的节点，则在样本尾部追加一个单元字符，若存在匹配的节点，则转到3），若不存在匹配的节点，则增加样本的长度，若直到该起始点遍历结束仍不存在匹配的节点，转到5）；若存在则转到3）。

5）起始点的位置向后迭代，移动一个字符单元长度的距离，如以“圳”为起始点；“圳市”为当前起始点的搜索样本，重复3），4），5）循环搜索过程。

单元最大长度匹配法区别于地址编码中常见的最大正向匹配算法，是以地址块为研究单元，字符单元集合为搜索数据对象，在城市地址树中寻找匹配地址块字符信息的地址要素。

2.2 地址块递归匹配

地址块“深圳市宝安区”经过初始的单元最大长度匹配流程，匹配到地址要素“深圳市”后，地址块中剩余待匹配的地址信息“宝安区”需递归进行匹配流程，其中“深圳市”匹配完成的地址要素节点可作为“宝安区”匹配流程的启发知识，通过节点搜索快速遍历能提高检索的效率，若节点搜索未能在子孙节点中查找到匹配节点，则通过全局搜索进行遍历，递归匹配地址块直到结束。具体流程如图3所示。

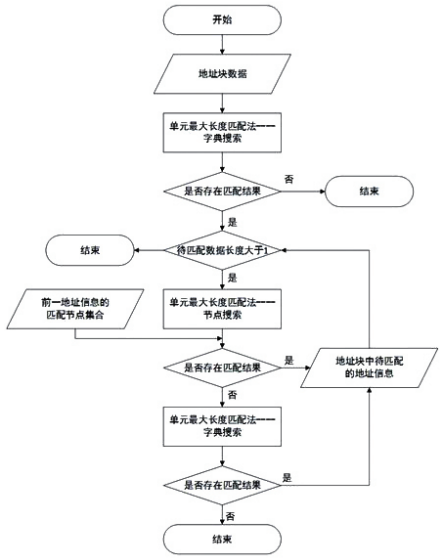

图3　地址块匹配流程图

Fig. 3 Address block matching flow chart

地址块递归查找城市地址树与其中地址信息相匹配的节点，匹配结束时地址块会产生单个或多个地址信息的匹配结果，本文定义最小的匹配结果单元为地址元，地址元中包含原始搜索地址字符串、匹配成功的字符串和地址树中的节点集合，连续排列的地址元构成地址块的匹配结果，地址元是地址数据匹配流程的基石。

相邻的地址元、地址块之间会存在地址树形式的层级关联，即相邻地址信息的匹配节点会存在树型的父子（孙）关系，子（孙）节点的匹配内容涵盖了父节点的匹配内容，是更加精确的地址描述，可以进行地址元和地址块的合并。如“深圳市”和“宝安区”分别成功匹配到地址树中的地址要素节点，节点间是父子关系，

两者在同一个地址块中，但分属不同的地址元，支持合并为子节点“宝安区”。

如图4中实验匹配到5个地址元，其中“宝安区”地址元是“深圳市”地址元的子类，同理“洪田路”是“沙井镇”的子类，地址元间的层级关系可以传导，因此“106号”是“沙井镇”的子类，地址元合并为“宝安区”和“106号”两个地址节。

图4 地址元层级关系

Fig 4 Hierarchical relationship of address unit

2.3 地址节权重和回溯评价

地址元、地址块合并之后形成地址节，结构如图5所示，地址节中包含独立的匹配节点集合和成功匹配的字符串信息，如图中地址节的字符串信息“深圳市宝安区”是地址元匹配结果的并集。

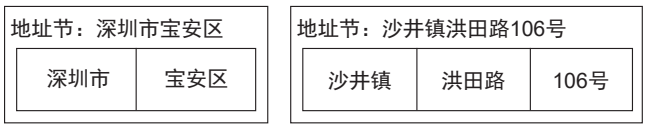

图5 地址节结构

Fig5 Address section structure

字符串信息的长度表明地址节匹配原地址数据的信息数量，表达在综合评价过程中的相对重要程度，字符串在原地址数据的索引位置表明地址节匹配的信息精细程度，索引位置越接近地址数据末尾，则越接近真实的匹配结果，索引位置是按照该地址节中最后地址元的字符信息在原地址数据的位置，因为有时地址数据表达不规范，会出现错别字或缺省字，地址元组合的字符信息无法精确匹配原地址数据，参照末尾地址元的索引位置代表整体地址节的索引位置。地址节权重因子表达中引入长度因素和位置因素，计算地址节在地址匹配评价中的重要性，计算过程如下：

假设原地址数据为D，其长度为D_l，地址节中字符串为S，其长度为S_l，在D中的索引位置为S_i，其中S_l，$S_i \leqslant D_l$。长度因素L和位置因素I的计算公式如式（1）、式（2）：

$$L=\frac{S_i}{D_l} \quad (1)$$

$$I=\frac{S_i}{D_l} \quad (2)$$

长度因素和位置因素与地址节在匹配时的权重值是正相关，而且值域均在(0,1)内，地址节的权重W计算公式如式（3），地址节的权重只依赖于匹配的字符串信息。

$$W=L\times I \quad (3)$$

城市地址树中的节点向上回溯时，构成了局部的地址数据，如图1中节点“鹊山工业区”回溯时则构成了地址数据“深圳市宝安区大浪街道鹊山工业区”，地址树中的节点不仅表示自身的特征信息，也涵盖了更广泛的地址描述信息。原始地址数据的地址元合并之后形成相对独立的地址节集合，地址节中的节点向上回溯时，可能会经过其他地址节中的匹配节点集合。图6中地址节中的节点“106号”向上回溯时经过上一地址节“宝安区”，表示“106号”节点中包含“深圳市宝安区”信息。

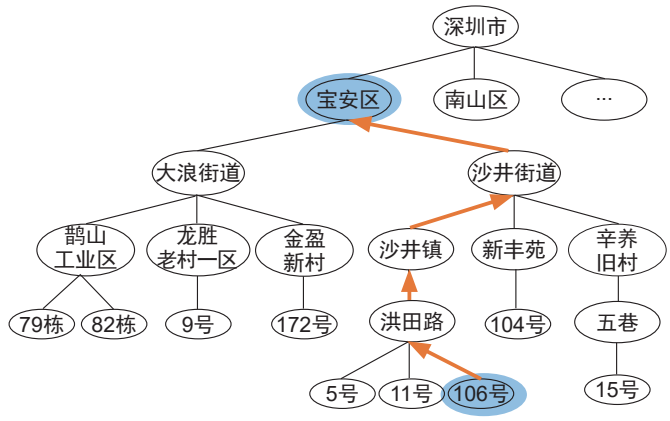

图6 地址树的节点回溯

Fig6 Node backtracking of address tree

当节点经过其他地址节时，表示节点成功匹配该地址节中的字符串信息，累计节点经过地址节的权重值，即是该节点匹配原地址数据的匹配度，匹配度最高的节点为最优的匹配结果，本示例数据中匹配度最高的节点为“106号”，回溯的地址信息为匹配最优的地址数据。

3 试验结果和分析

本文中使用的实验数据是由深圳市政府单位提供的深圳市建筑物地址普查数据，共有518 948条地址文本数据，在此基础上进行地址分词构建城市地址树和地址匹配的试验，城市地址树基本上覆盖深圳市的主要街道和村落，为了适应中文地址数据的复杂性，根据地址

树的特征自主编程实现基于地址信息文本自适应的匹配方法，单元最大长度匹配法查找地址数据中各部分地址信息在地址树中匹配的节点，让地址数据中局部正确的信息也能够在地址树中得到有效表达，根据层级关系适应性形成地址元和地址节，并计算地址节的权重和评价。本文采用未参与构建城市地址树的建筑普查数据进行验证，随机抽取其中500条地址数据进行检验，检验过程依赖于各部分地址信息在地址树中的匹配节点，因此地址节的回溯过程中存在与原地址数据匹配度最高的节点，因此判读地址数据的回溯中间过程，验证评价方法能否返回最优的匹配结果，并统计验证结果，其中有428条地址数据能成功匹配地址树中最优节点，准确率为85.6%，部分地址数据匹配成功的过程见表1。

表1　地址数据成功匹配流程

Tab1 The flow of address data matching successfully

原地址数据	地址块	节点回溯路径	匹配结果
龙岗区新三村东区五巷1号	龙岗区|新三村|东区|五巷|1号	龙岗区←布吉镇←新三村←东区←五巷	龙岗区布吉镇新三村东区五巷
福田区保税区新桂花村金大厦东侧	福田区|保税区|新桂花村|金大厦东侧	福田区←福田保税区←新桂花村	福田区福田保税区新桂花村
龙岗区低山北路1号	龙岗区|低山北路|1号	龙岗区←横岗←低山北路←1号	龙岗区横岗低山北路1号
南山区中山园路北同乐村117号	南山区|中山园路|北同乐村|117号	南山区←南头街道←中山园路←同乐村	南山区南头街道中山园路同乐村
南山区下白石三坊5号	南山区|下白石三坊5号	南山区←白石洲←下白石←三坊	南山区白石洲下白石三坊
龙岗区罗庚丘罗丰路44东边	龙岗区|罗庚丘罗丰路|44东边	龙岗区←坪山街道←竹坑社区←罗庚丘←罗丰路	龙岗区坪山街道竹坑社区罗庚丘罗丰路
宝安区公明街道红星社区第二工业区志峰厂	宝安区|公明街道|红星社区|第二工业区|志峰厂	宝安区←公明街道←红星社区	宝安区公明街道红星社区
罗湖区清水河村48-1栋西侧	罗湖区|清水河村|48-1栋|西侧	罗湖区←清水河街道←清水河村←61-1栋	罗湖区清水河街道清水河村61-1栋
龙岗区赐昌路8号五厂C栋北面	龙岗区|赐昌路|8号|五厂C栋|北面	龙岗区←横岗街道←赐昌路←8号	龙岗区横岗街道赐昌路8号

在地址数据和地址树的层级关系下，依赖地址数据中各部分地址信息匹配最优结果，本文匹配方法可以适应中文地址数据大多数的复杂性，允许地址数据中常见的信息缺省、冗余和部分错误情况，能应用于地址标准化和地址匹配流程。

表2是抽样中错误匹配结果的典型集合。

表2　地址数据匹配错误分析

Tab. 2 Analysis of address data matching error

原地址数据	正确匹配结果	错误匹配结果
宝安区松岗街道沙浦沙一村旧村37号	宝安区松岗街道沙浦社区沙浦	宝安区新安街道41区安乐一街旧村37号
宝安区公明街道合水口第二工业区B-29栋	宝安区公明街道合水口	宝安区公明街道经济发展总公司第二工业区
宝安区观澜街道桂花樟企路云之彩厂	宝安区观澜镇观澜街道樟企路	宝安区观澜街道桂花
宝安区沙井街道沙三村蚝三旧村九巷15号	宝安区沙井街道沙之社区蚝三旧村	宝安区沙井街道沙三村

地址数据中门牌号和地址通名[14]不能唯一表达城市内的地理实体造成匹配错误，门牌号如“39号”“1栋”，地址通名如“旧村”“第二工业区”，此类信息经常出现在地址数据尾部，是匹配深度的节点不可或缺的部分，但同时需要深入研究其信息特征并在地址数据中识别；城市地址树不完善和地址数据的不规范也会造成匹配错误。综上分析，规范地址树结构和地址数据描述[15]，匹配过程中识别门牌号和通名信息能够提高地址匹配的准确率。

4　结束语

地址信息中单个字符的差异会导致表达地理实体的不同，基于传统文本相似度计算和模糊匹配的逻辑难以保证匹配的准确度。为了适应中文地址数据的复杂性，本文根据城市地址树和地址数据的特征提出基于地址信息文本自适应的匹配方法，依赖各部分地址信息匹配的节点评价选择最优匹配结果，试验表明基于文本自适应的匹配方法能够达到较高的准确度，可以应用于地址标准化和地址匹配流程中；分析表明进一步提高地址匹配的准确率，需要在规范地址数据、地址树和识别门牌号、地址通名方面做进一步的研究。

参考文献

[1]　谭侃侃．基于规则的中文地址分词与匹配方法[D]．青岛：山东科技大学，2011

[2]　周海．基于条件随机场和空间推理的地理编码方法[D]．郑州：信息工程大学，2015.

[3]　臧英斐．基于语义分析的地址匹配研究[D]．重庆：重庆大学，2015.

[4]　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. 北京：中国地理信息与系统协会，2007.
[5] 叶海波. 城市地址编码的技术及应用[D].北京：中国石油大学，2009.
[6] 于滨. 面向经济普查项目需求的模糊中文地址匹配方法研究[D].长沙：中南大学，2010.
[7] 庄海东，张鸿恩. 基于规则的中文地址匹配系统[J].福建电脑，2013(9):130-132，146.
[8] 马照亭，李志刚，孙伟，等. 一种基于地址分词的自动地理编码算法[J].测绘通报，2011(2):59-62.
[9] 徐聪，张丰，杜震洪，等. 基于哈希和双数组trie树的多层次地址匹配算法[J].浙江大学学报：理学版，2014,41(2):217-222.
[10] 王勇，刘纪平，郭庆胜，等. 顾及位置关系的网络POI地址信息标准化处理方法[J].测绘学报，2016,45(5):623-630.
[11] Tian Q, Ren F, Hu T, et al. Using an Optimized Chinese Address Matching Method to Develop a Geocoding Service: A Case Study of Shenzhen, China[J]. ISPRS International Journal of Geo-Information, 2016,5(5):65.
[12] 亢孟军，杜清运，王明军. 地址树模型的中文地址提取方法[J].测绘学报，2015,44(1):99-107.
[13] 国家测绘地理信息局. CH/Z 9010-2011 地理实体与地名地址数据规范[S]. 北京：中国标准出版社，2011.
[14] 邬伦，刘磊，李浩然，等. 基于条件随机场的中文地名识别方法[J].武汉大学学报：信息科学版，2017，42(2):150-156.
[15] 李永恒. 澳门地理数据的整合进程——以街道门牌数据为例[J].地理信息世界，2013,20(1):87-91.

（上接第80页）

条件。只有建立完善的数据体系，才能开展信息化审计分析工作。审计工作需要测绘地理信息技术支撑，更要建立有关部门之间良好的自然资源数据共享与交换机制，深入研究自然资源审计评价指标体系，加快建设自然资源资产数据库和审计信息化平台，为全面开展领导干部自然资源资产离任审计工作提供强有力的决策支持服务。

参考文献

[1] 张峻. 浅谈在领导干部自然资源资产离任审计中信息化审计的运用[J].财会学习，2016(24):154-154.
[2] 戴斌. 自然资源资产离任审计的理论思考[J].中国市场，2016(29):120-123.
[3] 姚霖，侯冰. 我国自然资源资产负债表编制的问题与思考[J].国土资源情报，2015(7):27-30.
[4] 安徽省审计厅课题组. 对自然资源资产离任审计的几点认识[J].审计研究，2014(6):3-9.
[5] 蒋爱华. 智慧城市地理信息数据体系研究[J].地理空间信息，2016，14(3):15-17.
[6] 张衡，成毅，王晓理，等. 云GIS下智慧城市地理空间信息共享平台构建[J].地理信息世界，2016，23(03):71-76.
[7] 张宏亮，王秀华. 我国政府自然资产审计理论框架的构建[J].财会月刊，2007 (5):47-49.
[8] 安家鹏，程月晴，安广实. 自然资源资产离任审计评价指标体系构建[J].南京财经大学学报，2016(5):67-76.