中图分类号：F208　　密级：公开
UDC：　　本校编号：

兰州交通大学

硕士学位论文

论文题目：基于多源POI数据的匹配融合方法研究

研究生姓名：陈瑞　　学号：0211556

学校指导教师姓名：刘纪平　　职称：研究员

申请学位等级：理学硕士　　专业：地图学与地理信息系统

论文提交日期：2014年5月　　论文答辩日期：2014年5月

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含获得 兰州交通大学 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：陈瑞　　签字日期：2019 年 6 月 20 日

学位论文版权使用授权书

本学位论文作者完全了解 兰州交通大学 有关保留、使用学位论文的规定。特授权 兰州交通大学 可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：陈瑞　　导师签名：刘君平

签字日期：2019 年 6 月 20 日　　签字日期：2019 年 6 月 20 日

硕 士 学 位 论 文

基于多源 POI 数据的匹配融合方法研究

Study on the method of matching and fusion based on the multi-source POI data

作 者 姓 名：陈 瑞

学科、专业：地图学与地理信息系统

学 号：0211556

指 导 教 师：刘纪平 研究员

完 成 日 期：2014 年 5 月

兰 州 交 通 大 学

Lanzhou Jiaotong University

摘　要

随着早期 Google Earth、互联网瓦片地图带给人们的新鲜感逐渐下降，电子地图仅满足浏览需求或者作为背景的时代已经过去，人们对地理位置服务提出了更高的需求，地图也亟需融入到大众的日常生活中。

POI 数据是网络电子地图的最重要内容，也是互联网位置服务的立足之本。由于互联网上的 POI 数据来源不一，采集与处理过程各不相同，从而导致了这些数据在空间位置、属性信息以及丰富程度上存在着一定的差异，因此如何有效地消除数据间的不一致性，并把它们组织成一套内容准确、可供用户使用的数据成为了当前研究的热点，本文针对目前多源 POI 数据融合过程中存在着效率低下且精度难以保障等问题，首先对当前国内外相关部门与互联网电子地图网站各自的 POI 数据分类体系进行了分析和研究，对当前存在的不同分类体系进行了融合处理，最后提出了一种基于网格划分自动提取控制点的多源 POI 位置纠正方法，有效地提高了多源 POI 空间位置融合的精度和效率。论文具体内容有几下几点：

(1)首先对互联网多源数据产生的背景及原因进行了分析总结，然后针对多源数据融合技术的国内外研究状况做了论述和对比。

(2)重点介绍了多源空间数据融合的几种方法，包括基于坐标几何位置的变换模型和地理实体的匹配过程，基于分类分级和属性编码方式以及文本相似度的计算方法，对方法的特点和不足做了分析与研究。

(3)介绍了一种互联网地图 POI 数据的自动提取技术，通过研究互联网 POI 数据的特点和 Html 网页内容的组织形式，编程实现了 POI 的自动解析与提取；在研究国内外大型地图网站 POI 兴趣点分类体系的基础上，结合其优缺点建立了一套较为完善的 POI 分类体系。

(4)针对目前多源数据融合过程中使用传统控制点选取方法存在着效率低下且精度难以保障等问题，提出了一种基于网格划分自动提取控制点的多源 POI 位置纠正方法，采用中文语义匹配的方式自动获取每个网格内的控制点与精度检核点，通过多项式变换与粗差剔除过程计算出相应网格内的坐标变换系数，实现了不同源数据在空间位置上的“吻合”，最后对计算结果进行了精度评价。

(5)最后以上海市售票点数据为例做了相关实验，实验结果达到了预期的效果，具有较高的准确度和效率。

关键词：POI；多源数据融合；分类体系；语义匹配；位置纠正
论文类型：应用研究

Abstract

As early times with Googl-Earth and internet tile map, the freshness which above brought to people has gradually decreased, the era has passed as electronic map browsing only meet people's needs or as the background, now people put forward higher demand for location service, the map also an urgent need to integrate into to the public's daily life.

POI data is the most important content of network electronic map,and also is foundation of Internation Location Based Service. Since POI data source on the internet has many different origins, the otherness of acquisition and processing eventually led to the data exist certain differences in spatial location, attribute information and richness of data, so how to eliminate the inconsistency effectively between them and organise them into a set with accurate content and available data has become a research focus. As the low efficiency and uncertain accuracy in multi-source data fusion process, this paper offered a kind of multi-source POI position correction method, which based on control points automatic-extraction by grid division, improve the accuracy and efficiency of multi-source spatial position POI fused effectively.

1) Firstly, analyzed and summarized the background and reasons on the causes of internet multi-source data,and then discussed and compared technology research status both here and abroad of multi-source data.

2) Mainly introduced several methods of multi-source spatial data fusion, including the matching process transformation model and geographic entity based on geometric coordinates, classification and attribute coding and text similarity computing method,and analysised and researched on the characteristics and shortcomings of the methods.

3) Introduces an internet map POI data automatically extracting method, through the study on characteristics of POI data and internet webpage http content organization forms, program to realize automatic analysis and storage technology of POI; studying the research POI interest points classification system of large scale map website in domestic and foreign , with its advantages and disadvantages to establish a set of more complete POI classification system.

4) As the low efficiency and uncertain accuracy in multi-source data fusion process using traditional control points selection methods, offered a kind of multi-source POI position correction method, which based on control points automatic-extraction by grid division, using the Chinese semantic matching method to automatic obtain control points and check points within each grid, calculated the coordinate transformation parameter of corresponding grid with the polynomial transform and error elimination, finally, the accuracy of the calculation results were evaluated.

5) Finally, based on the ticket data of Shanghai as an example to do the experiment, the experimental results achieved the expected results and have a high degree of accuracy and efficiency.

Key Words：POI；Multi-Source Data Fusion；Classification System；Semantic Matching；Position Correction

目　　录

1 引言

1.1 研究背景

随着社会经济、通信技术和互联网的高速发展，人们对身边位置搜索服务的需求越来与大，在美国最大的社交网络服务网站 Fackbook 上，用户可以快速找到离自己最近的肯德基快餐店，同时该快餐店近期推出的套餐和优惠活动也会自动推送到用户面前，位置服务的应用在这里得到了很好的体现。在国内，位置服务产业也发展势头猛进，包括腾讯、新浪等一些大型门户网站在其旗下产品中也相继推出了位置搜索服务，用户通过网络就能够发现附近的其他用户，实现了空间位置的网络共享。据相关机构调查显示，过去几年中国的 LBS（Location Based Service）市场规模呈现大幅度增长的趋势，预计到 2014 年市场规模将突破百亿。随着市场规模的逐渐扩大，LBS 成为了当今互联网行业中的巨大热点，越来越多的用户也将享受到地里位置服务带来的便利。

POI（Point Of Interest，兴趣点）是地图服务重要的矢量化形式表达方式，也是地图最鲜活的“血液”，与面向公众的基于位置服务密切相关，它代表一类真实地理实体的地理空间数据。POI 主要指人们日常生活中经常遇到的地理场所，如学校、医院、宾馆、餐饮、银行、景点和标志性建筑物等，其在智能交通、应急指挥、公共安全、物流管理、电子商务以及其它各类 LBS 服务领域发挥着重要作用。准确性、现势性、完整性是评价 POI 数据质量的重要指标，也是影响地理位置服务可用性的重要因素，当前互联网上存在着海量的 POI 数据资源，由于不同地图厂商的数据采集技术和分类体系各不相同，再加上根据各自的需求和目的对数据进行了相应的变换处理，最终导致了这些数据在空间位置、属性信息以及丰富程度上存在着一定的差异，如何有效地避免和消除这些差异，以更快速度、更低成本丰富 POI 数据内容和提高数据质量，是空间信息领域相关研究者和服务商持续关注的热点。

多源数据融合的过程就是保持多源数据一致性的过程，从广义上讲，多源数据融合是针对数据的空间几何位置、属性信息、时间属性进行一致性融合；从狭义上讲，多源数据融合是从不同的数据精度、比例尺、数据模型、拓扑关系、语义匹配这几个方面来研究。作为一种特殊的点状地理实体，如何从不同的网络地图上获取需要的 POI 数据，利用多源数据融合技术实现数据融合是本文研究的重点。

1.2 国内外研究现状

从上世纪八十年代开始，中国地理信息产业开始飞速发展，基于地理信息系统的各种应用与需求不断加大，各种地理信息内容分类体系也随之应用而生。目前国内现行的

地理信息分类体系大概可分为三个层次[40]：第一个层次是基于国家颁布的相关标准和规范，其中有《基础地理信息要素分类与代码》（GB/T 13923-2006），《1:500,1:1000，1:2000 地形图要素分类与代码》（GB 14804-93），《地名分类与类别代码编制规则》（GB/T 18521-2001），《国土基础信息数据分类与代码》（GB/T 13923-92）；第二个层次是行业相关标准，其中有《城市市政综合监管信息系统管理部件和事件分类与编码》（CJ/T 214），《城市基础地理信息系统 1:500,1:1000,1:2000 地形图要素分类与代码》（CJJ100-2004）；第三个层次即专用分类体系，是指基于某一个特定的地理信息系统所形成的地理信息分类编码，专用分类体系下的内容数量非常丰富，涉及到地理信息系统的各个应用领域，地理信息位置服务中的 POI 数据分类就属于这一层次。

在地理信息科学相关问题研究领域，针对多源空间数据融合技术的解释是通过保留准确数据、剔除冗余数据的方式，确保数据的一致性原则。目前数据融合方法主要包括：基于非空间属性的数据融合、基于空间几何位置的数据融合、基于本体论的数据融合。

（1）非空间属性数据融合即语义融合，是把同名实体各自的文本属性信息融合在一起，Cobb 等(1998)研究基于属性值的语义相似度和线性要素的几何相似度相结合的方法进行要素的匹配并进行融合[1]。张桥平等(2002)对于点状要素，通过点结构相似度结合同名点匹配距离阈值确定设计了同名点匹配算法[2]。章莉萍(2008)以城市居民地为研究对象，基于语义对照、多尺度和地形图精度，提出了居民地空间目标多尺度匹配的新方法[3]。唐文静(2009)等提出了多评价因素的点状要素融合变换算法[4]。

（2）在基于空间位置的空间数据融合方面，Doytsher 等(2001)提出了基于线状要素的矢量数据合并框架[5]。Hoseok Kang 等(2004)提出根据几何和拓扑一致性的矢量数据融合模型[6]。郭庆胜、丁虹(2004)通过研究多尺度地理空间方向关系给出了相似性的描述与计算方法[7]。ASHOK S 等(2004)通过采用图论结合基于要素的方式从不同源数据中提取出共同对象，但对于结合要素语义信息的提取方面考虑不足[8]。Deng S 等(2006)将最小二乘原理运用到矢量数据融合，调整了合并后的空间位置，取得了较好的融合效果[9]。Song(2006)等采用 snake 算法用于 TIGER 道路数据进行了数据融合试验[10]。

Saalfeld 最先提出了三角剖分目标同名点的图形数据合并方法。该方法通过将待处理图和标准图上的同名点同时进行 Delaunay 三角网划分，在每个三角形内利用坐标变换公式将三个顶点建立起一定的转换关系，则落入三角形内的其他点群根据这一转换关系同时进行坐标转换。但由于同一个三角形的顶点同时跨越多个三角形，每个三角形都有不同的偏移量，势必会造成扭曲现象的出现，影响了计算精度[11,12]。张桥平将拓扑关系引入了匹配算法。通过确定未匹配点与已调整点是否互相影响，再计算影响值的大小来确定位置调整关系。

C Beeri 等提出了基于空间位置的一些空间数据融合技术：片面最近邻居连接方法(one-sided nearest-neighbor join method)在商业地理信息系统中的应用较为常见[13]。假如一个对象 A，若在附近所有对象集中与 A 距离最近的对象为 B，我们就称对象 B 为对象 A 到的对应对象，所有对应对象可组成一个融合集。但是要注意 A 与 B 对应连接和 B 与 A 对应连接的结果集有可能是不同的，即片面最近邻居连接算法并不具有对称性。C Beeri 在此基础上做了一定程度上的改进，提出了单一融合集的概念，即使对象 A 在所有数据集中都找不到与之相匹配的对象也会自动生成一个单一集合，此时该集合中只有对象 A 作为融合集的唯一元素。集合中的要素个数越少或是某一集合被其它集合所覆盖时该方法融合效果最好。

相互最近方法(mutually-nearest method)在片面最近邻居连接方法的基础上又进行了改进。有时会出现一个对象同时有几个最近邻居的情况，这样称之为相互最近对象。C Berri 把互相最近的每两个对象都当做一个融合集，这样就生成了不同的几个二元融合集，同时还提出了计算融合集可信度的数学方法，通过可信度计算进一步过滤了误差较大的融合集。相比片面最近邻居连接方法，相互最近法的优点在于处理两个覆盖在一个的集合时会表现的更好。

概率方法(probabilistic method)主要通过计算对象之间的距离来计算融合集中的对象是否是对应对象的概率，C Berri 给出了概率计算公式，并指出影响计算结果的几个参数。总的来说，当两个集合的覆盖度较小时，概率方法并不能得到准确的计算结果，因此该方法适用于集合之间覆盖度比较的大的情况，融合集中每一个对象的概率值之和应该等于 1。

标准化权重方法(normalized—weights method)在概率方法的基础上又做了进一步改进，它使得在两个集合覆盖度较小时也能产生较好的效果。该方法引入了权重分配的思想，将相应的权重值赋给不同的融合集，再通过经典迭代算法对初始化权重进行标准处理，形成不同融合集之间的相互作用关系。标准化权重方法在集合之间覆盖度不大或者很小的情况下会取得较好的计算结果，不过由于单一融合集不能进行标准化处理，所以当集合之间覆盖度较大时仍需选用概率方法进行计算[14]。

（3）本体论(Ontology)一词是由 17 世纪的德国经院学者郭克兰纽(Goclenius,1547-1628)首先使用的，是指一切存在事物的最真实本性，包括事物存在的规律和表现形式，可以作为各领域研究的一种方法和工具，同时本体论又是一种科学思想和理论。

从上个世纪 90 年代末开始，本体论的思想开始在地理信息科学研究中被提出，地理本体的概念也相机产生。地理本体领域的研究主要有：地理本体的构成 内容、地理

本体的构建方式、地理本体的形式化表达、地理本体的集成方式、地理本体的应用领域等。在基于本体的空间数据融合研究方面，Frederico 等(2002)将本体引入空间数据融合，使得影像与矢量数据融合后的结果既具有面对对象的特性，也具有语义特征[15]。Shahram 等(2002)将 Multi-Agent 系统引入地理数据融合方面做了一些尝试[16]。彭煜玮(2007)等将基于本体的技术运用到空间数据融合，取得了一些成果。郭黎(2008)等提出了尺寸与形状相近情况下匹配线状与面状目标相似度的计算模型，建立了同名实体匹配识别、几何与属性融合的框架[17]。

1.3 论文研究内容

目前互联网上的 POI 数据都包含着三个重要属性：名称、位置和属性信息，分别表示了该兴趣点叫什么、在哪里、是什么[18]，因此应该从这三个方面进行 POI 融合研究，先基于目前国内地图网站的现有兴趣点分类成果建立一套较完整的分类体系，再根据名称信息进行自动分类，最后采用数学变换模型对不同类别的多源 POI 数据依次进行坐标变换，进而基于空间位置实现多源 POI 的融合。

1.3.1 相关技术与方法研究

对多源空间数据融合的相关技术和方法做了分析与研究，总结了目前几何位置融合技术中坐标系统的变换原理和几种坐标变换模型，并分析了地理实体匹配的三种基本方法；针对属性信息融合技术中不同来源数据的分类分级方式和属性编码进行相关研究，介绍了两种常用的属性信息相似度计算方法。

1.3.2 基于 HtmlParser 的 POI 自动提取

由于互联网上的 POI 数据资源数量相当庞大，并且作为 POI 数据载体的电子地图网站也有不少，因此只靠人工手动去采集 POI 是相当费力费时的，本文提出了一种基于 Java 语言编写用于分析和提取网页内容的辅助工具 HtmlParser 来实现互联网 POI 数据的自动提取，可为本文实验阶段提供地图网站最新的 POI 数据资源。

1.3.3 多源异构 POI 分类体系的融合

既然 POI 是用户感兴趣的点，那么该点是什么、属于什么类别便成为了我们当前研究的重点，对当前兴趣点进行合理、科学地分类，有助于保证数据的实用性和合理性，能很大程度上提高数据融合的质量。

本文首先从于五个方面总结了多源 POI 数据的特点，分析归纳了数据信息分类的指导思想与基本原则，研究了当前国内外相关部门和一些大型地图网站针对各自地理信息

数据和 POI 所形成的数据分类体系，基于以上众多不同分类体系融合了一套较为完整 POI 分类体系。

1.3.4 无人工干预的坐标位置纠正

多源 POI 位置纠正是多源空间数据融合的一个重要过程，无论是线性方程法还是多项式变换法，都需要利用公共点来实现先整体后局部的几何纠正过程，由于 POI 数据量较大，控制点的数量以及点位分布选取在很大程度上影响着纠正精度和效率，如果选取不合理，在计算法方程系数矩阵时平差处理会呈现不良状态，导致地图局部变形，产生纠正精度不高的问题。由于待处理地理目标范围较大时如何自动选取控制点存在着不小的难度，本文按照一定的地理网格间隔将待处理 POI 地理范围拆分为 M*N 个单元网格，每个单元网格我们称之为一个纠正单元，在每个纠正单元内再进行子单元格拆分，利用中文语义匹配的方法自动获取每个子单元内不同数据源的同名同址点作为控制点与检核点，通过多项式变换与粗差剔除过程计算出相应纠正单元内的坐标变换系数，实现 POI 位置纠正。

1.4 本文的结构安排

论文分为五章，具体安排如下：

第一章，介绍了本文的研究背景、国内外研究现状、主要的研究内容和结构组织安排。

第二章，对常见的多源空间矢量数据融合技术做了介绍，包括几何位置的融合与属性信息的融合。几何位置融合技术介绍了坐标系转换方法、实体的匹配方法、坐标的转换方法；属性信息融合技术介绍了属性数据的分类分级方法和属性编码方法、属性数据的相似度计算方法。

第三章，总结了 POI 数据获取的几种方法以及如何利用网络爬虫技术自动获取地图网站 POI 数据，并基于 POI 数据的特征和各个电子地图网站的不同分类标准构建了一套新的 POI 分类体系。

第四章，本章是本文的核心章节，在目前多源数据融合研究的基础上提出了基于网格划分并自动提取控制点的多源 POI 位置纠正方法，并做出了实验验证，证明网格划分目标范围的方法提高了坐标变换计算的准确性，控制点的自动提取方法提升了算法应用的效率。

第五章，对论文内容进行了总结，对下一步工作内容进行了展望。

2 多源空间数据融合基本方法

2.1 几何位置的融合

在数据采集过程中由于采集部门、采集标准规范和比例尺不同，业务人员的操作能力参差不齐，数据后续规范处理的方法、数据集更新程度等差异，同一地区的不同来源数据集往往存在着一定程度上的几何位置差异[19]。为了消除数据的差异性与不一致性并让更多的用户加以使用，需要对多源数据的几何位置融合进行深入研究。

2.1.1 坐标系换算

2.1.1.1 欧勒角与旋转矩阵

两个直角坐标系进行相互变换时产生的旋转角叫做欧勒角，假设从直角坐标系（X_1，Y_1）转换为直角坐标系（X_2，Y_2），坐标系间的欧勒角为 θ 、尺度系数为 γ，坐标系（X_2，Y_2）的原点坐标是坐标系（X_1，Y_1）内的点（X_0，Y_0）（如图 2.1 所示）。那么根据两套坐标系可以计算出四个转换参数，从而完成坐标变换。

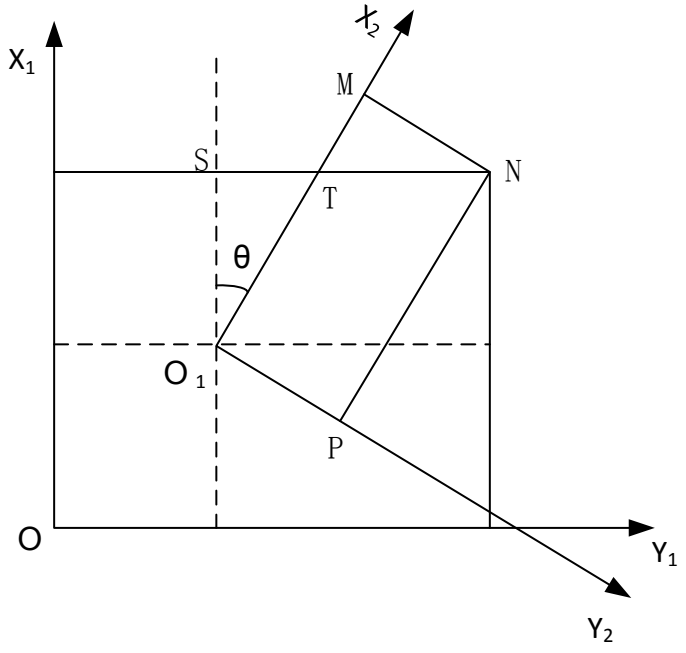


图 2.1 坐标系转换示意图

根据图中直角三角形 O_1ST 和直角三角形 TMN 的特性可计算得到两个坐标系之间的转换关系：

$$\begin{cases} x_2 = \gamma\big((y_1 - y_0)sin\theta + (x_1 - x_0)cos\theta\big) \\ y_2 = \gamma\big((y_1 - y_0)cos\theta - (x_1 - x_0)sin\theta\big) \end{cases} \tag{2.1}$$

将 2.1 式变形得到：

$$\begin{cases} x_1 = x_0 + \frac{1}{\gamma}x_2cos\theta - \frac{1}{\gamma}y_2sin\theta \\ y_1 = y_0 + \frac{1}{\gamma}x_2sin\theta + \frac{1}{\gamma}y_2cos\theta \end{cases} \tag{2.2}$$

用矩阵将上面两式分别表示为：

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \gamma \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} x_1 - x_0 \\ y_1 - y_0 \end{bmatrix} \tag{2.3}$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \frac{1}{\gamma} \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} \tag{2.4}$$

当旋转角度 θ 比较小时，近似的可以看作：

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \gamma \begin{bmatrix} 1 & \theta \\ -\theta & 1 \end{bmatrix} \begin{bmatrix} x_1 - x_0 \\ y_1 - y_0 \end{bmatrix} \tag{2.5}$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \frac{1}{\gamma} \begin{bmatrix} 1 & -\theta \\ \theta & 1 \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} \tag{2.6}$$

因此 2.1、2.2 式可表示为：

$$\begin{cases} x_2 = \gamma\big((y_1 - y_0)\theta + (x_1 - x_0)\theta\big) \\ y_2 = \gamma\big((y_1 - y_0)\theta - (x_1 - x_0)\theta\big) \end{cases} \tag{2.7}$$

$$\begin{cases} x_1 = x_0 + \frac{x_2}{\gamma} - \frac{\theta}{\gamma}y_2 \\ y_1 = y_0 + \frac{y_2}{\gamma} + \frac{\theta}{\gamma}x_2 \end{cases} \tag{2.8}$$

2.1.1.2 不同空间直角坐标系转换

对于两个空间直角坐标系进行坐标换算时既有平移、缩放以及旋转的情况，期间会产生 7 个变换参数，分别是 3 个平移参数、3 个旋转参数以及 1 个变换尺度参数，如图 2.2 所示，坐标变换公式为：

$$\begin{bmatrix} X_2 \\ Y_2 \\ Z_2 \end{bmatrix} = (1+m)\begin{bmatrix} 1 & \varepsilon_Z & -\varepsilon_Y \\ -\varepsilon_Z & 1 & \varepsilon_X \\ \varepsilon_Y & -\varepsilon_X & 1 \end{bmatrix}\begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \end{bmatrix} + \begin{bmatrix} \Delta X_0 \\ \Delta Y_0 \\ \Delta Z_0 \end{bmatrix} \tag{2.9}$$

上式中 ΔX_0，ΔY_0，ΔZ_0是坐标平移参数；ε_X，ε_Y，ε_Z是坐标旋转参数，m是变换尺度参数。

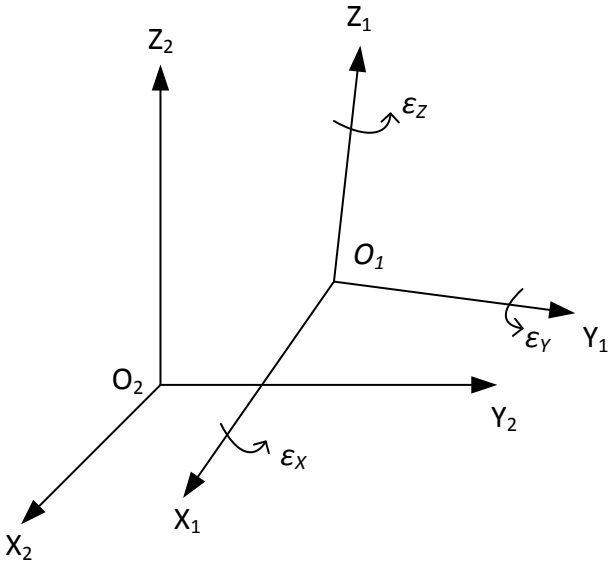


图 2.2

2.9 式表示的计算模型反映了两个空间直角坐标系的变换过程，该过程的求解需要计算出参数 ΔX_0，ΔY_0，ΔZ_0，ε_X，ε_Y，ε_Z，m。这种情况下的公共点个数应该不少于 3 个，若公共点多于 3 个时可依据最小二拟合法计算出以上 7 个参数的值。

令$a_1 = 1 + m$，$a_2 = a_1\varepsilon_X$，$a_3 = a_1\varepsilon_Y$，$a_4 = a_1\varepsilon_Z$，2.9 式可表示为：

$$\begin{bmatrix} X_2 \\ Y_2 \\ Z_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & X_1 & 0 & -Z_1 & Y_1 \\ 0 & 1 & 0 & Y_1 & Z_1 & 0 & -X_1 \\ 0 & 0 & 1 & Z_1 & -Y_1 & X_1 & 0 \end{bmatrix} \begin{bmatrix} \Delta X_0 \\ \Delta Y_0 \\ \Delta Z_0 \\ \mathrm{a}_1 \\ \mathrm{a}_2 \\ \mathrm{a}_3 \\ \mathrm{a}_4 \end{bmatrix} \tag{2.10}$$

取

$$\begin{bmatrix} V_{X2} \\ V_{Y2} \\ V_{Z2} \end{bmatrix} = \begin{bmatrix} X_2 \\ Y_2 \\ Z_2 \end{bmatrix}_{\text{以知值}} - \begin{bmatrix} X_2 \\ Y_2 \\ Z_2 \end{bmatrix}_{\text{转换值}} \tag{2.11}$$

误差方程可表示为如下形式：

$$\begin{bmatrix} V_{X2} \\ V_{Y2} \\ V_{Z2} \end{bmatrix} = -\begin{bmatrix} 1 & 0 & 0 & X_1 & 0 & -Z_1 & Y_1 \\ 0 & 1 & 0 & Y_1 & Z_1 & 0 & -X_1 \\ 0 & 0 & 1 & Z_1 & -Y_1 & X_1 & 0 \end{bmatrix} \begin{bmatrix} \Delta X_0 \\ \Delta Y_0 \\ \Delta Z_0 \\ \mathrm{a}_1 \\ \mathrm{a}_2 \\ \mathrm{a}_3 \\ \mathrm{a}_4 \end{bmatrix} + \begin{bmatrix} X_2 \\ Y_2 \\ Z_2 \end{bmatrix}_{\text{以知值}} \tag{2.12}$$

改写成矩阵形式：$V = B \cdot \delta X + L$

$\delta X = (\Delta X_0, \Delta Y_0, \Delta Z_0, \mathrm{a}_1, \mathrm{a}_2, \mathrm{a}_3, \mathrm{a}_4)^T$ 为待求的转换参数向量，$V = (V_{X2}, V_{Y2}, V_{Y2})^T$为改正数向量，$L = (X_2, Y_2, Z_2)^T_{\text{已知值}}$，$B$为系数阵。

依据最小二乘法$V^T PV = min$原则，列出的法方程为：

$$B^T PB\delta X + B^T PL = 0$$

其解为：

$$\delta X = -(B^T \mathrm{P}B)^{-1} B^T \mathrm{P}B$$

由δX可进一步求得：

$$m = \mathrm{a}_1 - 1,\ \varepsilon_X = \frac{\mathrm{a}_2}{\mathrm{a}_1},\ \varepsilon_Y = \frac{\mathrm{a}_3}{\mathrm{a}_1},\ \varepsilon_Z = \frac{\mathrm{a}_4}{\mathrm{a}_1}$$

因为在公共点的坐标中存在一定误差，转换参数的求解会受到影响，公共点的个数多少与点位的分布状况很大程度上影响着转换参数的精度，因此在计算过程中选取公共点的数量要尽可能的多，公共点分布的位置也需尽可能的均匀，这样才能求得准确的转换参数。

2.9 式为相似变换模型，公共点个数在 3 个以上时求解转换参数会存在多余观测，在实际工作中通常需要保持所有已知点的坐标不发生改变，但在坐标变换过程中由于公共点存在着误差会使计算结果与已知值不完全一致，所以通过配置法来解决这个问题，把公共点的计算结果当做已知值，然后在对剩余的点进行相应配置，最后计算坐标转换。具体方法是：

（1）公共点转换值的改正值为已知值减去转换值，公共点的坐标采用已知值。

（2）通过配置法计算非公共点转换值的改正值为：

$$V' = \frac{\sum_1^n P_i V_i}{\sum_1^n P_i}$$

上式中n表示公共点的个数，P表示权重值，通过计算公共点与非公共点的距离（S_i）来决定权值，一般取$P_i = \frac{1}{{S_i}^2}$。

2.1.1.3 不同大地坐标系换算

当两个不同大地坐标系进行坐标换算时除了有平移、旋转和变换尺度参数外，还存在地球椭球元素变换参数，因此有 3 个平移参数、3 个旋转参数、一个变换尺度参数和 2 个椭球变换参数，共 9 个变换参数。不同大地坐标系的变换公式推导如下：

空间直角坐标与大地坐标可用如下公式表示：

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} (N+H)cosBcosL \\ (N+H)cosBsinL \\ [(1-e^2)+H]sinB \end{bmatrix} \tag{2.13}$$

取全微分并经过一系列计算：

$$\begin{bmatrix} \mathrm{d}L \\ \mathrm{d}B \\ \mathrm{d}H \end{bmatrix} = \begin{bmatrix} \frac{sinL}{(N+H)cosB}\rho'' & \frac{cosL}{(N+H)cosB}\rho'' & 0 \\ -\frac{sinBcosL}{M+H}\rho'' & -\frac{sinBsinL}{M+H}\rho'' & \frac{cosB}{M+H}\rho'' \\ cosBcosL & cosBsinL & sinB \end{bmatrix} \begin{bmatrix} \Delta X_0 \\ \Delta Y_0 \\ \Delta Z_0 \end{bmatrix} +$$

$$\begin{bmatrix} tanBcosL & tanBsinL & -1 \\ -sinL & cosL & 0 \\ -\frac{Ne^2sinBcosBsinL}{\rho''} & \frac{Ne^2sinBcosBcosL}{\rho''} & 0 \end{bmatrix} \begin{bmatrix} \varepsilon_X \\ \varepsilon_Y \\ \varepsilon_Z \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{N}{(M+H)}e^2sinBcosB\rho'' \\ N(1-e^2sin^2B)+H \end{bmatrix} \cdot$$

$$m \begin{bmatrix} 0 & 0 \\ \frac{N}{M+H}e^2sinBcosB\rho'' & \frac{M(2-e^2sin^2B)}{(M+H)(1-\alpha)}sinBcosB\rho'' \\ -\frac{N}{\alpha}(1-e^2sin^2B) & \frac{M}{1-\alpha}(1-e^2sin^2B)sin^2B \end{bmatrix} \begin{bmatrix} \Delta a \\ \Delta\alpha \end{bmatrix} \tag{2.14}$$

通常人们把 2.14 式称为广义大地坐标微分公式，如果忽略尺度参数与旋转参数产生的影响，可化为一般通用大地坐标微分公式，这种情况下的公共点个数应该不少于 3 个，通过带入式中联立方程组，最终可解得 9 个转换参数 ΔX_0，ΔY_0，ΔZ_0，ε_X，ε_Y，ε_Z，m，Δa，$\Delta\alpha$[20]。

2.1.2 实体匹配

实体匹配是在两个不同源数据集中找到表示同一地理实体的对象并抽取出来的过程，匹配的方式包括几何形状、度量距离、拓扑关系、文本属性、图形结构等[21]。大致可分为以下三大类：

2.1.2.1 几何匹配

矢量法：矢量法是地理实体几何匹配研究的常用方法，通过计算候选实体几何属性之间的相似度值来判断是否为同名实体，一般依据距离、方向、长度、形状特征、位置关系来衡量。我们可以设定一个阈值，如果两个实体之间的欧氏距离小于当前阈值，则认为表示同一实体。长度和方向有时候也是判断实体间相似度的度量条件，当候选实体是线型时，在某一范围内两条线状实体的长度和方向相似度大于当前阈值，则认为是同一实体。形状特征和位置关系也是匹配实体的几何条件，包括形状的描述与比较等方法。不同实体的几何匹配采用不同的几何匹配条件，可以独立以一个条件来度量，也可以与其他指标共同作用实现匹配。

栅格法：栅格匹配是借助其他资料或工具（比如遥感影像）进行匹配的方法。如果该遥感影像与候选图中的几何精度相似的话，便可借助影像中的同一地理实体进行实体匹配。

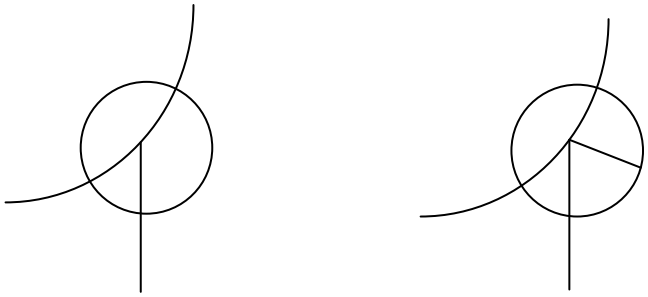
在同名实体的匹配过程中使用几何匹配的方式有着一定程度上的难度：不同源的地理实体由于数据采集方的水平差异往往会不能准确测得其几何坐标，几何位置不精确；数据集中的每个实体都要与另一数据集的实体进行匹配，工作量较大[22]。

2.1.2.2 拓扑匹配

拓扑匹配的过程是计算候选实体之间的拓扑不变性的方式来匹配同名实体。拓扑不变性指地理实体在拓扑变换（平移、缩放、旋转等）过程中实体之间拓扑关系保持不变的空间特性，拓扑匹配只有在候选实体具有拓扑关系可用性的情况下才起作用，否则需要建立实体间的拓扑关系[23]。

拓扑匹配过程由于与实体的几何精度无关，成功避免了由于几何精度不高引起的错误匹配，很好的补充了几何匹配的不足，缺点是候选实体的拓扑关系一旦发生变化后，微小的差异都会导致匹配的失败。

在图 2.3 中，第一幅图表示了某一节点有三条分支，第二幅图表示了在第一幅图节点上多出了一个分支，变成了四分支节点，由于多出来的分支造成了该节点的拓扑结构的变化，尽管变化不大，也使得在拓扑匹配过程中得不到正确的匹配结果。为了克服这一缺点，有人在该基础上又提出了“超节点”的概念，考虑在某个领域内的整体拓扑结构，而不是单独考虑这一节点，如图中一个圆形即表示一个领域。类似的，在边的拓扑匹配过程中也可加入“超边”[24]、面的拓扑匹配中加入“超面”的概念。



第一幅图（三分支节点）　　第二幅图（四分支节点）

图 2.3 拓扑匹配失败示意图

2.1.2.3 语义匹配

语义匹配的过程是通过比较候选实体之间的文本属性相似程度作为匹配的依据，主要用到的技术有中文切词技术，语义相似度计算技术等。

2.1.2.3.1 相似度的概念

不同的中文结构对相似度来说有着不同的含义，词语作为汉语语法的最小粒度单位，对相似度计算起着至关重要的作用[25]。词语的相似度计算在没有上下文的情况下存在不确定性，容易造成错误的匹配，一般是根据上下文确定不同词语的角色和权值，再做出相应计算。词语的相似度计算中涉及语义、语法、词性、句法以及众多语言因素，这就给计算过程带来了很大的困难，其中最主要的是语义因素。

我们可以把待匹配文本之间的语义相似程度用数学上的一个闭区间[0,1]来表示，区间内的数值越大表示相似程度越高，数值越小表示相似程度越低，当数值为 1 时，表示待匹配文本在语义上完全相同，数值为 0 时，即完全不相同，也失去了继续比较的意义。在实际计算中可以将词语相似度转化成计算词语间的距离来得到它们的相似度。假如可以设两个词语w_1和w_2，距离表示为$distance(w_1,w_2)$，相似度为$sim(w_1,w_2)$，则：

$$sim(w_1,w_2)=\frac{\alpha}{distance(w_1,w_2)+\alpha} \tag{2.15}$$

其中α是一个自定义的可变参数，该变换只是一种线性关系，并没有考虑到词语深度、密度等因素的影响，又比如相似度与距离成指数关系：

$$sim(w_1,w_2)=\exp\left(-\beta distance(w_1,w_2)\right) \tag{2.16}$$

其中β也是一个自定义的可变参数。

2.1.2.3.2 中文切词技术

我们知道，在英文的句子中是把空格作为单词之间的自然分隔符，而中文则是句、段能通过明显的分隔符来简单划界，没有一个形式上的真正的分界符，因此切词只有在没有明显的自然基础词语分隔符的语言中才有研究意义，如中文、日文等。中文分词就是把一个汉语句子划分成一个个单独的基础词，再将划分得到的词按照一定的组合规范重新组合的过程[26]。比如汉语句子“我准备下周去故宫”，经过切词后的结果为：“我\准备\下周\去\故宫”。

人们可以通过自己的理解判断哪些是词哪些不是词，但如何也让计算机也清楚的理解呢？所以有必要研究中文自动分词算法，目前的分词算法通常有三大类：针对字符串匹配的分词方法、基于理解的分词方法、基于统计的分词方法[27]。

（1）字符匹配法又称为机械匹配分词法，它是将待匹配的汉字串按照某种规则与一个足够完整和丰富的词典进行模版匹配，若词典中有这个字符串，则成功匹配。按扫描方向和不同长度优先匹配的情况分词方法可分为正向最大匹配法、反向最大匹配法、最短路径分词法[28]。

（2）理解法是用计算机引擎模拟人的大脑思维对语句做出理解，从而对词语进行识别，方法的基本思路就是对语句进行句法理解、语义分析、歧义消除的过程，该方法一般包括三个内容：分词子系统、句法语义子系统、总控部分。由于中文语言相比英语等其他语言具有较为复杂的特性，所以难以通过计算机模拟法对言语进行有效的识别，因此目前基于理解分词法的研究还处于试验性的阶段。

（3）统计法是通过计算字符串出现在统计库中的频率来找出分词的规律。假设 X，Y 为两个汉语汉字，当 X、Y 相邻并且出现的频率较高时，则可认为它们可构成为一个词，字与字相邻出现的频率能够较好的反映它们是否构成此的可信度，这种方法只需要对出现的词进行频率统计，无需切分词典。但是这种方法也存在一定的局限性，比如“无一”、“某一”、“之一”，这些相邻字出现的频率很高，但并不是词的常用字组。

2.1.3 坐标转换

一般来说，坐标产生的误差是由于坐标系的定位、尺度定义，椭球参数定义等造成的，此外还有大地测量产生的局部误差和控制网的测量累计误差。因此在坐标变换过程中首先应考虑坐标系定位的差异，再进行数学方法转换，对产生的误差进行拟合处理，避免大量的平差计算[29]。地图数据坐标变换的实质是建立两个不同源点集之间的坐标映射关系，可解决多源数据坐标信息不一致的问题。为了保证坐标变换的良好效果，不但要求原始数据的采集精度要高，控制点选取要合理，还要选取适当的坐标变换数学模型，否则在计算过程中会产生相应的误差，使计算结果受到影响[30]。

2.1.3.1 常用的坐标变换模型

从自身模型和转换参数两方面来讲，有两类坐标转换方法：一种是坐标与坐标之间的转换关系可以用参数表示，且参数具有明确的几何特性，如平均位移变换法、相似变换法；另一种是坐标之间的转换关系与参数无关，且参数不具有明确的几何特性，实质上是一种数值逼近方法，比如仿射变换方法、多项式变换方法。以下是这两类变换模型，(r,c)为旧坐标系中点的经纬度坐标，(x,y)为新坐标系中点的经纬度坐标。

（1）平均位移法

该方法是取整个区域公共点的经纬度平均位移量作为该区域的经纬度位置量，偏移量分别为m_0，n_0，再把原坐标系内所有点的经纬度值都加上得到的偏移值。在实际运用中，该方法一般在区域较小、公共点坐标位移量差异不大、对转换精度要求较低的情况下才予以使用，平移模型如下所示：

$$\begin{cases} x = r + m_0 \\ y = c + n_0 \end{cases} \tag{2.17}$$

（2）相似变换模型

在欧几里得空间中把图形变成相似图形的变换，诸如平移、旋转以及缩放等变换。公共点的个数与位置分布状况直接影响着变换精度，公共点的个数越多，变换精度越高；公共点分散的越均匀，变换精度也越高。该方法至少需要代入两个公共点计算出变换参数。相似变换具有几何关系明确、变换公式规则、可以适当外推的优点，平移缩放模型如下所示：

$$\begin{cases} x = m_1 \times r + m_0 \\ y = n_1 \times c + n_0 \end{cases} \tag{2.18}$$

（3）多项式变换

该方法用一个多项式表示两系统之间的变换关系，包括一阶仿射变换模型、二价多项式变换模型等。仿射变换是建立在仿射坐标系基础上坐标变换模型，是经过对原点进行平移，通过对两条坐标轴进行旋转和对两条坐标轴分别进行尺度变换实现的，不过仿射坐标系的两条坐标轴的夹角不一定都是 90^{o}，直角坐标系可以看作是仿射坐标系的特例。仿射变换如下图所示。

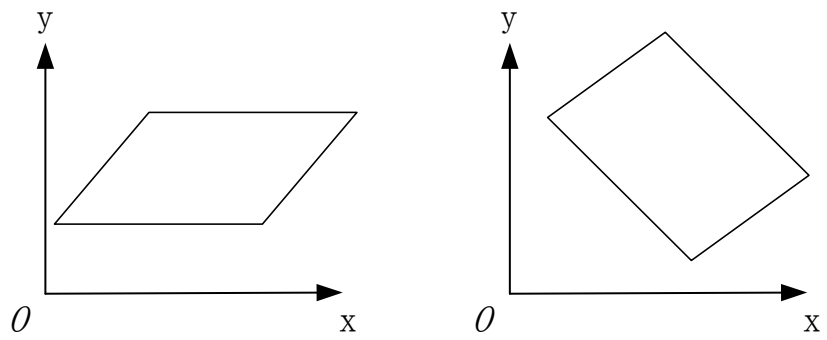


图 2.4 放射变换示意图

多项式变换存在着解析式不规则、无明显几何意义、分区后存在接边处理等问题，公共点的个数和分布状况也直接影响坐标变换的精度，公共点个数越多，变换精度越高；公共点分散的越均匀，变换精度也越高。

一阶仿射变换模型：

$$\begin{cases} x = m_1 \times r + m_2 \times c + m_0 \\ y = n_1 \times r + n_2 \times c + n_0 \end{cases} \tag{2.19}$$

二阶多项式变换模型：

$$\begin{cases} x = m_1 \times r + m_2 \times c + m_3 \times r^2 + m_4 \times c^2 + m_5 \times r \times c + m_0 \\ y = n_1 \times r + n_2 \times c + n_3 \times r^2 + n_4 \times c^2 + n_5 \times r \times c + n_0 \end{cases} \tag{2.20}$$

此外，在相似变换计算后再使用多项式曲面拟合方法对剩余的残差进一步处理，使处理后的低精度框架点坐标与高精度系统框架具有较好的一致性[31]。

借鉴摄影测量中已经成熟的遥感影像校正思想，基于二阶多项式变换模型，通过选取一系列控制点，利用最小二乘法(generalized least squares,简称 GLS)计算出变换模型参数，最终建立坐标间的映射关系。所谓的最小二乘法是一种数学优化技术，又叫做最小平方法(least square method)，通过计算误差的平方之和的最小值来寻找最佳匹配数据。利用最小二乘法可以用最简单的方法求得一些绝对不可能得到的真值，并使求得的数据与实际的数据之间的误差平方和最小，最小二乘法通常用于曲线拟合，很多其他的问题也可以通过最小化能量或最大化熵用最小二乘形式表达[32]。

对二阶多项式变换模型通过矩阵形式进行表达为：

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & r & c & r^2c^2 & r\times c & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & r & c & r^2c^2 & r\times c \end{pmatrix} \begin{pmatrix} m_0 \\ m_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \\ n_0 \\ n_1 \\ n_2 \\ n_3 \\ n_4 \\ n_5 \end{pmatrix} \tag{2.21}$$

误差方程为：

$$\begin{pmatrix} v_x \\ v_y \end{pmatrix} = \begin{pmatrix} 1 & r & c & r^2c^2 & r\times c & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & r & c & r^2c^2 & r\times c \end{pmatrix} \begin{pmatrix} \Delta m_0 \\ \Delta m_1 \\ \Delta m_2 \\ \Delta m_3 \\ \Delta m_4 \\ \Delta m_5 \\ \Delta n_0 \\ \Delta n_1 \\ \Delta n_2 \\ \Delta n_3 \\ \Delta n_4 \\ \Delta n_5 \end{pmatrix} - \begin{pmatrix} x - x^0 \\ y - y^0 \end{pmatrix} \tag{2.22}$$

即：

$$v = B\Delta - l$$

$$\text{其中，} B = \begin{pmatrix} 1 & r & c & r^2c^2 & r\times c & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & r & c & r^2c^2 & r\times c \end{pmatrix} \tag{2.23}$$

$$l = \begin{pmatrix} x - x^0 \\ y - y^0 \end{pmatrix}$$

利用间接平差的最小二乘原理，由：

$$\Delta = (\mathrm{B^T B})^{-1}\mathrm{B^T l}$$

计算出坐标变换系数，实现点的位置映射。

2.1.3.2 基于三角剖分的坐标变换

该方法利用 Delaunay 三角剖分技术分别将“参考图”和“目标图”中的同名点作为顶点进行三角划分，依据仿射变换原理对每个三角形的三个顶点进行坐标转换计算，然后将得到的转换关系应用与落入该三角形的其他点进行转换，具体过程见下图：

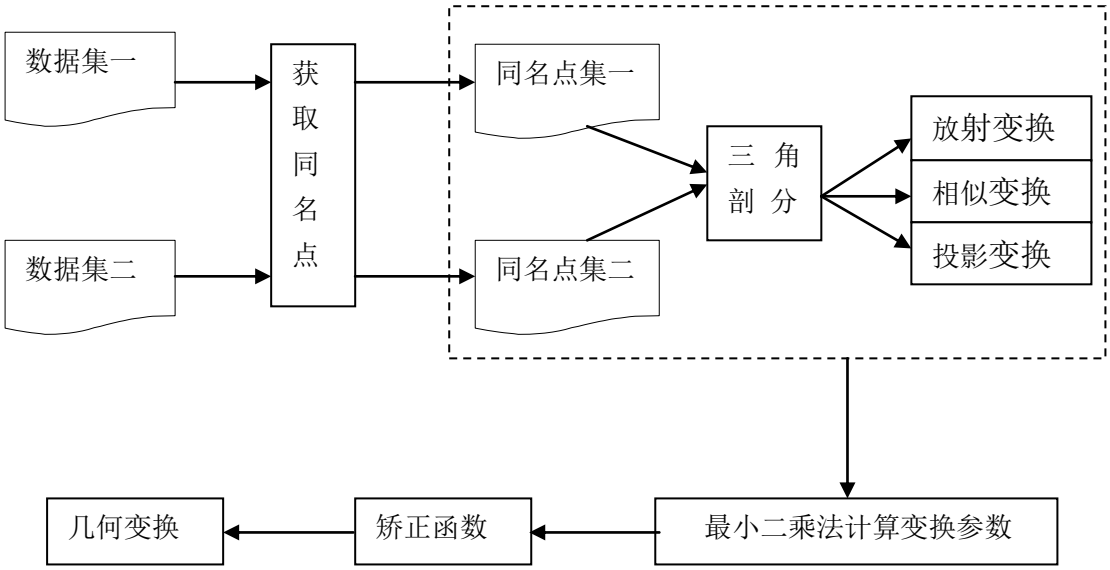


图 2.5 过程示意图

（1）以同名实体集合为顶点，对数据集一和数据集二的空间范围分别进行 Delaunay 三角剖分，形成如下图所示的三角网 T。

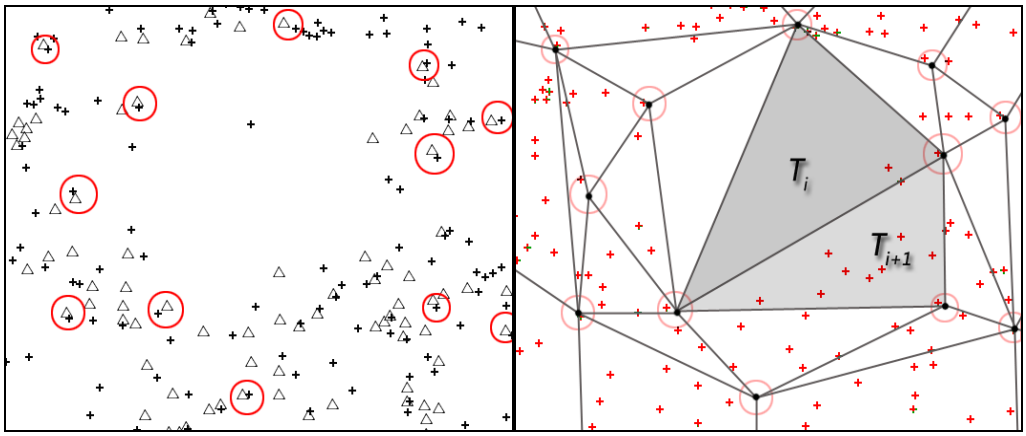


图 2.6 三角剖分过程

（2）对集合 T 中的每一个三角形的顶点进行变换计算，变换参数为$D_i=(a_i,b_i,m_i,n_i)$，纠正函数为：

$$\begin{cases} X_a = a_i X_b + b_i \\ Y_a = m_i Y_b + n_i \end{cases} \tag{2.24}$$

选取三个顶点在数据源一和数据源二中的相应对象，依据最小二乘法原理，需要满足如下条件：

$$M(a_i,b_i) = \sum_{k=0}^{n}(X_a - a_i X_b - b_i)^2 = \min \tag{2.25}$$

$$M(m_i,n_i) = \sum_{k=0}^{n}(Y_a - m_i Y_b - n_i)^2 = \min \tag{2.26}$$

因此a_i，b_i的计算公式为：

$$\begin{cases} (\sum_{k=0}^{n} X_{bk}^2)a_i + (\sum_{k=0}^{n} X_{bk})b_i = \sum_{k=0}^{n} X_{bk}X_{ak} \\ (\sum_{k=0}^{n} X_{bk})a_i + (n+1)b_i = \sum_{k=0}^{n} X_{ak} \end{cases} \tag{2.27}$$

m_i，n_i的计算公式为：

$$\begin{cases} (\sum_{k=0}^{n} Y_{bk}^2)m_i + (\sum_{k=0}^{n} Y_{bk})n_i = \sum_{k=0}^{n} Y_{bk}Y_{ak} \\ (\sum_{k=0}^{n} Y_{bk})m_i + (n+1)n_i = \sum_{k=0}^{n} Y_{ak} \end{cases} \tag{2.28}$$

还可以进一步增加变换参数$D_i=(a_i,b_i,c_i,m_i,n_i,o_i)$，纠正函数为：

$$\begin{cases} X_a = a_i + b_i X_b + c_i Y_b \\ Y_a = m_i + b_n X_b + o_i Y_b \end{cases} \tag{2.29}$$

在求解出集合 T 中每个三角形的空间位置变换参数后，依次取出数据集二中每个三角形所包含的点集对象，基于求得的变换参数实现坐标变换。

由于剖分三角形的一个顶点同时也是其他三角形的顶点，即多个三角形共用一个顶点，这样采用上述方法进行转换后势必会出现局部扭曲现象，不可避免的会使计算出现一定的误差，因此还需要进一步的研究。

2.2 属性信息的融合

多源空间数据除了有表示地理位置的几何属性之外，还具备大量的文本描述信息，如何把同一个地理目标的不同源描述信息融合为一套信息并且“为我所用”，是实现多源数据融合的重要手段。多源空间矢量数据的属性融合过程即尽可能地丰富其属性信息的过程，可以将不同来源数据各自的属性信息添加到同一个数据源中，以达到属性融合的目的[33]。例如数据源一中的 POI 数据描述信息有名称、地址、经纬度、联系电话，数据源二中的同一 POI 数据描述信息有：名称、地址、经纬度、所属类别、所属城市行政编码，数据源三中的同一 POI 数据描述信息有：名称、地址、采集时间、街景照片、背景资料等，所以可以把数据源二和三的额外描述信息添加到数据源一中，因此该条 POI 最终包含如下属性信息：名称、地址、经纬度、联系电话、所属类别、采集时间、街景照片、所属城市行政编码、背景资料。这样就避免了单一数据源信息量少、可用性不强的缺点，很大程度上满足了不同用户、不同行业部门的需求，实现了 POI 属性数据的融合。

2.2.1 分类分级和属性编码

由于不同行业研究地理问题的用途不同，因此分类和属性编码等地理特征会有所不同。即使是相同的地理实体，其分类特征和属性信息编码也会有相当大的差别，所以有必要针对不同来源数据的分类方式和属性编码进行相关研究。

2.2.1.1 分类分级

地理数据针对地理实体的表达的前提是具备一定的逻辑性，但为了成功创建逻辑上的概念就要求人们系统地归纳这些地理知识，地理分类信息是一个非常重要的地理知识系统化方法。信息分类一般是归纳那些具有相同属性或共同特征的现象或实体，分开那些特征或属性不同的现象或实体，从而形成一个多层次、并逐步扩大的分类系统[34]。

分类与分级是构成分类体系的两个方面，分类是根据不同的的类别属性将地理现象或实体归类的过程，分级是针对各个类别中的实体划分不同的等级，分类与分级描述了各地理现象或实体之间的类别、等级和从属关系。

2.2.1.2 属性编码

属性编码是基于地理实体各自不同的属性而设置的编码，对数据的合理组织和自动检索提供了帮助。编码应包括码的内容、字符格式、长度等，如可对道路要素的不同级

别分别对每一级赋予相应的属性编码，在数据处理时可按不同编码进行提取，科学合理的编码有利于数据存储与更新。编码形式一般有三种：字母型、数字型、字母数字混合型。

表 2.2 属性编码形式

编码形式	定义	优点	缺点
字母型	利用一个字母或几个字母表示	便于快速识别	占用空间相对较大
数字型	利用一个数字或几个数字表示	简单、方便排序	直观描述性差
字母型+数字型	利用字母和数字组合表示	描述直观、便于记忆	处理难度大

地理数据的属性编码还应遵循以下几项原则[35]：

（1）科学性。根据地理对象的特征赋予编码，既能对该对象进行唯一标识，又要反映其客观规律。

（2）统一性。统一性是为了符合所有用户和相关部门的需求，对编码的内容、字符格式、长度实行严格的标准化统一。

（3）实用性。在用户使用编码过程中不但能够便于操作和使用，还能够产生积极的效果。

（4）可扩展性。随着数据量的进一步增大，编码还应满足持续扩展新内容的需求。

由于 POI 数据的位置属性是坐标数值，不具备文本分类条件，所以起不到分类的作用；属性信息作为描述该数据的文本信息，虽然具有分类条件，但因其篇幅较长，类别属性不容易直接提取出来，还会对分类结果产生干扰，具有较大的噪声，所以名称信息就作为分类的唯一要素。

中文 POI 名称一般是各种商业场所、机构名、地名或者机构名+地名的方式构成，有一定的规律性，在形式上可以由[装饰词]+[特征词]来表示，通常情况下[特征词]回答了该兴趣点到底是什么的问题，因此需要采用对名称进行中文切词和语义相似度匹配的方式匹配出[特征词]与 POI 基础分类库作对比，实现对数据的分类处理。

2.2.2 属性数据相似度计算

2.2.2.1 哈罗-温克勒距离算法

在计算机科学和统计学中，哈罗-温克勒距离是衡量两个字符串之间相似性的数学指标，它是温克勒距离算法的一种变形，主要应用于数据组织和处理等领域[36]。它比较适合计算比如名称短语这样较短的字符间的相似程度，经过距离计算后得到的值越高说明相似程度越大，0 分表示没有任何相似度，1 分表示完全匹配。

算法计算公式：

$$d_j = \frac{1}{3}\left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{m}\right) \tag{2.30}$$

其中：s1、s2表示待匹配的字符，m 是匹配的字符数，t 是换位的数目

$$MW = \left(\frac{Max(|s1|,|s2|)}{2}\right) - 1 \tag{2.31}$$

其中；MW 为匹配窗口值。

（1）字符串 s1 与字符串 s2 在做匹配计算时，当两个字符的距离不大于公式二的最后结果(匹配窗口)即认为是匹配的。

（2）当 s1、s2 中字符相匹配但是字符位置不一样时会发生换位操作，而公式一中换位的数目 t 为不同顺序的匹配字符数目的一半。比如把两个字符串 TRACE 和 CRATE 进行匹配处理，结果是字符串中只有'R' 'A' 'E'三个字符匹配成功，即 m=3。为什么'C', 'T' 没有成功匹配呢，因为虽然两个字符串中都出现有'C', 'T'，但是通过公式二得出匹配窗口值为 (5/2)-1=1.5。而两个字符串中'C', 'T'字符的距离均大于 1.5，所以不算做匹配，因此 t=0。在另一组字符串 DwAyNE 与 DuANE 。匹配的字符 D-A-N-E 在两个字符串中有相同的字符顺序，所以不需要进行换位操作，因此 t=0，m=4。

2.2.2.2 词共现相似度计算方法

该方法是基于传统向量空间模型（VSM）和词共现模型的文本相似度计算方法，传统向量空间模型是把文本表示为空间向量，组成该文本的不同词分别赋予一定的权重，然后通过计算向量与向量夹角的余弦值得到文本之间的相似度，例如 α 、β 分别表示文

本向量，φ表示不同词的权重值，λ表示组成文本的词语个数，即可将相似度值表示如下：

$$\sin(\alpha,\beta)=\cos(\alpha,\beta)=\frac{\sum_{i=1}^{\lambda}\varphi_{\alpha i}\varphi_{\beta i}}{\sqrt{\sum_{i=1}^{\lambda}\varphi_{\alpha i}^{2}}\sqrt{\sum_{i=1}^{\lambda}\varphi_{\beta i}^{2}}} \tag{2.32}$$

词共现模型研究是统计学研究领域的一个重要课题，根据词共现模型可以找到与一个中心词经常搭配的一组词语集，在一定程度上表达了这个词的上下文和文本的语义信息[37]。词共现模型是建立在一个假设基础上的：在大量的文本资料中如果经常出现几个词，便认为这几个词具有相互关联性，例如：在一篇文章中如果大量出现“股票”、“投资”、“期货”这样的词语，则可断定这是一篇金融投资方面的文章。而且共现词出现的频率越高，上下文语义信息越紧密。

词共现相似度计算过程如下：先对待处理文档进行分词预处理，则文档可表示为$T_i=(t_{i1}, t_{i2}, t_{i3}, \ldots, t_{im})$，其中$t_{im}$为基础词；然后基于每个词出现的频率计算各自的权重，若权重值较大，则说明该词在文本中起着较大影响作用；接着对权重结果进行统计，删除低频词，也就是去掉出现次数较少的词；最后带入 VSM 模型计算公式。

词的频率分为绝对频率（tf）和相对频率（idf）两个部分，绝对频率指该词在文中出现的频率，相对频率指该词在全部文档中出现的频率[38]。所以词频的权重$\varphi=tf\times idf$，文本采用最常用的 TF-IDF 公式：

$$\varphi_{ik}=\frac{(\lg(tf_{ik})+1.0)\times\lg(N/n_k)}{\sqrt{\sum_{k=1}^{\iota}[(\lg(tf_{ik})+1.0)\times\lg(N/n_k)]^2}} \tag{2.33}$$

其中$\lg(N/n_k)$即为 idf，N 是文档中的文本数。

下面给出基于传统向量空间模型（VSM）和词共现模型的文本相似度计算公式：

$$\varphi'_{\alpha i}=\varphi_{\alpha i}\times(1+\lg(1+S_{\alpha i})) \tag{2.34}$$

$$\varphi'_{\beta i}=\varphi_{\beta i}\times\left(1+\lg\left(1+S_{\beta i}\right)\right) \tag{2.35}$$

$$\sin(\alpha,\beta)=\cos(\alpha,\beta)=\frac{\sum_{i=1}^{\lambda}\varphi'_{\alpha i}\varphi'_{\beta i}}{\sqrt{\sum_{i=1}^{\lambda}\varphi'^{2}_{\alpha i}}\sqrt{\sum_{i=1}^{\lambda}\varphi'^{2}_{\beta i}}} \tag{2.36}$$

3 数据获取与分类处理

地理数据的获取和采集工作是建立地理信息系统过程中必不可少的一个部分，如果少了数据的支撑系统的建立就无从谈起，而数据质量的高低也直接影响着地理信息系统的内容成果和研究价值；在获取数据后，作业人员会进行相应的数据处理，一方面对输入的数据进行检查和评估，另一方面做进一步的分类、归纳和分析处理。由于互联网上的 POI 数据来源众多，同一条不同来源的 POI 所定义的类别各有差异，所以在数据融合过程中会产生该 POI 具体划分到哪一个数据类别下的问题，因此首先需要总结和建立一套较为完整的数据分类体系，以便于接下来的数据分类处理。本章根据上述问题首先提出了一种基于网络爬虫的 POI 数据自动提取技术，可支持对互联网电子地图 POI 数据进行实时抓取，然后对目前存在的众多 POI 分类体系做了收集和比较，并在此基础上对这些分类体系进行了归纳和融合处理，最后形成了一套新的数据分类体系，为多源 POI 融合方法的研究提供了支持。

3.1 数据获取

地理信息数据的来源非常广泛，既可通过人工外业采集获得，又可通过航空遥感、GPS 等现代化手段获取，在网络地理信息系统发展日新月异的今天，互联网上存在着海量的地理数据开放资源，所以有必要研究如何能够从互联网快速获取批量、高时空性能的地理数据。

3.1.1 多源 POI 数据的特点

（1）多语义性。GIS 研究对象的多种类型特点决定了地理信息的多语义性。对于同一个地理 POI 对象，在现实世界中其几何特征是一致的，但是却对应着多种语义，如地理位置、海拔高度、气候等自然地理特征；同时也包括经济社会信息，如行政区界限、人口、产量等。不同 GIS 解决问题的侧重点也有所不同，因而多源 POI 会存在语义差异。

（2）多时空性。地理空间数据具有很强的时空特性。地理信息系统中的数据源既有同一时间不同空间的数据系列，也有同一空间不同时间序列的数据。这一特性在 POI 数据中尤为明显，随着城市不断的规划和土地的更加合理利用而使得 POI 数据的生命周期极为短暂，因此 POI 的数据更新频率很大程度上影响着 GIS 数据质量。。

（3）获取手段多源性。目前获取 POI 数据的方法有多种多样，包括来自现有系统、图表、遥感手段、GPS 手段等。这些不同方式获得的数据其数据模型及处理手段都各不相同。

（4）存储格式多源性。地理信息应用系统在较长一段时间内都是以具体项目为中心的孤立发展状态存在的，数据采集商各自的数据存储方式与数据处理模型没有统一标准，而现在常用的 GIS 处理软件也都有自己的数据格式。

（5）空间基准不一致。不同来源的空间数据有着不同的坐标参考体系和不同的投影方式，这使得 POI 的数据共享问题变得尤为突出。

3.1.2 POI 获取方式

（1）扫街式实地采集

传统的 POI 数据采集方式主要是基于地图底层数据而言，大多数 POI 信息的采集手段还都是比较原始的，需要投入大量的外业测绘人员通过一系列设备对目标地区进行“扫街式”采集，这一采集方式具有采集成本较高、效率较低、数据维护成本较大的特点。测量人员通过使用手持 GPS 接收机等设备测得单个 POI 数据的空间位置信息，这些 GPS 接收机可随时与笔记本电脑或个人掌上平板电脑进行连接，实现数据的实时传输。在采集数据过程中，操作人员通过记录单个 POI 数据的文本描述信息来获取该条 POI 的属性信息，与此同时，还可以使用数码相机将该兴趣点的名称、门牌号、联系电话、经营时间等信息拍摄下来，在内业人员输入数据时可利用图像识别技术将所拍摄的内容一一识别并记录下来。

（2）基于 VGI 的 POI 数据共享

自发性地理信息（Volunteered Geographic Information，VGI）是以城市各行各业人员为载体的城市空间数据，其思想是用户自发的贡献地理数据，该手段不受时间、地域、和网路的限制，且具备数据量大、社会化信息丰富、现势性强、覆盖区域广等特点。目前国内外一些地图网站比如谷歌地图和百度地图都开通了用户地图标注功能，网友可以把自己身边存在而地图上没有的建筑物标绘在图层上，同时还可以指出该地图上出现的一些错误和问题，这样大大降低了传统行业数据采集制作的成本，提高了数据采集的效率和质量，丰富了数据的内容和形式。

（3）基于网络爬虫的 POI 垂直搜索

互联网上目前存在着大量的地理空间数据，网络爬虫作为一种互联网搜索引擎，通过连接到某个特定网站并对网页内容进行读取，对所需要的内容进行连续自动抓取，具有全自动化、无人工干预、效率高等特点。当前互联网各大地图网站拥有着庞大的 POI 数据资源，总数超过了 1 亿条，因此从海量的网络资源中通过使用搜索引擎采集技术是获取 POI 的一种有效方法。对于网络爬虫实现的困难体现为选择和实现存储的数据结构、分配临界资源、实现多线程等。

3.1.3 POI 自动提取

互联网上的 POI 数据都是通过各个电子地图网站作为载体并且展现的，人们浏览地图的过程中经常会通过关键字搜索的方式来获取自己感兴趣的地理目标，这些兴趣点的空间位置与属性描述信息会实时的显示在地图网页中（如图 3.1），即 POI 数据会保存在 Html 半结构化文本中，通过对该半结构化文本内容的分析和解析，最终提取出所需要的 POI 数据。

图 3.1 百度地图学校搜索

3.1.3.1 Html 网页结构解析

Html 即超文本标记语言，“超文本”就是指网页页面内容可以包括图片、链接、甚至是音乐等非文本信息，网页文件本身是一种文本文件，它通过一些标签来标记网页中显示的不同部分，这些非文本信息被标签按照一定格式和顺序有序的组织起来，用户使用浏览器打开网页文件的过程即顺序读取 Html 文本的过程。

Html 文本中使用大量的标签来标记不同的页面显示内容，标准的超文本标记语言文件都具有成对出现的标签，符号“<”和“>”用来标识一个标签的起始边界和结束边界，如<Html>…</Html>即为一个完整标记，<Html>表示文件的开头，</Html>表示文件的结尾，该标签由头部（head）和主体（body）两个部分组成。头部中包含该页面的标题、类型、字符编码、网页内容关键字等信息，主题中包含该页面的布局、正文文本、

超链接等方面的信息。由于 Html 网页文件是不同标签不断嵌套所形成的，因此可通过如图 3.2 所示的树状结构来表示 Html 文件，这种树状形式很容易被计算机识别并处理。

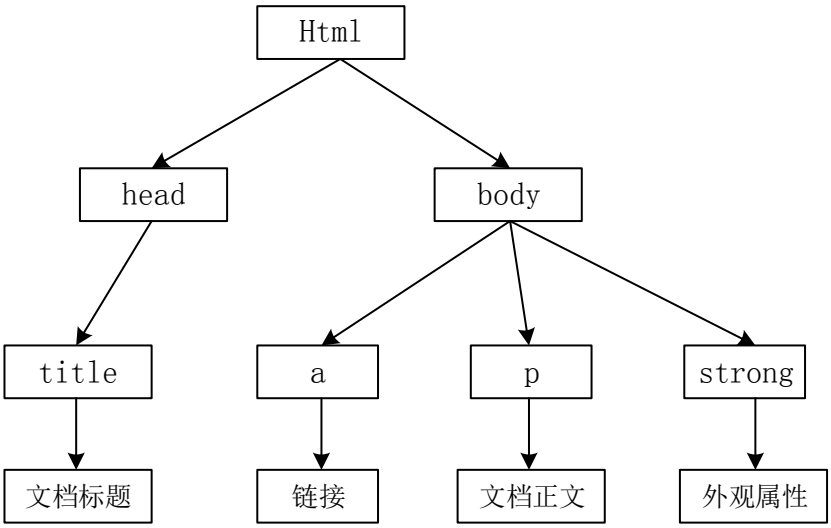


图 3.2 Html 树状结构

由上图可以看出，Html 文件被转换成了一个树状结构，标签<Html>为树的根节点，它的两个孩子节点为<head>和<body>，<title>、<a>、<p>互为兄弟节点，Html 文件标签从功能上大体分为三个类别[39]：

（1）控制外观属性的标签，如，<i>，<b>，<font>等。

（2）设置网页布局的标签，如 ，<tr>，<td>，<ul>等。

（3）引入外部内容的标签，如<a>，<img>，<map>等。

3.1.3.2 基于 HtmlParser 的自动提取

HtmlParser 是一个解析 Html 文档的辅助工具，它是用 Java 语言写的辅助包，能够超高速解析 html 网页内容，并且不会出错，它主要提供以下两种功能：

（1）信息提取

a、文本信息抽取，例如对 Html 进行有效信息搜索

b、链接提取，用于自动给页面的链接文本加上链接的标签

c、资源提取，例如对一些图片、声音的资源的处理

d、链接检查，用于检查 Html 中的链接是否有效

e、页面内容的监控

（2） 信息转换

a、链接重写，用于修改页面中的所有超链接

b、网页内容拷贝，用于将网页内容保存到本地

c、内容检验，可以用来过滤网页上一些令人不愉快的字词

d、Html 信息清洗，把本来乱七八糟的 Html 信息格式化

e、转成 XML 格式数据

HtmlParser 的核心模块是 org.htmlparser.Parser 类，该类提供针对网页中的每个标签进行遍历搜索的一些方法，通过调用这些方法便可匹配出网页内容中的 POI 信息并存储在数据库中。例如提取百度地图北京地区的学校信息，通过图 3.1 中的操作后，查看该页面的 html 源码（图 3.3），发现 POI 信息都存在“p_left”这个 class 中，通过遍历所有的“p_left”类，提取出包含 POI 信息的字符串，部分实现代码为：

```
if (tag.Attributes["CLASS"] =="p_left" ) {
    if (tag.Attributes["CLASS"].ToString() == "poi-fav"){
        string temp = tag. ToPlainTextString();
    }
//获取节点间的内容
if (htmlNode.Children != null && htmlNode.Children.Count > 0)
{
        this.RecursionHtmlNode(current, htmlNode.FirstChild, true);
        //content = new TreeNode(htmlNode.FirstChild.GetText());
        //treeNode.Nodes.Add(content);
}
//the sibling nodes
if (siblingRequired)
{
    INode sibling = htmlNode.NextSibling;
    while (sibling != null)
    {
        this.RecursionHtmlNode(treeNode, sibling, false);
        sibling = sibling.NextSibling;
      }
   }
```

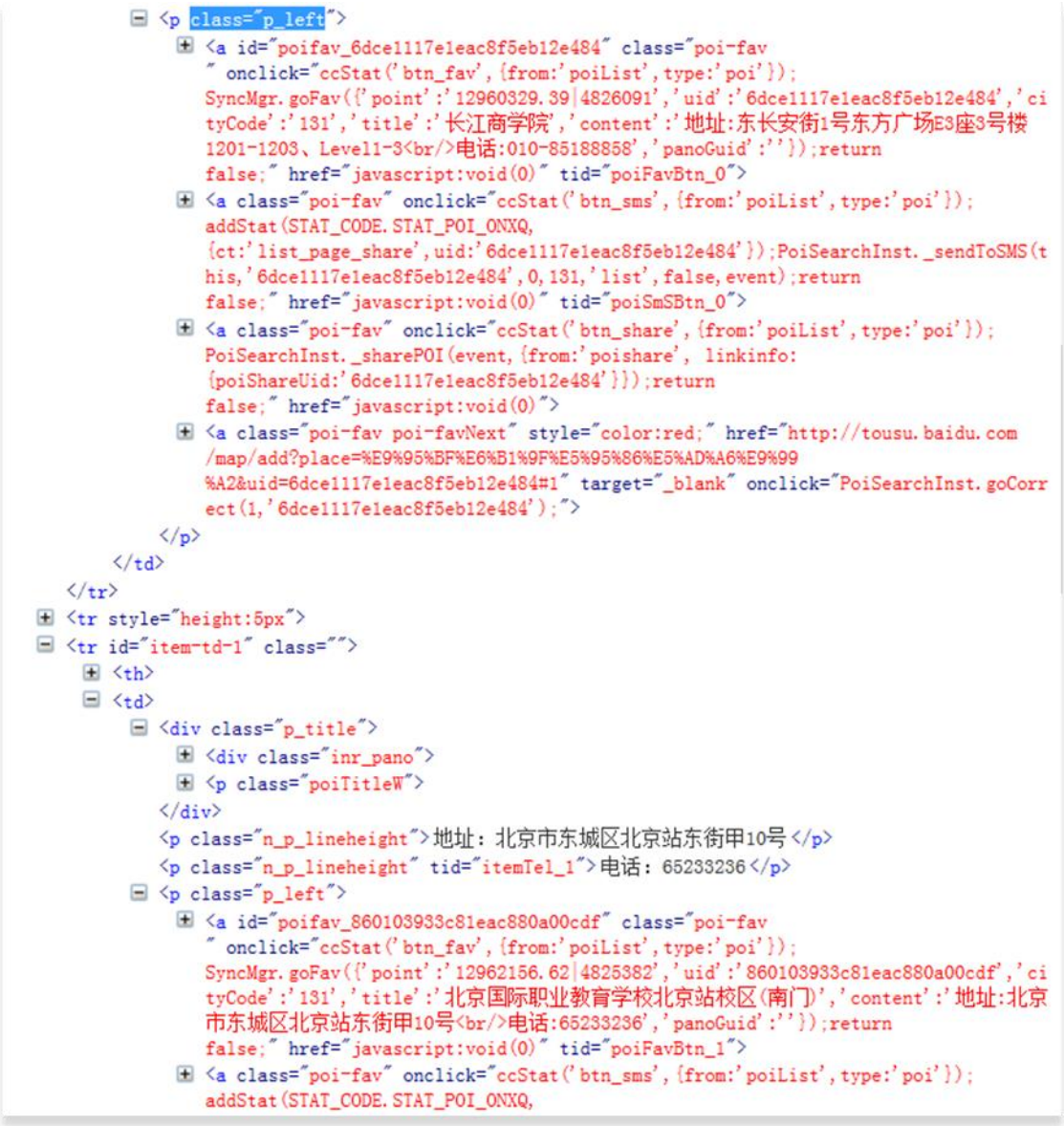


图 3.3 Html 页面结构

匹配出的前两条 POI 字符串结果如下：

ccStat('btn_fav',{from:'poiList',type:'poi'});SyncMgr.goFav({'point':'12960329.39|4826091','uid':'6dce1117e1eac8f5eb12e484','cityCode':'131','title':'长江商学院','content':'地址:东长安街 1 号东方广场 E3 座 3 号楼 1201-1203、Level1-3
电话:010-85188858','panoGuid':''});return false;

ccStat('btn_fav',{from:'poiList',type:'poi'});SyncMgr.goFav({'point':'12962156.62|4825382','uid':'860103933c81eac880a00cdf','cityCode':'131','title':'北京国际职业教育学校北京站

校区(南门)','content':'地址:北京市东城区北京站东街甲 10 号
电话:65233236','panoGuid':''});return false;

这时匹配出的结果只是包含有 POI 信息的文本字符串，还需借助正则表达式匹配出标准 POI 信息，经过匹配并组合后的结果如下：

POI1：长江商学院，东长安街 1 号东方广场 E3 座 3 号楼 1201-1203，12960329.39，4826091，010-85188858。

POI2：北京国际职业教育学校北京站校区(南门)，北京市东城区北京站东街甲 10 号，12962156.62|4825382，65233236。

本文通过该方法自动提取了百度地图和谷歌地图中北京和上海地区的部分 POI 数据，作为本文中所需的实验数据，数据提取过程如图 3.4 所示。

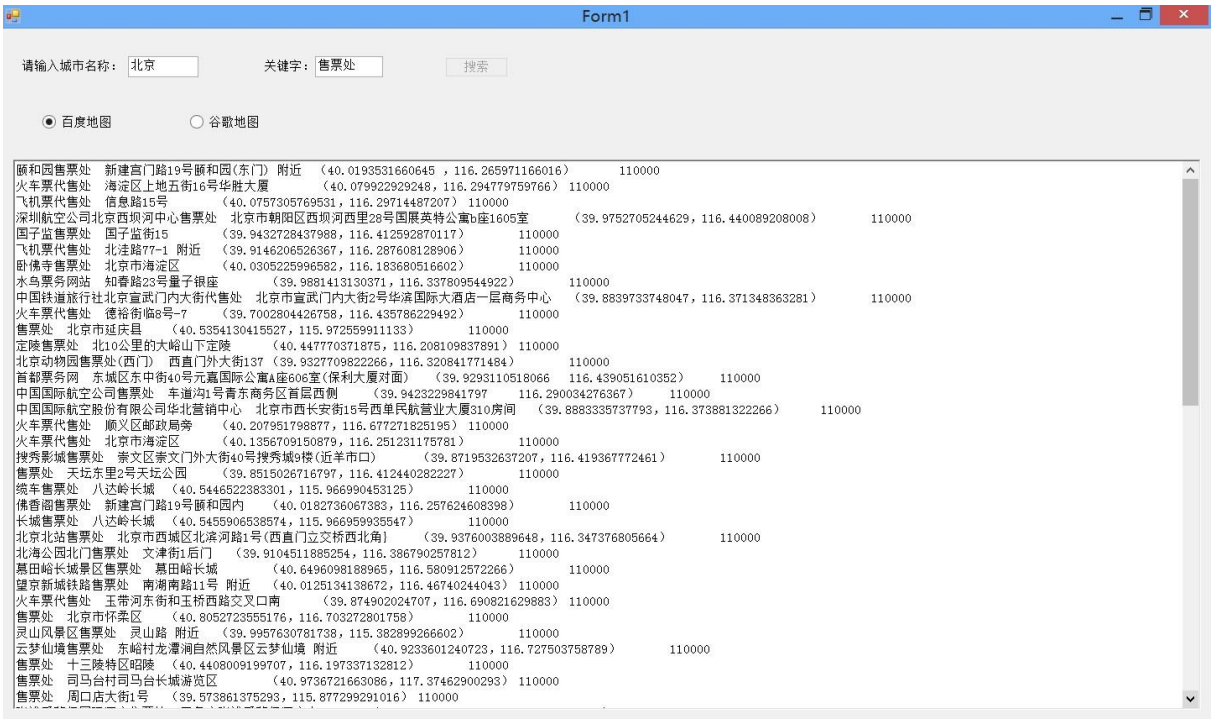


图 3.4 POI 自动提取过程

3.2 多源 POI 数据分类体系融合

目前互联网上的电子地图网站针对 POI 信息都有自己的一套数据分类体系，这些各不相同的分类体系给 POI 融合带来了一定的影响，往往使一部分数据没有得到融合处理。POI 数据分类体系的融合处理是一个系统的工程，既要涵盖原分类中所有涉及的内容，保证分类融合结果的完整性，又要建立新的层次与逻辑关系，不产生冗余数据。下图为多元异构分类体系融合的过程。

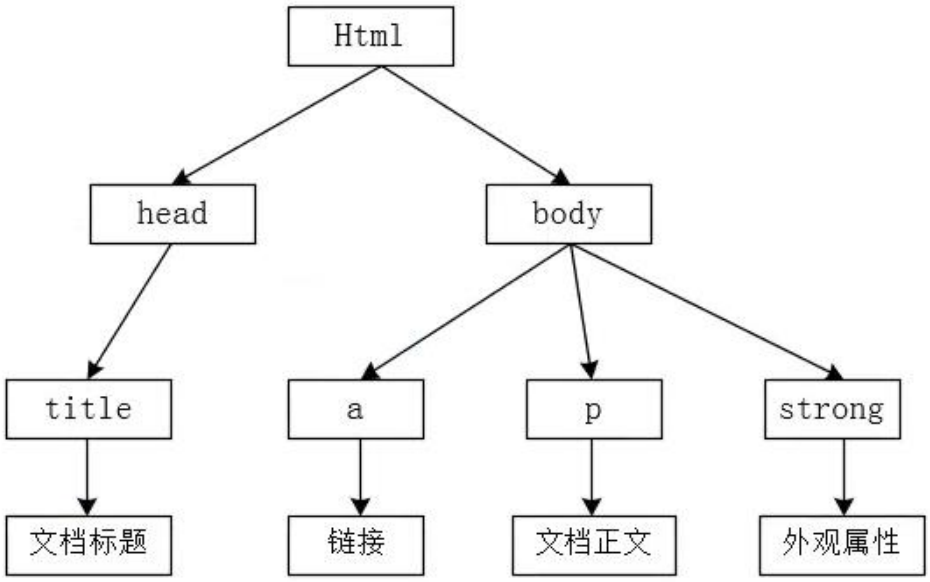


图 3.4 分类体系融合过程

3.2.1 分类的指导思想与基本原则

POI 的一个重要属性是对什么感兴趣，在调查分析后发现，大众用户对餐饮、住宿、购物、交通路线、医疗设施、娱乐场所等类型的位置服务更感兴趣，这些位置服务极大地方便了人们的生活方式。由于数据来源于不同的大型电子地图网站，其各自的分类标准有着一定的差异，比如百度地图 POI 的一级分类有：餐饮、丽人、景点、地铁、超市、KTV、银行、加油站、酒吧、购物、ATM、网吧、咖啡厅、停车场、学校、医院、洗浴，谷歌地图 POI 一级分类则是：餐饮、购物、住宿、出游、娱乐、服务、教育、健康、企业、国家机构，所以在构建 POI 分类体系的过程中要尽可能的涵盖每个网站自身现有的分类体系，形成一套合理、完整的分类成果，这样才能确保在多源 POI 数据分类处理中不遗漏每一条数据，在构建分类体系过程中应该遵循以下 5 条基本原则[41]。

（1）普遍性。POI 分类应该基于普通大众用户的基本需求，重点关注他们的衣、食、住、行等一切日常生活与之相关的活动，既要满足一般的普遍性，又要尽量结合自然、人文地理信息，分类体系要做到清晰、合理。

（2）一致性。本文分类体系应与国家《基础地理信息分类》规定的内容具有一致性。《基础地理信息分类》采用科学的分类体系，从基础地理信息角度对地理信息要素进行了系统而全面的整理、归类与补充，通过要素的分类和编码，确定类别、等级明确的代

码结构，最终形成我国统一和协调一致的基础地理信息要素分类代码标准文本[42]。POI 分类体系可依照该信息分类，充分利用和深度挖掘这些信息，与基础地理信息的分类保持一致。

（3）稳定性。POI 分类体系根据各个不同类别稳定的特征和属性，建立比较稳定的分类标准，能在较长时间内不发生大的变化。

（4）完整性与可扩展性。POI 分类体系包含庞大的地理信息要素实例，内容涉及许多方面，既要反映要素的自身特征，又要体现不同要素之间的相互关系，完整性必不可少。此外，分类体系还应满足持续扩展新内容的要求，当前我们正处于一个信息持续增长的时代，POI 内容的更新速度和频率越来越快，及时丰富扩展分类体系至关重要。

（5）公共性。POI 信息内容应只服务于普通大众用户或相关专业机构，不能涉及危害国家安全的相关敏感内容。POI 信息的分类内容不应该包含国防、军事设施、军事单位，武器弹药、爆炸物品、剧毒物操、危险化学品、铀矿床和放射性物品的集中存放地等与公共安全相关的设施，未经公开的机场、港口和机关、单位等。

3.2.2 本文采用的分类方法

基本的信息分类方法包括线分类方法和面分类方法。

1）线分类法也称层次分类法。它将要分类的对象（被划分的事物或概念）按其所选择的若干个属性或特征，按最稳定本质属性逐次地分成相应的若干层类目，并排列成一个层次的逐级展开的分类体系。

在这个分类体系中，同位类的类目之间存在并列关系，且不重复，也不交叉；下位类与上位类目之间存在着隶属关系。

所谓上位类，即在线分类体系中，一个类目相对于由它直接划分出来的下一级类目而言，称为上位类。下位类即在线分类体系中，由上位类直接划分出来的下一级类目相对上位类而言，称为下位类。同位类，即在线分类体系中，由一个类目直接划分出来的下一级类目，彼此称为同位类。

2）面分类法，又称平行分类法，是指将所选定的分类对象的若干标志视为若干个面，每个面划分为彼此独立的若干个类目，排列成一个由若干个面构成的平行分类体系。面分类法分类时所选用的标志之间没有隶属关系，每个标志层面都包含着一组类目。

表 2.1 两种地理信息分类方法比较

方法	优点	缺点
线分类法	1)层次性好,能较好地反映类目之间的逻辑关系; 2)使用方便,既符合手工处理信息的传统习惯,又便于电子计算机处理信息。	1)结构弹性较差，分类结构一经确定，不易改动; 2)效率较低，当分类层次较多时，代码位数较长。
面分类法	1)有较大弹性，一个“面”内类目改变，不影响其他“面”；2)适应性强，可视需要组成任何类目； 3)易于添加和修改类目。	1)不能充分利用容量，可组配的类目很多，但有时实际应用的类目不多； 2)难于手工处理信息。

3.2.3 多源 POI 分类体系融合

目前国内外的主要大型地图门户网站已经针对各自的 POI 数据建立了相应的数据分类体系，例如高德地图根据用户的不同需求将数据分为餐饮、酒店、医院、学校、电影院、超市、商场、银行、景点、地铁十个一级分类（如图 3.4），将餐饮又细分为中餐馆、西餐馆、日本料理、韩国料理、快餐小吃、咖啡馆、蛋糕房、肯德基、麦当劳、必胜客等二级分类。

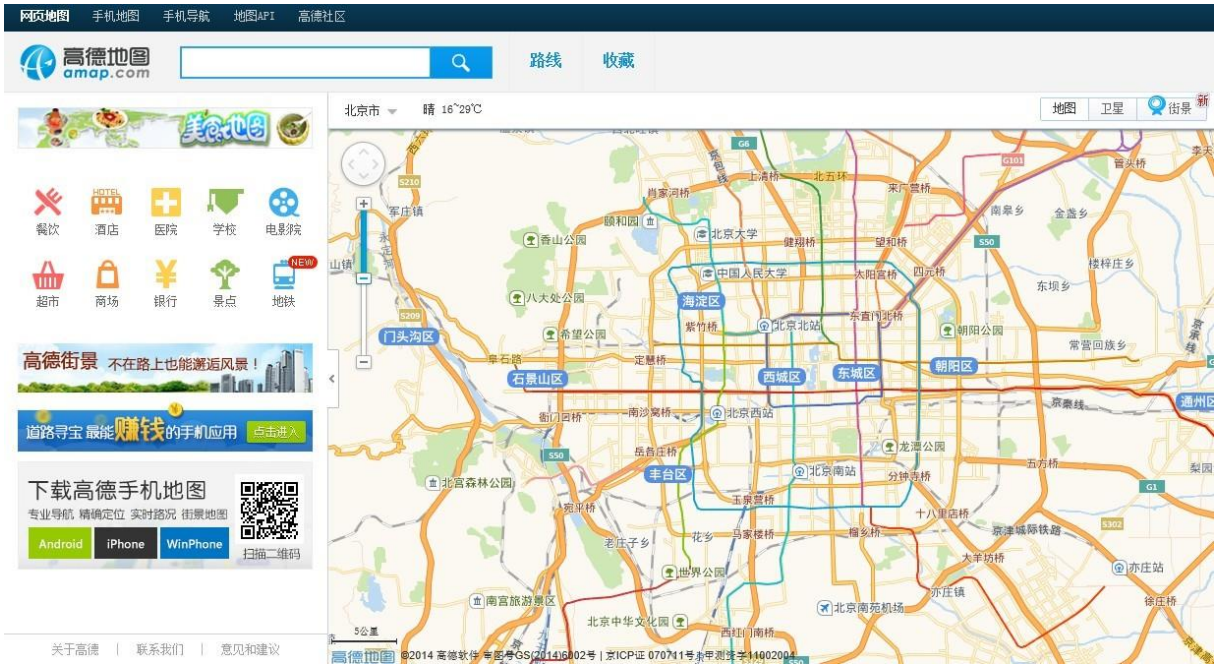


图 3.4 高德地图 POI 分类

本文收集并归纳总结了国内外相关部门或机构针对 POI 以及地理空间数据的分类体系，其中包括 11 家大型地图网站，新国家标准《基础地理信息要素分类与代码》，民政部颁发的《地名分类与类别代码编制规则》，由于以上各自的分类体系各不相同并且数据量较大，在此不一一列举，只列出各自不同分类级别的节点个数，如表 3.1 所示：

表 3.1 各部门与网站分类体系统计表

类别个数 / 分类来源	一级分类	二级分类	三级分类	四级分类
百度地图	14	44	54	40
腾讯地图	9	0	0	0
高德地图	9	39	0	0
谷歌地图	10	73	125	5
天地图	10	51	0	0
搜狗地图	8	43	94	0
51 地图	8	77	0	0
图盟	4	29	0	0
图吧	13	16	83	0
中国搜索	10	0	0	0
必应地图	10	74	0	0
Esri 分类	15	139	0	0
民政部	2	10	62	168

根据 POI 分类应遵循的基本原则，采用线分类法与面分类法相结合的方法，对以上不同分类体系和标准进行了融合处理，最终得到一级分类下的类别数为 13，二级分类类别个数为 109，三级分类类别个数为 465，四级分类类别个数为 69。详细分类体系内容见附录 A。

4 基于网格划分的多源 POI 位置纠正

本章研究基于网格划分地理目标范围的多源 POI 位置纠正技术，其基本思想是以一种坐标值精度较为可信的数据源作为参考标准，把其余的“非标准”数据源中 POI 数据的坐标值依照参考标准纠正为“标准”坐标值，以实现多源数据几何位置的融合。在实际应用中，为了减少控制点选取不准确以及位置分布不均匀带来的影响，往往尽可能多选取一些控制点以保证精度要求，目前控制点的选取方式主要分为三种：(1)纯人工选取控制点(2)人机交互式选取控制点(3)计算机自动选取匹配控制点。

首先对待处理地理范围进行单元格划分，分别对每个单元格利用中文语义匹配的方法提取出控制点集合与检核点集合，再通过粗差剔除过程进一步过滤控制点集合，利用二阶多项式模型计算出各单元格纠正系数，然后结合检核点集合进行精度评定，最后将计算出的纠正系数写入到数据库，具体过程如图 4.1 所示。

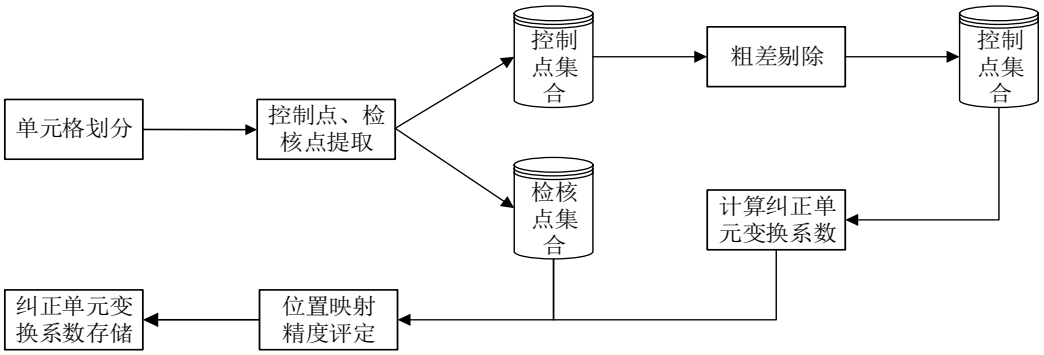


图 4.1 基本流程图

4.1 地理范围网格划分

由于采用多项式坐标变换的方法对较大范围内的大量 POI 点群进行位置纠正会导致误差增大，因此本文采用分而治之的思想：把待处理数据的地理范围划分成一个个的小单元格，针对落入不同单元格内的 POI 点群分别进行坐标变换处理，求得每个单元格内的坐标变换系数，从而最大可能的减小计算中产生的误差。

首先计算目标点群的地理范围（Bounding Box），再按照一定的地理格网间距，将该地理范围拆分成 M*N 个正方形单元格，在这里我们称每个单元格为一个“纠正单元”，为了确保纠正单元内用于计算变换系数的控制点合理分布，再进一步对每个纠正单元分别按 3*3 和 6*6 进行均匀拆分，这样同一个纠正单元格就被划分了两次，分别形成了 9 个子网格与 36 个子网格。我们将这 9 个子网格命名为控制点子网格，作用是将纠正单

元平均地划分为了 9 个均匀的小区域，在每个小区域内找到一个 POI 点作为多项式计算过程中的公共点，即坐标控制点，这样划分的目的是更好地实现坐标控制点的自动选取。

有时经过自动选取得到的控制点自身也存在着较大的精度误差，为了避免坐标换算过程中变换系数误差增大，需要按照粗差过滤原理对控制点进行进一步精度过滤，方法是把之前划分得到的 36 个子网格命名为检核点子网格，在每个检核点子网格中找到 POI 检核点，根据坐标检核点计算出整体的中误差，最后剔除精度大于三倍中误差的坐标控制点。

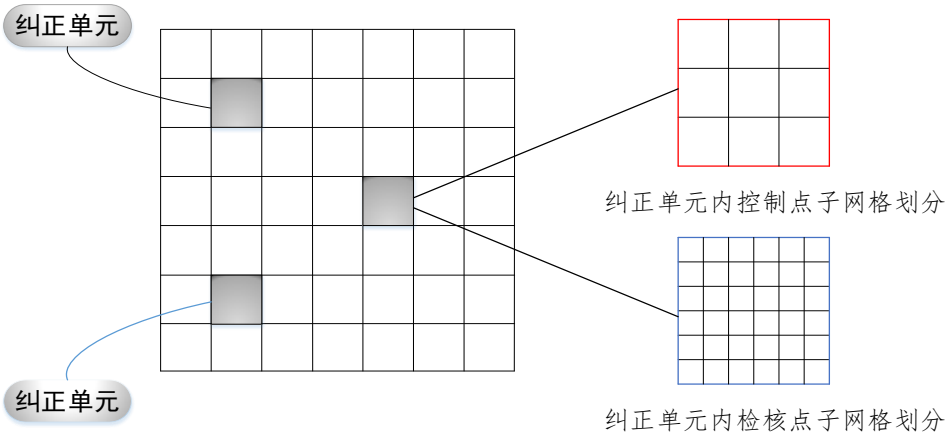


图 4.2 纠正单元划分

网格划分代码如下：

```
void DivideGridByBLInterval()
{
    double intervalb = degree;      //定义经度方向的划分间隔
    double intervall = degree;       //定义维度方向的划分间隔
    //计算经度方向划分的网格数
    int row_count = Convert.ToInt32(Math.Ceiling((bmax - bmin) / intervalb));
    //计算纬度方向划分的网格数
    int col_count = Convert.ToInt32(Math.Ceiling((lmax - lmin) / intervall));
    int[,] id = new int[row_count, col_count];
    for (int i = 0; i < row_count; i++)      //划分地理网格
        for (int j = 0; j < col_count; j++)
        {
            id.SetValue(col_count * i + j + 1, i, j);
            bmin_temp = bmin + intervalb * i;
```

```
            bmax_temp = bmin + intervalb * (i + 1);
            lmin_temp = lmin + intervall * j;
            lmax_temp = lmin + intervall * (j + 1);
            if (bmax_temp > bmax)
            {
                bmax_temp = bmax;
            }
            if (lmax_temp > lmax)
            {
                lmax_temp = lmax;
            }
    }
```

4.2 控制点与检核点自动提取

利用中文语义匹配的方法，对不同源两套 POI 数据的名称和地址进行相似度计算，匹配出共指 POI 对象。

4.2.1 共指对象的定义

共指对象是指不同资料、数据或文献中描述的同一现象或实体，本文中提到的共指对象是不同源数据集中的同一地理实体。有时候人们习惯称呼某一地理对象的简称或缩略语，比如“中国人民大学”，会称之为“人民大学”、“人大”或“中国人大”，尽管名称不同，但都表示统一对象，这样就造成了名称的歧义现象。共指对象在 POI 数据集中也普遍存在，多源 POI 数据融合技术可以有效地消除这种现象。

4.2.2 共指 POI 自动匹配

对于每个控制点子网格范围内的两套不同源 POI 点集，利用中文分词和语义匹配的方法获得两套点集中的任意一组共指对象作为该子网格内的控制点，这里的同名同址点指名称和地址语义相似度超过 80%，此时我们认为这两个不同源点表示同一地理对象，如点集一：外航服务公司海洋大厦机票销售处，地址：延安东路 550 号海洋大厦；点集二：上海外航服务公司海洋大厦销售处，地址：延安东路 550 号；尽管点 1 与点 2 名称和地址都不完全相同，但相似度匹配都超过 80%，即表示同一地理对象。采用同样的方法对检核点子格网提取出其范围内的检核点，作为所属纠正单元内的检核点。

4.3 控制点粗差过滤

对于以上纠正单元内获取的控制点，再采用二阶多项式变换模型计算出变换系数，把当前所有控制点代入计算出每个控制点的残差以及中误差；利用粗差过滤的方法，对于残差大于三倍中误差的控制点，直接删除，否则保留下来；利用循环迭代的方法，直至所有点的残差都小于三倍中误差，最后保留的点为该纠正单元内的最终控制点。具体流程如图所示。

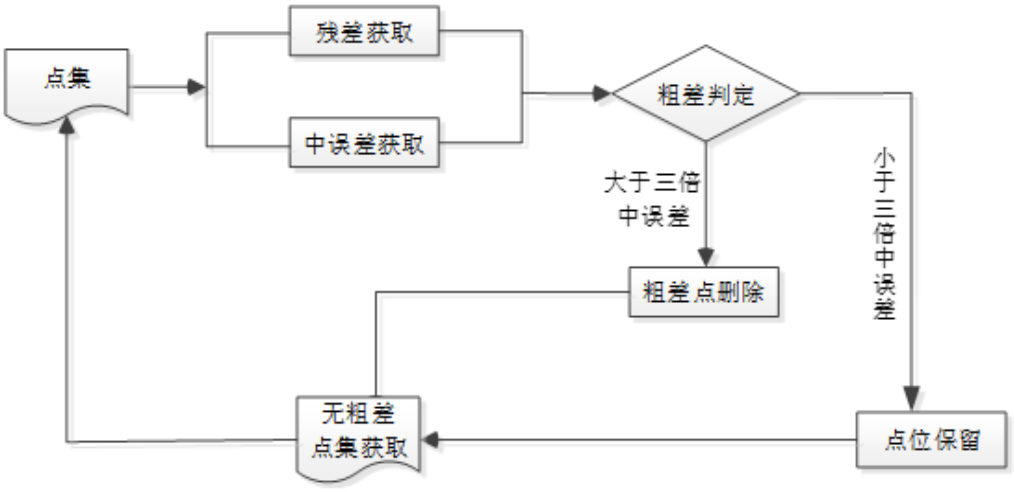


图 4.3 粗差剔除流程

计算残差代码如下：

```
//计算残差
double CalculateResRCToXY(double []r,double [] c, double[] x, double[] y)
{
    resx = new double[r.Length];
    resy = new double[r.Length];
    //残差数组
    double[] residual = new double[r.Length];
    for(int i=0;i<r.Length;i++)
     {
          resx[i]=this.m00+this.m10*r[i]+this.m20*c[i]+this.m30*r[i]*r[i]+
          this.m40*c[i]*c[i]+this.m50*r[i]*c[i]-x[i];
         resy[i] = this.n00 + this.n10 * r[i] + this.n20 * c[i] + this.n30 * r[i] * r[i] +
            this.n40 * c[i] * c[i] + this.n50 * r[i] * c[i] - y[i];
          residual[i] = resx[i] * resx[i] + resy[i] * resy[i];
```

```
    }
    double rms = 0;
    for (int i = 0; i < r.Length; i++)
    {
        rms += residual[i];
    }
    rms = rms / r.Length;
    rms = Math.Sqrt(rms);
    return rms;
}
```

计算中误差代码：

```
{
        //计算中误差
        public double GetRMSE(List<double> x1, List<double> x2)
        {
            double sum = 0;
            for (int i = 0; i < x1.Count; i++)
            {
                x1[i] = Math.Abs(x1[i]);
                x2[i] = Math.Abs(x2[i]);
            }
            for (int i = 0; i < x1.Count(); i++)
            {
                sum += (x1[i] - x2[i]) * (x1[i] - x2[i]);
            }
            return Math.Sqrt(sum / (x1.Count()));
        }
```

粗差剔除代码：

```
//粗差剔除
public void ErrorProcess()
{
    double rmse3x=this.calrmse.Get3RMSE(this.x_rec,this.x_normal);
    double rmse3y =this.calrmse.Get3RMSE(this.y_rec, this.y_normal);
    List<int> recordCount = new List<int>();
     bool flag = true;
    while (flag)
```

```
    {
        bool flag_x = false;
        bool flag_y = false;
        //X 方向遍历
        for (int i = 0; i < x_rec_afterdel.Count; i++)
        {
            double res = this.x_rec_afterdel[i] - this.x_normal_afterdel[i];
            if (Math.Abs(res) > rmse3x)
            {
                flag_x = true;
                recordCount.Add(i);
            }
        }
        //Y 方向遍历
        for (int i = 0; i < x_rec_afterdel.Count; i++)
        {
            double res = this.y_rec_afterdel[i] - this.y_normal_afterdel[i];
            if (Math.Abs(res) > rmse3y)
            {
                flag_y = true;
                recordCount.Add(i);
            }
        }
        //按照 recordCount 删除粗差点
        this.deleteByID(recordCount);
        recordCount.Clear();
        rmse3x = this.calrmse.Get3RMSE(this.x_rec_afterdel, this.x_normal_afterdel);
        rmse3y = this.calrmse.Get3RMSE(this.y_rec_afterdel, this.y_normal_afterdel);
        if (!flag_x    && !flag_y )
        {
            flag = false;
        }
    }
}
```

4.4 坐标变换计算

对于纠正单元内完成粗差点剔除后剩余的控制点，再一次采用二阶多项式变换模型计算出该纠正单元的变换系数，依据该纠正单元内的检核点集计算出该单元的最大、最小误差和中误差，并写入到数据库中。

按照以上方法对每一个纠正单元进行计算，最终得到各个单元的坐标变换系数与误差数据，统一写入到数据库，通过点群落入各自相应的纠正单元得到变换系数，最终实现点群位置纠正。

二阶多项式计算过程部分代码：

```
while (ii < 5)
{
    ii = ii + 1;
    for (int i = 0; i < pn; i++)
    {
        double rr = r[i];
        double cc = c[i];
        double xx = x[i];
        double yy = y[i];
        B.SetValue(1, 2 * i, 0);
        B.SetValue(rr, 2 * i, 1);
        B.SetValue(cc, 2 * i, 2);
        B.SetValue(rr*rr, 2 * i, 3);
        B.SetValue(cc*cc, 2 * i, 4);
        B.SetValue(rr*cc, 2 * i, 5);
        B.SetValue(0, 2 * i, 6);
        B.SetValue(0, 2 * i, 7);
        B.SetValue(0, 2 * i, 8);
        B.SetValue(0 * rr, 2 * i, 9);
        B.SetValue(0, 2 * i, 10);
        B.SetValue(0, 2 * i, 11);
        B.SetValue(0, 2 * i + 1, 0);
        B.SetValue(0, 2 * i + 1, 1);
        B.SetValue(0, 2 * i + 1, 2);
        B.SetValue(0, 2 * i + 1, 3);
        B.SetValue(0, 2 * i + 1, 4);
        B.SetValue(0, 2 * i + 1, 5);
```

```
            B.SetValue(1, 2 * i + 1, 6);
            B.SetValue(rr, 2 * i + 1, 7);
            B.SetValue(cc, 2 * i + 1, 8);
            B.SetValue(rr*rr, 2 * i + 1, 9);
            B.SetValue(cc*cc, 2 * i + 1, 10);
            B.SetValue(rr*cc, 2 * i + 1, 11);
             L.SetValue(xx - m00 - m10 * rr - m20 * cc-m30*rr*rr-m40*cc*cc-m50*rr*cc, 2
             * i, 0);
             L.SetValue(yy - n00 - n10 * rr - n20 * cc - n30 * rr * rr - n40 * cc * cc - n50 * rr
             * cc, 2 * i + 1, 0);
        }
    double[,] Nbb = matrix.MatrixMulti(matrix.MatrixTurn(B), B);
    double[,] NbbInv = matrix.MatrixInv(Nbb);
    double[,] temp = matrix.MatrixMulti(NbbInv, matrix.MatrixTurn(B));
    delta = matrix.MatrixMulti(temp, L);
    m00 = m00 + delta[0, 0];
    m10 = m10 + delta[1, 0];
    m20 = m20 + delta[2, 0];
    m30 = m30 + delta[3, 0];
    m40 = m40 + delta[4, 0];
    m50 = m50 + delta[5, 0];
    n00 = n00 + delta[6, 0];
    n10 = n10 + delta[7, 0];
    n20 = n20 + delta[8, 0];
    n30 = n30 + delta[9, 0];
    n40 = n40 + delta[10, 0];
    n50 = n50 + delta[11, 0];
}
```

4.5 实验结果

文本以上海地区为例，分别采集了两套该地区的售票点信息，其中数据源一中有售票点 518 处，数据源二中有售票点 923 处：

id	name	addrss	x	y	city	keyword	tel
6457558	民航售票处	金高路2200号	31.264552	121.60205	上海市	售票处	021-61017993
6457560	航空售票处	曹安公路4671号	31.278334	121.21291	上海市	售票处	021-59593317
6457562	航空售票处	浦城路172号	31.2318	121.50732	上海市	售票处	021-38870119
6457564	鹰航售票处	青年路48号	31.157794	121.35268	上海市	售票处	021-64787118
6457566	航空机票售票处	仙霞路88号	31.207384	121.39915	上海市	售票处	021-62700466
6457567	高智航空售票处	宜山路888	31.173862	121.39843	上海市	售票处	021-64854742
6457568	羽翔民航售票处	水城南路37号万科商务楼北楼	31.196422	121.38893	上海市	售票处	021-62783532
6457569	上海航空售票处	商城路660号	31.23269	121.51425	上海市	售票处	021-58793133
6457570	好乐航空售票处	浙江中路386号附近	31.238008	121.47362	上海市	售票处	13801648515
6457571	上海南站售票处	动力北三路	31.156408	121.42535	上海市	售票处	
6457588	火车票售票处	灵岩南路1138弄7	31.143074	121.49529	上海市	售票处	
6457589	冠华航空售票处	常德路940	31.236874	121.43738	上海市	售票处	021-56443604
6457590	安亭航空售票处	墨玉路59-1	31.29444	121.1591	上海市	售票处	
6457591	铁路上海站售票处	梅园路385号	31.24934	121.45303	上海市	售票处	021-63179090
6457592	东航佳鼎售票处	博乐南路125	31.37877	121.25686	上海市	售票处	021-69919868
6457593	上海站北广场售票处	上海市闸北区	31.25268	121.45218	上海市	售票处	021-95105105
6457594	上海星云航空售票处	西藏南路760号501室	31.216848	121.48029	上海市	售票处	021-63458009
6457595	JAL日本航空售票处	陆家嘴环路1000	31.242186	121.50289	上海市	售票处	
6457614	上海航空公司售票处	飞虹路646号	31.268414	121.50687	上海市	售票处	021-65628156
6457615	上海航空公司售票处	延安西路1371	31.210222	121.41132	上海市	售票处	021-62292929
6457616	上海航空公司售票处	延安西路2633	31.199332	121.38783	上海市	售票处	021-32230830
6457617	上海航空公司售票处	吴淞路205	31.249048	121.48494	上海市	售票处	021-65215990
6457618	上海航空公司售票处	安远路775-1	31.23683	121.42943	上海市	售票处	021-62327889
6457619	上海航空公司售票处	向城路58号	31.225816	121.52593	上海市	售票处	021-68406566
6457620	上航假期光复路售票处	光复路1	31.24235	121.46705	上海市	售票处	021-53550418
6457621	中国东方航空售票处	花园石桥路33	31.235692	121.49586	上海市	售票处	021-62062559
6457622	中国东方航空售票处	东方路135	31.241062	121.5163	上海市	售票处	021-68609888
6457630	中国东方航空售票处	乍浦路71号	31.24767	121.48311	上海市	售票处	
6457631	上海野生动物园售票处	南六公路178号	31.05577	121.71166	上海市	售票处	021-61180000
6457632	上海航空公司南汇售票处	城东路35	31.04545	121.75874	上海市	售票处	021-68018540
6457633	上海天龙外贸航空售票处	临平北路5弄1号	31.265224	121.48929	上海市	售票处	021-65628890
6457634	芷新长途汽车总站售票处	秣陵路79	31.24974	121.45346	上海市	售票处	
6457635	ANA全日空航空售票处	陆家嘴环路1000	31.242186	121.50289	上海市	售票处	
6457636	上海火车站第一自助售票处	秣陵路100号附近	31.250396	121.45105	上海市	售票处	021-63179090
6457637	上海火车站第二自助售票处	秣陵路203号天达商务中心	31.250162	121.45158	上海市	售票处	021-63179090
6457638	东美航空售票处NO.005	崂山路526	31.22985	121.51866	上海市	售票处	021-68868520
6457639	东方航空虹梅路售票处	虹梅路3297弄78A	31.190326	121.38153	上海市	售票处	021-64055037
6457654	中国东方航空奉贤售票处	南桥路303	30.917584	121.45034	上海市	售票处	021-57415113
6457655	中国东方航空打浦路售票处	打浦路38弄1～6号附近	31.206408	121.46532	上海市	售票处	021-51098208
6457656	上海航空公司广西路售票处	广西南路41号	31.2298	121.47712	上海市	售票处	021-63361666
6457657	中成航空服务公司航空售票处	宣桥北路358号4楼	31.24381	121.59166	上海市	售票处	021-58661283

数据源一

图 5.1 数据源一售票点

id	name	address	x	y	city	keyword	
4182178	售票亭	宜川路300-2 附近	31.262148	121.437947	上海市	售票处	
4182179	航空售票处	延长中路720	31.267026	121.446953	上海市	售票处	021-66610400
4182180	航空售票处	汉口路271号 附近	31.235744	121.485186	上海市	售票处	021-55666666
4182181	航空售票处(南桥路)	南桥路72号 附近	30.911812	121.456881	上海市	售票处	
4182182	火车票售票处	许昌路1312	31.26861	121.516035	上海市	售票处	021-65152222
4182183	航空售票处(吴淞路)	吴淞路207号 附近	31.247329	121.489427	上海市	售票处	021-63258877;
4182184	豫园售票处	安仁街132号豫园	31.226421	121.492554	上海市	售票处	
4182214	东方票务售票点	成都南路127-1 附近	31.221028	121.468119	上海市	售票处	
4182215	大观园售票处	金商公路701号	31.072365	120.908296	上海市	售票处	021-59262089
4182216	锦江乐园售票点	虹梅路201	31.140705	121.410911	上海市	售票处	021-54204956
4182217	南京路	汉口路581	31.233949	121.47957	上海市	售票处	021-63526805;
4182218	春申站售票处	站前路 附近	31.079199	121.354119	上海市	售票处	
4182219	火车票代售点(抚顺路)	抚顺路363弄4号103	31.278273	121.510944	上海市	售票处	021-65625336
4182220	火车票代售处	年家浜路149	31.114979	121.582368	上海市	售票处	021-68115001
4182221	火车票代售处(贵州路)	贵州路126	31.236373	121.476201	上海市	售票处	021-63220697
4182222	中国票务在线	宁海东路200号申鑫大厦8层	31.229204	121.481829	上海市	售票处	
4182223	上海航空公司售票处(延安西路)	延安西路1731-1	31.208442	121.416151	上海市	售票处	021-62292929
4182254	太阳岛售票处	沈太路2588	31.03526	121.090043	上海市	售票处	
4182255	铁路车票代售点(四平路)	四平路2559-1	31.296354	121.51343	上海市	售票处	021-65109910
4182256	中国东方航空售票处	秣陵路85号	31.248097	121.457676	上海市	售票处	
4182257	上海马戏城售票处	共和新路2266号	31.277981	121.451741	上海市	售票处	021-66300000;
4182258	火车票代售处(桂林路)	桂林路396号	31.168502	121.418108	上海市	售票处	021-54978233;
4182259	火车票代售处(桃园公寓西南)	川沙路5273	31.188226	121.699492	上海市	售票处	021-58989923
4182260	火车票代售处(东平南路)	东平南路568	30.725576	121.332997	上海市	售票处	021-67965356
4182261	国内航空机票售票点(桂林路)	桂林路100-2	31.161406	121.41936	上海市	售票处	
4182262	航空机票售票处(梅陇路)	梅陇路415号梅陇文化馆	31.138099	121.419122	上海市	售票处	
4182263	火车票代售处(中华路)	中华路1057-4号110室	31.214619	121.489638	上海市	售票处	021-63773558
4182293	火车票代售处(长宁路)	长宁路707-1	31.21973	121.422852	上海市	售票处	021-52390244
4182294	火车票代售处(曹杨路)	曹杨路1000-2	31.243114	121.409937	上海市	售票处	021-62543183
4182295	火车票代售处(峨海小区东南)	峨山路137	31.214669	121.528202	上海市	售票处	021-58756448
4182296	铁路车票代售点(龙吴路)	龙吴路400-23	31.160449	121.448656	上海市	售票处	021-54351869
4182297	火车票代售处(水城路)	水城路787号	31.213537	121.391662	上海市	售票处	021-64694408;
4182298	火车票代售处(灵石路)	灵石路606	31.282999	121.448163	上海市	售票处	021-66315019
4182299	上海航空公司售票处(四川北路)	四川北路2211号上海天鹅宾馆大堂左侧	31.268073	121.482053	上海市	售票处	021-56665666;
4182300	松江站售票处	人民南路1号	31.000093	121.228373	上海市	售票处	
4182301	火车票代售处	盈港路1580号	31.160727	121.096688	上海市	售票处	021-59200265
4182302	火车票代售处(内环高架路)	岚皋路10	31.246326	121.426491	上海市	售票处	021-62030399
4182333	火车票代售处	包头路375	31.317135	121.53834	上海市	售票处	021-65585891
4182334	火车票代售处(梅园五街坊东南)	张杨路1071号	31.231612	121.529035	上海市	售票处	021-68672066
4182335	火车票代售处(环庆中路)	环庆中路126号	31.23319	121.720818	上海市	售票处	021-58978146
4182336	火车票代售点	北青公路4652号 附近	31.195512	121.231586	上海市	售票处	

数据源二

图 5.2 数据源二售票点

经过自动匹配得到共指 POI 数据 175 条：

	数据源一				数据源二			
2	豫园售票处	安仁街132号豫园	31.226421	121.49255	豫园售票处	安仁街132号豫园	31.229022	121.487536
3	大观园售票处	金商公路701号	31.072365	120.9083	上海大观园售票处	金商公路701号	31.07473	120.904152
4	锦江乐园售票点	虹梅路201	31.140705	121.41091	锦江乐园售票处	虹梅路201	31.142558	121.406264
5	中国东方航空售票处	秣陵路85号	31.248097	121.45768	中国东方航空售票处上海	秣陵路85	31.249986	121.453064
6	上海马戏城售票处	共和新路2266号	31.277981	121.45174	上海马戏城售票处	共和新路2266号	31.27979	121.447128
7	中国东方航空售票处	中原路22-3	31.306627	121.53219	中国东方航空中原路售票	中原路22-3	31.30864	121.527824
8	火车票代售处	望园南路与南奉公路交叉口东	30.916863	121.48879	火车票代售处S28268	望园南路与南奉公路	30.91921	121.484032
9	火车票代售处	虹梅南路2099号	31.10402	121.43117	火车票代售处	虹梅南路2099号	31.105856	121.426408
10	火车票代售处	共富路385号	31.348754	121.42469	火车票代售处S28102	共富路385号	31.350528	121.420032
11	上海航空公司售票处	华山路1859号	31.198775	121.43701	上海航空公司华山路售票	华山路1859号	31.200556	121.432376
12	火车票代售处(桃浦路)	桃浦路177号	31.260638	121.39976	火车票代售处	桃浦路177	31.26247	121.394928
13	火车票代售处	电台路640	31.35739	121.39099	火车票代售处S28202	电台路640	31.359262	121.386376
14	火车票代售处	金高路1296弄83号	31.277602	121.6093	火车票代售处S28150	金高路1296弄83号	31.279796	121.605128
15	上海科技馆售票处	世纪大道2000号	31.217982	121.54109	上海科技馆售票处	世纪大道2000号	31.22017	121.53668
16	火车票代售处	前曹公路12	31.432522	121.30813	火车票代售处S28276	前曹公路12号	31.434576	121.303584
17	火车票代售处	碧波路573弄2号	31.202153	121.58341	火车票代售处S28153	碧波路573弄2号	31.20446	121.579224
18	火车票代售处	奉浦大道111	30.942086	121.45953	火车票代售处S28280	奉浦大道111	30.944024	121.454888
19	火车票代售处	金陵东路2号	31.231922	121.49212	火车票代售处S28026	金陵东路2号	31.233944	121.487968
20	火车票代售处	泰兴路218	31.230808	121.45895	火车票代售处S28072	泰兴路218号	31.232664	121.454416
21	中国东方航空(威海路售票	威海路258	31.22763	121.46601	中国东方航空威海路售票	威海路258号	31.22954	121.461464
22	火车票代售处	沪宜公路187号3号楼121A	31.325514	121.28794	火车票代售处S28158	沪宜公路1878号3楼	31.327572	121.283624
23	火车票代售处	星华公路1606	31.26968	121.29971	火车票代售处S28177	星华公路1606号	31.271614	121.295192
24	火车票代售处	沣坊路243号	31.154243	121.31929	火车票代售处S28277	沣坊路243号	31.156246	121.3148
25	铁路火车票代售处	肇嘉浜路873	31.197273	121.44596	铁路火车票代售处S28012	肇嘉浜路873号	31.199172	121.441304
26	火车票代售处	新泾路319号	30.888815	121.01272	火车票代售处S28216	新泾路319号	30.89096	121.008456
27	火车票代售处	长江南路490号	31.33578	121.48259	火车票代售处票S28139	长江南路490号	31.33761	121.478048
28	天蓝票务	国定路366	31.297651	121.50931	天蓝票务	国定路366号	31.299728	121.50488
29	上海海洋水族馆售票处	陆家嘴环路1388	31.240339	121.50172	售票处	陆家嘴环路1388	31.241126	121.496312
30	朱家角古镇旅游售票处	翔宁浜763号	31.11425	121.05369	朱家角古镇旅游区售票处	课植园路朱家角翔宁	31.116154	121.04924
31	火车票代售处	金运路523号	31.242151	121.32021	火车票代售处S28003	金运路523号	31.24401	121.315784
32	高智航空售票处	宜山路888号	附近	31.172095	高智航空售票处	宜山路888	31.173862	121.398432
33	高智航空售票处	宜山路888号新银大厦5层0505A	31.172251	121.40289	高智航空售票处	宜山路888	31.173862	121.398432
34	上海航空公司售票处(飞虹	飞虹路646	31.266412	121.51124	上海航空公司售票处	飞虹路646号	31.268414	121.506872
35	厦门航空(上海营业部)	陕西北路58-66号科恩国际中心	31.224604	121.45666	厦门航空上海营业部	陕西北路58	31.226458	121.452216
36	火车票代售处	胶州路1118弄17-5	31.239904	121.43431	火车票代售处S28292	胶州路1118弄17-5	31.241796	121.429632
37	火车票代售处	平型关路175号底层119室	31.267499	121.46733	火车票代售处S28149	平型关路175号底楼	31.269252	121.462792
38	火车票代售处	场中路3119丙	31.303892	121.43188	火车票代售处	场中路3119丙	31.30572	121.42724
39	火车票代售处(江川路)	江川路1550号	30.99505	121.37457	火车票代售处	江川路1550号	30.996	121.369976
40	火车票代售处	淞兴西路37号	31.369959	121.49253	火车票代售处	淞兴西路37号	31.37193	121.488056
41	火车票代售处	浦东大道1686	31.245186	121.54352	火车票代售处S28002	浦东大道1686	31.24728	121.539184
42	中国东方航空(商城路售票	商城路101	31.150583	121.11090	中国东方航空春浦商城路	商城路101	31.153486	121.115304

图 5.3 自动匹配结果

基于上述方法分别按八度、四度、二度、一度、三十分和十五分划分格网后计算得到的误差如表 5-1 所示：

表 5-1 误差统计表　　单位：km

精度评价 / 网格间隔	经度方向中误差	纬度方向中误差	经度方向最大误差	经度方向最小误差	纬度方向最大误差	纬度方向最小误差
八度	1.25312	0.07479	42.75098	0.19052	0.32527	0.000032
四度	0.88173	0.00717	1.93567	0.02346	0.03003	0.000053
二度	0.13623	0.00139	0.28953	0.00107	0.00447	0.00001
一度	0.01857	0.00001	0.03903	0.00006	0.00039	0.00001
三十分	0.00242	0.00001	0.00505	0.00001	0.00003	0.00001
十五分	0.00088	0.00001	0.00182	0.00001	0.00001	0.00001

误差走势图如下：

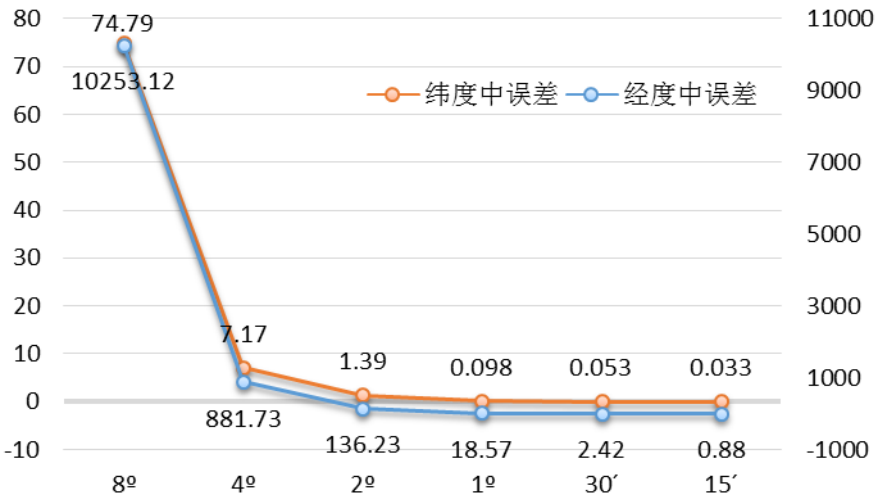


图 5.4 误差走势图

从以上计算结果可以看出，网格划分程度与误差大小成正比关系，划分程度越细，误差越小，在以十五分为单位划分地理网格的情况下，纠正精度达到了 1m 范围以内，较好的实现了多源地理点群的位置纠正。下面二张图为部分 POI 位置纠正前和纠正后的示意图，可以看出在位置纠正之前两套数据源存在着明显位置偏差，数据源二经过纠正后较好的与数据源一在位置上“吻合”，最终实现了两套 POI 数据的几何位置融合。

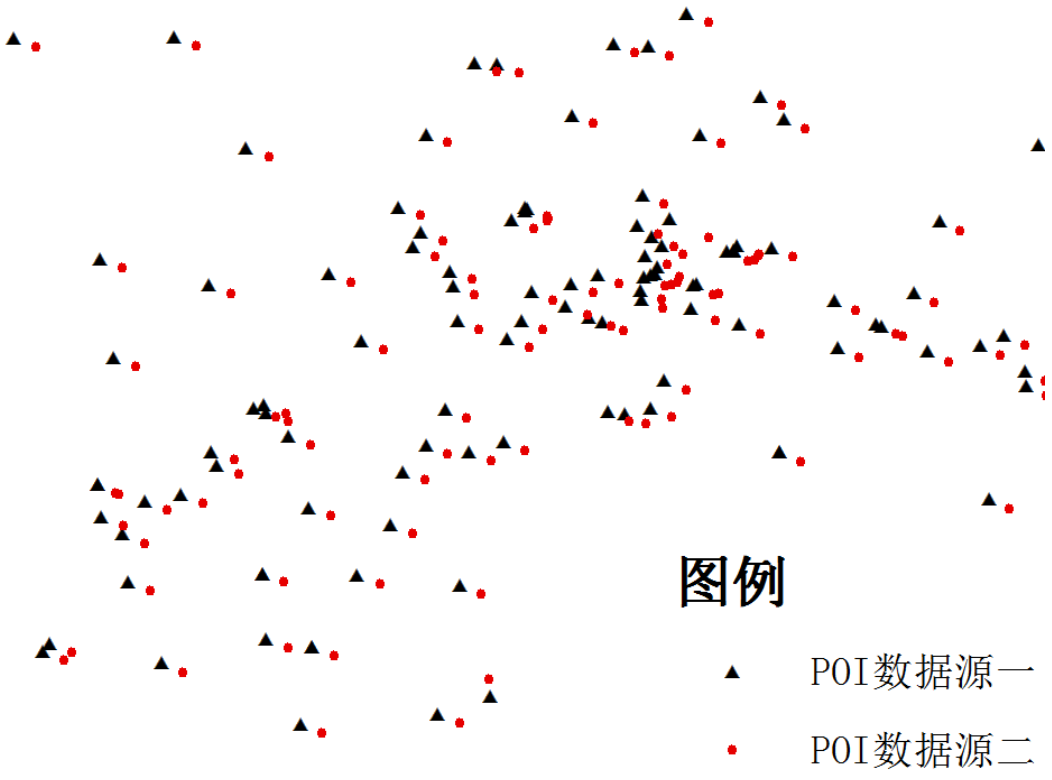


图 5.5 处理前示意图

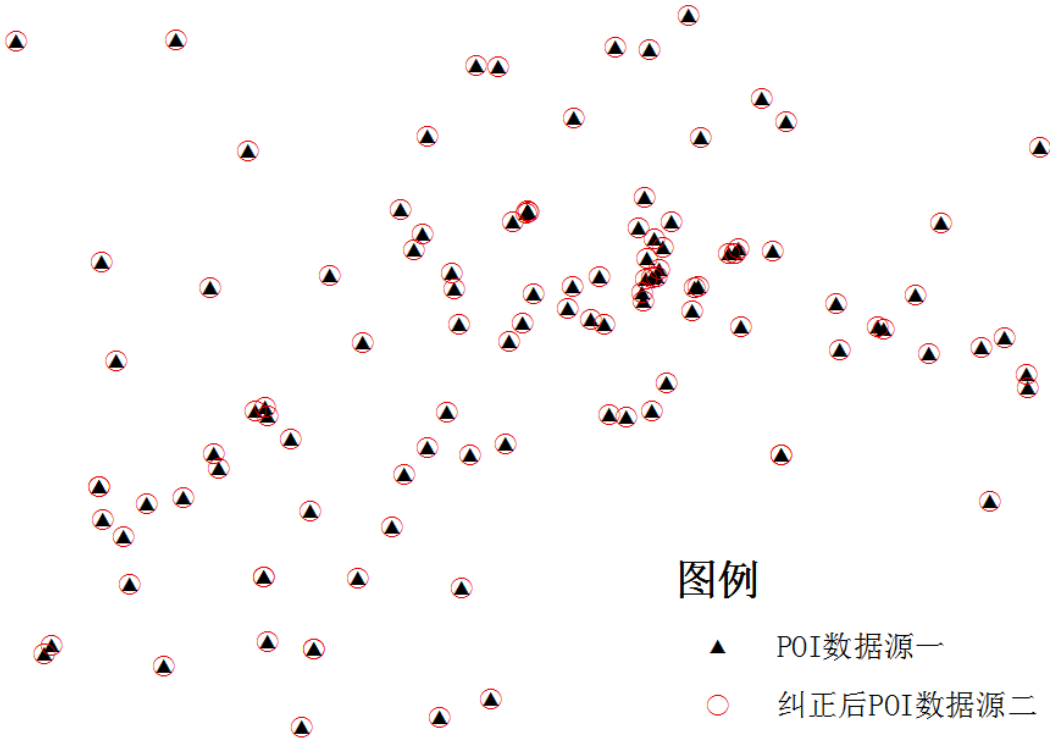


图 5.6 处理后效果图

5 总结

随着互联网地理信息系统的不断发展壮大，如何有效地获取和利用互联网上的各种地理数据资源成为越来越受关注的热点，多源数据融合技术更显得尤为重要。本文首先提出了一种自动提取与分析互联网地图 POI 数据的方法，针对以半结构化网页为载体的 POI 数据利用面向对象的编程方式进行了发现、提取以及存储；然后基于目前互联网电子地图兴趣点数据的分类方法建立了一套相对完善的 POI 内容分类体系，为多源 POI 数据的融合处理提供了有利的帮助；最后提出了基于网格划分自动提取控制点的多源 POI 位置纠正方法，通过划分地理目标区域和中文语义匹配的方式对多源 POI 数据的几何位置进行了纠正处理，通过实验验证，控制点的自动选取极大地提高了纠正效率，对地理目标的网格划分方法也保证了 POI 点群位置纠正的精度，对多源数据融合提供了支持。本文的创新点有：

（1）在研究各大互联网电子地图网站 POI 分类结构的基础上，构建了一套 POI 分类体系，为 POI 融合和集成共享提供了有效支撑；提出了一种基于互联网地图 POI 的自动提取分析方法，实现了从互联网快速获取批量、高时空性能的地理数据。

（2）将语义匹配与位置纠正的方法引入多源 POI 数据融合构建过程中，提出将语义匹配与位置纠正相结合的多源 POI 数据融合自动构建解决方案，实现了坐标控制点的自动提取，并利用网格划分法针对大范围内的 POI 目标进行空间位置套合。与现有的手动/半自动构建方案相比，该方法能够有效地利用语义知识和位置纠正方法自动进行 POI 数据融合。

由于本文针对空间位置对多源 POI 数据进行了匹配融合研究，只是把具有不同坐标标准的 POI 点的经纬度坐标统一了起来，而并没有涉及到属性描述信息方面的融合，例如兴趣点的电话、背景介绍等描述属性还没有参与融合研究，因此对 POI 数据的融合只研究了一半，这些都是在以后的学习和工作过程中需要研究和解决的问题。

致　　谢

时光飞逝，岁月如梭，转眼间三年的研究生学习生涯就要结束了，回想起来，在这期间我有着太多的经历，从一个专业基础薄弱的本科毕业生成长为一名对地图学与地理信息系统领域有一定认知的研究生，这期间不但有欢乐，也有汗水。在研究生学习生涯中，我不仅接受到了专业领域的培养和熏陶，而且还感受到了自己所在研究小组浓厚的学术研究氛围和生活中互帮互助的伟大友谊。在这里我要感谢在我研究生道路上不断帮助我前进的家人、老师和同学们，谢谢你们！

首先我要感谢我的父母，是你们的悉心培养造就了我的今天。

其次，我要衷心地感谢我的导师刘纪平研究员，刘老师严谨的治学态度和高尚的学术追求给我留下了深刻的印象，正因为您孜孜不倦地教导和培育，我才能够顺利地完成学业，刘老师不仅在学业上给我悉心的指导，在生活上也给了我很大帮助，在我遇到困难的时候总是给予我勇气和鼓励，在今后的学习和工作中我会以刘老师为榜样，不断向前。

感谢我的第二导师闫浩文院长，我从本科阶段就开始接受您的指导，闫老师在课堂授课过程中及其注重与同学们的交流，鼓励大家提出新思路、新方法，闫老师渊博的学识，一丝不苟的治学态度，对同学们创造性的思维培养模式都是我受益匪浅，您的教诲让我受益终生。

另外还要感谢在测绘研究院二年时间中伴随我学习和进步的课题组组长王勇老师，是您一次次专业并且详细的答疑解惑使我在很多方面都有所收获和提升，此外王老师对科学技术的崇高追究也感染了每一位小组成员。

感谢同一个课题研究小组的罗安、吴文毅、徐胜华、杨毅、纪莹莹、王克永、蔡地，感谢你们在我的科研和生活过程中帮我排忧解难。

特别要感谢我的女朋友游浩妍同学，你的支持和鼓励是我最大的动力，是你让我的生活充满了欢乐和希望，相信在未来的日子里我们一定能够幸福快乐的走下去。

感谢我身边的同学和朋友，感谢你们的关怀和帮助。

最后再一次衷心地感谢你们，谢谢！

参考文献

[1] Cobb M,Chung M,Foley H.A Rule-based Approach for the Conflation of Attributed Vector Data[J].Geoinformatica,1998,1:7-35.

[2] 张桥平.地图数据库实体匹配与合并技术研究[D]. 武汉:武汉大学博士学位论文.2002.

[3] 章莉萍,郭庆胜,孙艳.相邻比例尺地形图之间居民地要素匹配方法研究[J].武汉大学学报(信息科学版),2008,33(6):604-607.

[4] 唐文静.数字地图点状要素合并变换算法研究[J].系统仿真学报,2009,21(5):1399-1402.

[5] Doytsher Y,Filin S,Ezra E.Transformation of datasets in a linear-based map conflation framework[J].Surveying and Land Information Systems.2001,61(3):159-169.

[6] Hoseok Kang.Geometrically and Topographically Consistent Map Conflation for Federal and Local Government[J]. Journal of the Korean Geographical Society.2004,39(5):804-818.

[7] 郭庆胜,丁虹. 基于栅格数据的面状目标空间方向相似性研究[J]. 武汉大学学报(信息科学版),2004,39(5):604-607.

[8] Ashok S,Sharad S,Kevin C.A feature-based approach to conflation of geospatial sources[J]. International Journal of Geographical Information Science.2004,18(5):459-489.

[9] Deng S S,Tong X H. A New Least Squares Adjustment Method for Map Conflation[J].Geoinformatics. 2006,B4201-B4201.

[10] Song W,Haithcoat T L,J M.A snake-based approach for TIGER road data conflation[J].Cartography and Geographic Information Science,2006,33:287-298.

[11] Saalfeld A. Automated Map Conflation [D]. Washington DC, USA:The University of Maryland, 1993.

[12] Saalfeld A. Conflation: Automated Map Compilation [J]. International Journal of Geographical Information Systems (S1365-8816), 1988,2(3): 217-228.

[13] Beeri C, Kanza Y, Safra E, et al. Object Fusion in Geographic Information Systems[C]//Proceedings of the 30th VLDB Conference.2004: 816-827.

[14] 彭煜玮,彭智勇.空间数据融合技术的研究[J].计算机工程,2007,33(18):51-55.

[15] Frederico T，Fonseca M J,Egenhofer P A.Using Ontologies for Integrated Geographic Information Systems[J].Transactions in GIS 2002,6(3).

[16] Shahram R,Maria C,Dia A,Marcin P,Frederick E.A Knowledge-Based Multi-Agent System for Geospatial Data Conflation[J].Journal of Geographic Information and Decision Analysis.2002, 6(2):67-81.

[17] 郭黎.多源地理空间矢量数据融合理论与方法研究[D].郑州:解放军信息工程大学博士学位论文,2008.

[18] 罗安,王勇,张福浩等.基于角色标注的中文 POI 名称语义分类方法[J].测绘通报,2012,S1.

[19] 龙军.基于角色标注的中文 POI 名称匹配的研究及原型系统实现[D].重庆:西南大学硕士学位论文,2008.

[20] 郭黎,崔铁军，王玉海等.多源空间数据融合技术探讨[C]. 中国地理信息系统协会第九届年会

论文集,2005:848-851.
[21] 孔祥元.大地测量学基础.武汉:武汉大学出版社，2008.
[22] 崔铁军，郭黎.多源地理空间矢量数据集成与融合方法探讨[J].测绘科学技术学报,2007,24(1).
[23] 郭黎.空间矢量数据融合问题的研究[D].郑州:解放军信息工程大学硕士学位论文. 2004.
[24] 徐枫，邓敏，赵彬彬等.空间目标匹配方法的应用分析[J].地球科学信息学报,2009,11(5):657-663.
[25] 金博，史彦军，滕弘飞.基于语义理解的文本相似度算法[J].大连理工大学学报,2005,45(2):291-297.
[26] 韩月阳，邓世昆，贾时银等.基于字分类的中文分词的研究[J].计算机技术与发展,2011,21(7):29-31.
[27] 熊泉浩.中文分词现状及未来[J].科技广场,2009,11：222-225.
[28] Li, L. and Goodchild, M.F., 2010, Automatically and accurately matching objects in geospatial datasets. Theory, Data Handling and Modelling in GeoSpatial Information Science. (Hong Kong, 26-28 May, 2010).
[29] 丁士俊,张忠明.几种不同坐标变换方法问题的研究[J].四川测绘,2005,28(1):16-19.
[30] 杨恒山.地图数字化坐标变换模型的选择方法[J].湖南理工学院学报（自然科学版）,2005,18(4):66-68.
[31] 杨元喜，徐天河.不同坐标系综合变换法[J].武汉大学学报（信息科学版）,2001,26(6):509-513.
[32] 武汉大学测绘学院测量平差学科组.误差理论与测量平差基础.武汉大学出版社，2007.
[33] 张敏.多源矢量地图数据的集成与融合[D].哈尔滨工程大学硕士论文,2011.
[34] Marinos K, Margarita K. A method for the formalization and integration of geographical categorizations.International Journal of Geographical Information Science,2002,16(5):439-453P.
[35] 胡鹏等.地理信息系统教程.武汉大学出版社，2007.
[36] 徐健,方安,洪娜.一种基于词语相似度计算的本体映射方法[J].现在图书情报技术,2013,2:36-41.
[37] 曹恬,周丽,张国煊.一种基于词共现的文本相似度计算[J].计算机工程与科学,2007,29(3):52-53.
[38] 张贯虹,乌达巴拉,巩政.基于向量空间模型的网页文本句子对齐方法研究[C].第十一届全国人机语音通讯学术会议,西安,2011.
[39] 张晓雷.面向 Web 挖掘的主题网络爬虫的研究与实现[D].西安：西安电子科技大学硕士学位论文，2012.
[40] 李霖，王红，赵宁等.基于本体论的基础地理信息分类研究[J].地理信息世界，2004，2(6):21-25.
[41] 张玲.POI 的分类体系标准研究[J].测绘通报,2012,10:82-84.
[42] 国家测绘局.基础地理信息要素分类与代码. GB/T 13923-2006.

攻读学位期间的研究成果

攻读硕士学位期间发表的学术论文主要有：

《一种自动提取控制点的多源兴趣点位置纠正方法》，测绘科学 2014 年第 05 期，第一作者。

参与的项目：

参与国家 863 项目子课题“互联网地理空间信息探测发现与预警关键技术研究”。

附录 A POI 分类体系表

一级分类	二级分类	三级分类	四级分类
行政区域及其他区域	国家		
	一级行政区域	省、直辖市、自治区、行政区	
	二级行政区域	地区、地级市、自治州、盟	
	三级行政区域	县、县级市、市辖区、自治县、旗、林区、工农区、特区	
	四级行政区域	乡、镇、民族乡、苏木、街道办辖区	
	区域性群众自治组织辖区	村民委员会辖区、居民委员会辖区	
	非行政区域	矿区、农区、林区、牧区、渔区、工业区、开发区、边贸区、口岸、军事区、地片、居民小区	
居民点	城镇居民点	首都、首府	
	农村居民点	村、庄、屯、集、堡	
	工矿区居民点		
	农、林、牧场居民点	农场、林场、牧场	
交通运输设施	水上运输	泊地、海港、河港、船闸、渡口	
	公路运输	公路、长途汽车站、收费站、里程标志处	
	铁路运输	铁路、火车站	
	航空与管道运输	航空港、管道、管站、	
	城镇交通运输	道路、街巷、地铁、轻轨、公交站、停车场	

一级分类	二级分类	三级分类	四级分类
	交通运输附属设施	桥梁、隧道、道班、检查站、环岛、加油站、灯塔、导航台、公路出入口	
	其他	索道、扶梯	
水利、电力、通信设施	井		
	蓄水区	池塘、海塘、水库、人工湖、蓄洪区、泻洪区、灌区	
	排灌设施	灌溉渠、排水沟、渡槽、泵站、涵洞	
	提堰	海堤、河堤、湖堤、闸坝、拦河坝	
	运河		
	电力设施	输变电线路、发电站、输变电站	
	通信设施	通信线路、通信基站	
餐饮美食	中餐	家常菜、川菜、鲁菜、闽菜、湘菜、湖北菜、江浙菜、淮扬菜、粤菜、东北菜、徽菜、西北菜、云贵菜、素食清真、火锅、海鲜	
		北京菜	东来顺、爆肚王、北平楼、正院大宅门、京味面大王、京味楼、京味斋、民福局、小肠陈、大宅门、到家尝、翠满楼
		烤鸭	全聚德、便宜坊、红莲烤鸭店、鸭王、大董烤鸭店、天外天烤鸭店、金百万烤鸭店
		烧烤	

一级分类	二级分类	三级分类	四级分类
	西餐	披萨餐厅	必胜客、好伦哥、巴贝拉
		牛排馆	豪享来、豪客来、绿茵阁
		意大利菜、法国菜、德国菜、俄罗斯菜、拉美烧烤、中东料理	
	日本菜	日本料理、日式烧烤、寿司、日式自助	
	韩国料理	权金城、汉拿山	
	东南亚菜	泰国菜、越南菜、印度菜、菲律宾菜、印尼菜	
	快餐小吃	西式快餐、中式快餐、日式快餐、面食、粥店、肯德基、必胜客、永和大王、吉野家	
		老北京小吃	奶酪魏、爆肚满、都一处、左邻右舍褡裢火烧、稻香村、桂香村、门钉李、三元梅园、白魁老号饭庄、海碗居
	甜点冷饮	面包西点	稻香村、面包新语、味多美、好利来、元祖、贝尔多爸爸
		冷饮甜品	哈根达斯、DQ
	自助餐	金钱豹、阳光海岸、金汉斯、汉丽轩	
宾馆住宿	酒店	经济型酒店、豪华酒店、公寓式酒店、主题酒店、假日酒店	
		星级酒店	五星级、四星级、三星级、二星级、一星级

一级分类	二级分类	三级分类	四级分类
		快捷连锁酒店	如家快捷酒店、锦江之星、速 8、MOTEL168 连锁酒店、香格里拉、7 天假日
	公寓		
	宾馆		
	旅馆		
	招待所		
	青年旅社		
	度假村		
商场购物	商场	新世界百货、太平洋百货	
	超市	华联超市、家乐福超市、物美超市、美廉美超市、沃尔玛超市、好又多超市	
	电器	厨房电器、灯光照明、电动工具	
		家用电器	苏宁电器、国美电器、大中电器、卫浴电器
	市场	菜市场、花鸟市场、旧货市场、服装批发市场、小商品批发市场、农贸市场	
	商店	副食店、饰品店、便利店、文化用品店、烟酒专卖、化妆品店、保健品店、礼品店、母婴用品	
	数码电子	手机店、电脑城、数码产品店、软件店	
	家装	床上用品店、家具、建材、装饰城	
		家装连锁	好美家装潢建材、百安居、宜家家居

一级分类	二级分类	三级分类	四级分类
	汽车	汽车用品、汽车销售店、4s 店、汽配城、二手车交易市场	
	品牌专卖	品牌服饰、体育用品、专卖店	
休闲娱乐	酒吧		
	咖啡厅		
	娱乐中心	歌舞厅、夜总会、娱乐城、俱乐部、会所、迪厅、KTV 网吧、棋牌室、游乐场、电玩城	
	影剧院	电影院、剧院、戏院、音乐厅	
	体育场馆		
	休闲运动	高尔夫球场、保龄球馆、溜冰场、游泳馆、健身中心、垂钓、台球厅	
	洗浴按摩	洗浴城、桑拿、温泉、按摩推拿、海滨浴场、足疗	
	旅游景点	纪念地	陵园、纪念馆、纪念堂、古战场、寺、庙、教堂
		公园、广场、植物园、动物园、水族馆、自然保护区	
科研教育	科研机构	科研所、科研中心、科学院、设计院	
	高等教育	大学、学院、高等专科学校、民办高等院校、军校	
	初中等教育	中学、职高、中专、小学、涉外学校、私立学校	
	学前教育	幼儿园、托儿所	
	职业教育	高职、技校	
	特殊教育	培智、聋盲、工读学校	
	政治教育	党校、团校、政治学院、社会主义学院	

一级分类	二级分类	三级分类	四级分类
	业余教育	夜大、业余院校、进修院校、函授大学、老年大学、专修院校、成人学校	
	专业教育	体育学校、艺术学校	
	培训中心	驾校、干部培训学校	
医疗卫生	急救中心		
	综合医院	甲级医院、乙级医院、丙级医院	
	专科医院		
	妇幼保健		
	口腔医院		
	整形美容		
	康复中心	康复医院、康复俱乐部、康复中心	
	动物医院	宠物医院、宠物诊所、动物医疗保健中心	
	防疫控制	疾控中心、防疫站	
	保健院	保健院、保健中心	
	敬老院		
	疗养院	干休所、修养所(院)、护理院、护理中心	
	诊所	门诊部、卫生所	
	药店药房	医疗器械、同仁堂、金象大药店、连锁药店	
文化媒体	文化艺术	图书馆、博物馆、展览馆、会议中心、美术馆、科技馆	
	文化活动	青少年宫、文化馆、工人俱乐部、文化广场	
	新闻出版	报社、出版社、记者站、新闻中心、杂志社	
	广播电视	电视台、广播台、广播中心	
	影视	影视公司、电影制片厂、影视基地	

一级分类	二级分类	三级分类	四级分类
	文化用品	音像店、书店、文化办公用品	
	宗教	教堂、天主教、基督教、伊斯兰教	
	档案馆		
生活服务	日常服务	理发店、美容院、护理院、照相馆、婚纱摄影、大众浴室、图片社、干洗店、家电维修、婚姻介绍、典当铺、复印社、搬家公司	
	殡葬	殡葬服务、墓园服务	
	金融服务	银行、ATM、理财公司、期货公司、信用社、证券营业部、保险公司	
	宠物服务	宠物店、宠物医院、宠物寄养、宠物领养、宠物美容、宠物训练	
	企业服务	互联网、办公、保安公司、公关服务、广告公司、会计师事务所、写字楼、会展中心、评估服务	
	其他服务	翻译公司、回收公司、鉴定中心、打印服务社、设计、律师事务所、典当行	
	中介	房产中介、职业介绍所、婚姻介绍所、家政中心、留学中介、物业中介	
	汽车服务	汽摩维修、拖吊服务、汽车检验场、汽车护理中心、汽车美容中心、洗车场	
	邮政电信	邮电局、邮电所、刊发行、中国电信、中国联通、中国移动、电信营业厅、联通营业厅、移动营业厅	
单位机构	政府机构	人民政府、人大、政协、党委、街道委员会	

一级分类	二级分类	三级分类	四级分类
	公检法	公安局、警署、派出所、刑侦总队、交管局、交警队、劳教所、拘留所、戒毒所、检察院、司法局、法院、公证处	
	消防	消防局、消防支队、消防中队	
	机关单位	委员会、劳动局、工商局、税务局、财政局、测绘院、勘测院、规划院、海关缉私局、海关调查局、海关监管所(站)、海关办事处、出入境检验、教育局、教育委员会、所、站、队	
	涉外机构	大使馆、领事馆、国际组织在华机构、国外新闻机构	
	驻地机构	政府驻地办事处、机关驻地办事处	
	党派团体	民主党派、工会、共青团、妇联、侨联、工商联、残联、协会、红十字会、宗教团体、贸易促进委员会、福利会、科协、商会	
	福利机构	敬老院、福利院、基金会、希望工程	
	产权交易中心	产权交易所、资产调剂市场、承包市场、租赁市场	
	质量监督	质监局、药监局	
	企业单位	国有企业、国有股份企业、外资企业、合资企业、民营企业	