



中国海洋大学  
OCEAN UNIVERSITY OF CHINA

顺序号(硕): SS02160813

姓名: 王婷婷

学号: 21120233072

学院: 信息科学与工程学院

专业: [20085212]软件工程

# 硕士学位论文

MASTER DISSERTATION

论文题目: 基于位置与属性的多源 POI 数据融合的研究

英文题目: Multi-source POI Fusion Based on Geospatial  
and Natural Property

作者: 王婷婷

指导教师: 张巍

学位类别: 全日制专业学位

专业名称: 软件工程

研究方向: 现代软件工程学

2014 年 5 月 22 日

谨以此文献给尊敬的张巍副教授以及我亲爱的朋友和  
同学们！

-----王婷婷

# 基于位置与属性的多源 POI 数据融合的研究

学位论文答辩日期: 2011.5.22

指导教师签字: 张新

答辩委员会成员签字: 赵国良

徐建良

孙世科

曲国良


解翠

孙世科

陈高

## 独 创 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及取得的  
研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其  
他人已经发表或撰写过的研究成果，也不包含未获得  
(注：如没有其他需要特别声明的，本栏可空)或其他教育机构的学位或证书使  
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  
确的说明并表示谢意。

学位论文作者签名： 签字日期：2014 年 5 月 22 日

---

## 学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，有权保留并  
向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人  
授权学校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，可以采用  
影印、缩印或扫描等复制手段保存、汇编学位论文。同时授权中国科学技术信息  
研究所将本学位论文收录到《中国学位论文全文数据库》，并通过网络向社会公  
众提供信息服务。（保密的学位论文在解密后适用本授权书）

学位论文作者签名：

导师签字：



签字日期：2014 年 5 月 22 日

签字日期：      年      月      日

本论文得到山东省自然科学基金项目（No. ZR2012FM016）的资助

# 基于位置与属性的多源 POI 数据融合的研究<sup>1</sup>

## 摘 要

随着互联网技术的迅速发展以及网络电子地图的日益更新,以 POI 为代表的空间地理数据也出现了快速的增长。POI 即兴趣点,它一般包含名称、地址、经纬度、类别等信息。Web 中含有大量有价值的 POI 信息,这些 POI 信息对于人们的日常生活与工作具有重大的参考意义。POI 中的信息是需要不断丰富与完善的,而传统的数据完善方式主要是人工采集,这种费时费力且更新不够及时的采集方式,将逐步被取代。

另一方面,来源不同的网络电子地图所提供的数据内容可能存在差异,将这些存在差异的 POI 信息进行数据融合,形成一个更丰富的数据库,从而实现数据的有效复用,最终得到完善规整的数据。这方面的研究逐渐成为数据挖掘与融合中的重要问题。

本文对多源 POI 数据融合方面,在 POI 特征词的选择、POI 经纬度坐标统一、评论信息的添加融合、评论信息主题词的抽取等方面进行了一定研究工作,主要研究成果与工作如下:

(1) 网页数据抽取与坐标统一,主要通过网络爬虫抽取 POI 信息,然后将其通过网络电子地图扩大。POI 中的经纬度坐标信息来源于不同网络电子地图,同一实体在不同地图上的坐标不一致,对后续的 POI 融合工作造成了一定的影响。为解决这个问题,提出了统一经纬度坐标的方法。

(2) 通过分析 POI 字段中各属性的形式与特点,提出了 POI 融合的概念,主要包括空间地理信息、非空间属性两部分,其中的空间地理信息包括 POI 中的地址和经纬度,非空间属性主要是名称与评论信息。地址融合是通过地址切分等级然后通过短小字符串匹配实现,经纬度是通过计算两点间距离得到;名称则主要是一个别名化处理,评论信息的融合则主要是直接的添加。

(3) 通过对 POI 中评论信息的理解,提出了主题模型抽取(Topic Model)的概念,简单构建了一个基于分词的主题模型,构建过程中对文本文档进行了分词,为进一步做大规模的 Web 数据主题抽取提出了一个概念性解决方法。

---

<sup>1</sup>本文得到山东省自然科学基金 (No.ZR2012FM016)资助。

实验结果表明，本文提出的技术解决方法可以实现在基本无人工干预下自动、有效地完成 POI 数据的融合，并实现简单主题抽取，最终得到一个丰富完善的 Web 数据库。

**关键词：**POI 数据融合；地理信息；准确率；主题抽取

# **Multi-source POI Fusion Based On Geospatial and Natural Property<sup>2</sup>**

## **Abstract**

With the fast development of Internet technology and the growth of Web electronic maps, POI as the representative of geospatial data grow rapidly. POI is an abbreviation of Points of Interest, which includes title ,address, latitude and longitude, classification and so on. Web POI contains a lot of value information which is a great help to people's daily life and work. The POI information needs to be enriched and improved .The traditional way to get POI information mainly relies on manual. The method is laborious and time-consuming and it will be replaced someday.

In addition, there are differences between data, which are obtained from various electronic maps. And how to integrate and fuse these data from different electronic maps, making them together, getting a richer database and realizing the effective reuse of data, finally getting structured data ,which become an important problem in web data mining and fusing.

In this paper , we mainly do some research on the aspects of multi-source POI data fusion , including choosing the feature word of POI, unifying latitude and longitude, adding the comments, extracting the subject of the comments etc. The specific research and work are as follows :

(1)Web data extraction and unifying latitude and longitude. First, extract information such as title, address, latitude, longitude reviews, phone and other information of POI from some web sites, then expand the web database by the network of electronic maps on title. These latitude and longitude coordinates of the same entity are different as they are got from various electronic maps. Thus some negative impact will be caused. POI after fusion to work caused some impact. To resolve this problem, unifying latitude and longitude coordinates has be proposed.

---

<sup>2</sup> Supported by Natural Science Foundation of Shandong Province under Grant No.ZR2012FM016.



(2) Proposed fusion and integration of POI by analyzing the form and characteristics of each POI property. The main formal similarities are consists of two parts , one part is geographic information , the other is natural property. Geographic information include two parts of POI, address and coordinates; the natural property is also concludes two parts: title and comments. Address fusion is determined by calculating the similarity of two strings.The coordinate fusion is determined by calculating the distance between two points.Title fusion is mainly on aliasing and comments is mainly on addition.

(3) Proposed Topic Model extraction based on the understanding of the comments and build topic model through segmentation. The experiment process of building model use some segmentation. Build a good topic model and provide effective pretreatment for the next theme extraction from large-scale Web data.

Experimental results show that the proposed technical approach can complete fusion POI data automatically and efficiently , then we can get a rich Web database to further study.

**Keywords : POI Fusion ; Geographic Information ; Accuracy ; Topic Extraction**

# 目 录

<b>1 引言</b>	<b>1</b>
1.1 研究背景	1
1.2 国内外研究现状	2
1.3 本文主要研究内容	4
1.4 本文的组织结构	6
<b>2 实验数据的获取</b>	<b>8</b>
2.1 POI 数据的介绍	8
2.2 详细页面数据的抽取	9
2.3 地图数据的获取	11
2.3.1 谷歌地图 API 与地图模糊搜索	11
2.3.2 百度地图 API 与智能模糊搜索	13
2.3.3 Mapabc 地图 API 智能与模糊搜索	14
2.4 生成实验数据集	15
2.4.1 地理坐标统一问题	15
2.4.2 地图纠偏	17
2.4.3 代理服务的介绍与使用	20
2.4.4 生成实验数据集	21
2.5 本章小结	23
<b>3 POI 融合实现</b>	<b>24</b>
3.1 中文名称相似性	24
3.2 地理信息相似性	25
3.2.1 中文地址的相似度	26
3.2.2 空间地理信息相似度	30
3.3 各字段融合实现	31
3.3.1 名称相似度融合实现—名称别名化	31
3.3.2 地址相似度融合实现—地址规范化	32
3.3.3 评论信息的添加	33
3.4 本章小结	34
<b>4 基于 GibbsLDA++主题抽取原理与应用</b>	<b>35</b>
4.1 主题抽取介绍	35
4.2 Gibbs 抽样与 GibbsLDA++	36

4.2.1 Gibbs 抽样 .....	36
4.2.2 GibbsLDA++ .....	37
4.3 文本分词预处理 .....	40
4.4 实验结果与分析 .....	43
4.5 本章小结 .....	45
<b>5 总结与展望 .....</b>	<b>46</b>
5.1 总结 .....	46
5.2 展望 .....	46
5.2.1 存在问题 .....	46
5.2.2 未来工作安排规划 .....	47
<b>参考文献 .....</b>	<b>48</b>
<b>致谢 .....</b>	<b>51</b>
<b>个人简历 .....</b>	<b>52</b>

# 1 引言

## 1.1 研究背景

早在上世纪 70 年代就有人提出了数据融合<sup>[1]</sup> ( Data Fusion ) 的主张,但是这一主张在那时没有被人们重视。随着技术革命的产生和科学技术的飞速发展,在工业、商业、军事、互联网以及其他领域,人们都面临着海量的数据,人们不断采用新技术去理解、消化、解读、评价信息过载问题<sup>[2]</sup>,人们对数据融合的重要性也有了一个更深刻地理解。从数据校正与融合的研究内容以及活动来看,数据融合可以这样的释义,来自不同数据源的数据和信息,根据某一标准,结合成一个完整的数据,在此基础上提供信息给用户,满足用户的需求。简言之,数据融合是这样一个过程:通过对多个数据源里的信息进行的校正与整合,得到一个全面的信息,这个信息比任何一个单一数据源提供的信息都多。

另一方面,互联网技术的飞速发展,导致我们正处在一个大数据的信息时代,Web 互联网的内容快速增长,而用户对此的要求则是更加全面准确,这一状况势必会逐步引领信息融合的 Web 时代的到来。各种 Web 网站和网页数据不断涌现,Web 数据越来越复杂,越来越丰富,这些丰富的 Web 信息越来越多的受到人们的关注。如何挖掘与集成这些具有重要使用价值的 Web 数据,提高数据挖掘<sup>[3]</sup>与集成的质量变得越来越困难。另外,科技逐渐发展,在很多专业领域,各种专业数据已经由过去只有专业人士才能理解的情况变成大众可以灵活使用的现实,这在地理信息服务 ( GIS- Geographic Information service ) 中亦然。GIS 是一种基于计算机软硬件的空间信息分析、统计和处理的服务。

地理信息系统中有一个重要的词汇 POI<sup>[4]</sup> ( points of interest ),它是一个用来表示实体对象的常用术语,主要包括我们日常生活中经常用到的各种实体对象,它可以是一座政府工作大楼,一个旅游景点,一所学校,一个公交车站点,一个 ATM 提款机点,一个商铺等等。这些对象实体经常被看作一个点,一般一个对象实体包括四个方面的信息<sup>[5]</sup>:名称 ( Title )、地址 ( Address )、经度 ( Longitude )、纬度 ( Latitude ),同时还有可能具有电话、评价、等级、星级等信息。丰富而全面的 POI 信息是一个电子地图有利竞争的保证,也是其为用户提供高质量服务的保证。不断更新的 POI 资讯能够及时为用户提供路况介绍以及周边建筑的详细介绍。

绍,也能方便用户查询所要到达的地方的详尽信息,方便用户进行选择与规划。然而地理信息采集的传统方法是这样实现<sup>[6]</sup>的:采用精密的测绘仪器,通过人工的测量,得到一个 POI 点的纬度与经度等,然后测绘人员记录下来,整理之后反馈到电子地图。由此看来,传统的 POI 的采集是一个既浪费时间又浪费精力的工作,其采集的 POI 的质量与数量从某种方面来说关系着一个电子地图或者商业系统的未来。来源不同的 POI 数据包含的内容是不一样的,并且数据更新的速度也是不同的,这使得 POI 数据的准确性有待验证,将这些 POI 数据的内容进行一个简单的融合校正,提高数据的完整性与准确性,使 Web 上的 POI 内容更丰富、准确,最终实现 POI 信息的自动校正与更新,是一件值得深入研究的工作。

中国经济飞速发展的现代社会,人们持有的货币量在增长,同时人们的购买力也在增长。人们有了更多的意愿进行消费,且更关心美食、旅行、娱乐、运动健身等服务领域,在消费的过程中,不可避免地需要丰富的 POI 信息,例如商铺的名称、地址等信息,同样需要店铺简介、特色推荐、电话、环境、营业时间、公交查询、星级等信息。窝窝、58、美团等团购网站应运而生,现在全国各大城市,这些团购网站每天为用户提供着上述的大量信息,服务内容涵盖了吃穿住行,店铺数量巨大,店铺信息内容丰富,这些信息是良好的 POI 数据来源,如何利用数据挖掘的技术对这些不断更新的信息进行采集<sup>[7]</sup>,然后对数据校正、规整、融合,在地理信息服务与文本挖掘领域都有良好的研究意义。

综上所述,数据融合技术是解决 Web 空间数据准确性与简洁性的决定性手段之一,随着 POI 的信息在 Web 网页中的大量涌现,如何获得其中的有用信息,逐步提高 POI 数据的复用技术现在已经成为 POI 数据挖掘领域的一个热点话题。对采集的 Web 数据进行校正,融合,信息添加,最终得到一个全面丰富的 Web 数据,从而为用户提供更好的服务,在现实生活中同样具有重要的应用价值。

## 1.2 国内外研究现状

现在国内外对 POI 融合的研究包括两个方面,一个是非空间属性方法<sup>[8]</sup>,一个是空间位置方法。在 2004 年,C. Beerl 等人提出了一种基于空间位置的地理数据融合技术<sup>[9]</sup>。2006 年,E. Safra 等人通过 Google 地图与 Yahoo 地图抽取了 Soho, New-York 地区的全部的酒店信息,并在 C. Beerl 等人的基础上结合了 POI 的其他特征属性做融合实验<sup>[10]</sup>。后来,V. Sehal 等人利用地理实体的位置信息与其他

特征属性提出了一种地理实体识别技术,并对来源不同的阿富汗赫尔曼德省的重要目标做融合处理,取得了进展<sup>[11]</sup>。非空间属性方面的研究技术也日益成熟,而且有了良好的结果。但是只考虑非空间属性的话存在一定的问题,忽略了不同点之间经纬度坐标的不同,使空间数据的存在缺失了一定意义。

现在国内外对多源数据融合多有研究,并且主要是应用于军事领域与商业领域。在军事领域方面包括:多源图像复合、智能仪器导航系统、目标检测与跟踪、无人机驾驶、图像分析与处理、目标自动识别系统等等;在商业领域方面主要用于数据的互补与扩充。一般来说数据融合包含了多个过程,主要是特征选择、特征匹配、特征融合等步骤。本文研究的 POI 数据融合最重要的应用就是在地理信息服务的采集与更新过程。是在基于位置<sup>[12]</sup>与基于自然语言处理<sup>[13]</sup>的融合集寻找实验下进行的融合工作。

Web 网站与页面是一个丰富巨大的数据源,从 Web 上面得到数据主要包括数据获取、数据抽取和数据整合<sup>[14]</sup>三个过程。数据整合的过程也可以称作是数据提高融合的过程,其主要目的是提高数据的完整性,丰富数据,增加数据的间接性,提高数据的准确性。其中提高数据的准确性即为数据校正过程,这一过程通过对来源不同的数据源进行验证实现。提高数据的完整性即为数据融合实现,主要是将来源不同的数据源里的数据与信息合并到一起。数据融合,就是将来源不同的描述同一个现实世界的 POI 对象的内容合并成一条内容,另一方面,来自不同数据源的数据可能存在数据冲突,其最终目标是提高数据准确性与完整性,将 Web 上面丰富的数据融合,方便用户使用。

Web 网络的迅速发展以及地理信息系统被广泛应用,出现了多源空间数据,这不便于数据融合的实现,也为地理数据处理与共享带来了小的麻烦。不同的数据源,数据精度不同,加密程度不同,更新速度不同,因此同一个 POI 数据点表示方式也会不同。对同一个数据的更新以及校正工作也不同,数据采集的工作一般来说是由 POI 提供商完成的,而现在主要的采集方式仍然是依靠人力实现的,这种方式是一种纯人力的、效率低下、成本昂贵的采集方式,由于采集过程中主要是通过测绘与调查人员地毯式采集数据,因此数据更新的速度与质量在一定程度上受到人为的限制。

来源于 Web 的 POI 数据,可能是不真实的,电子地图商一般需要验证其准确性及有效性之后才可以发布其信息,但这些 POI 数据内容丰富,兼之有些 POI

点更新速率很快,仅依赖地毯式人工搜索,机械式人工校正与融合是很难实现的。为了解决上述出现的问题,我们采取空间位置属性与非空间位置属性相结合的方法,进行自动数据校正与融合。本文采用的 POI 数据融合实施具体如下:第一,通过网页爬虫工具得到基础数据,然后利用 POI 数据中的名称字段在百度、Google、Mapabc 等电子地图进行智能模糊搜索,并将搜索得到的数据与基础 POI 的名称、地址、经纬度等字段作对比,设置合适阈值,通过阈值的比较来判断是否为同一实体对象;然后根据不同来源的 POI 数据的这些字段的描绘确定该 POI 点是否真实存在;其次,观察该 POI 的有关点评更新时间、更新质量与速度、图片内容等其他的信息,进一步确定该 POI 的存在与否的正确性;最后实现数据的融合,将确定为真实存在的 POI 的不同来源信息合并到一起,若是相同则选择其一作为结果,不同则做融合,补充确实字段,完善评论信息等;最后通过主题抽取,将团购网站上面得到的评论信息采取主题抽取工作。

另外,在数据融合之前,我们对可以进行融合的 POI 数据做一次简单校正工作以提高融合实验的质量,将校正之后确认存在的 POI 信息做下一步融合工作,不存在的点,进行删除处理,这样就可以降低数据库中的冗余 POI 的信息量。本文工作的主要目的就是対 POI 数据进行融合,使之信息更加丰富全面,丰富并完善 POI 数据库。

### 1.3 本文主要研究内容

本文研究的主要内容如下:从多个数据源采集 POI 数据,进行数据校正,得到准确校正数据集,然后从校正数据集中找到可以融合<sup>[15]</sup>的数据对象;主要从名称、地址、经纬度等信息进行融合工作,并添加评论信息等,从而提高数据的准确性与完整性。最后对从团购网站采集的评论信息进行主题抽取,建立主题模型,得到正面或者负面主题信息,进行主题的分类,提高评论信息的可读性。

网络电子地图不同,来源不同,Web 数据或多或少会有差异,为了将不同出处的数据信息合并调整为一个整体,本文在国内外研究基础上提出一种技术解决方案,该方案使用了非空间属性方法与空间位置属性方法组合的技术来实现数据的融合。具体实现如下:首先从美团、口碑网、大众点评网站上面抽取 Web POI 数据,主要是地址、名称、评论、等级等信息,得到初步实验数据集,然后使用名称字段在地图上面模糊搜索,扩大数据源,然后使用空间位置属性方法(经纬

度相似度)排除部分对象,然后使用名称属性与地址属性相似度方法找出融合对应对象,接着对融合数据进行融合工作,包括名称的别名化,地址的标准化(补齐地址的各行政级别),经纬度的统一化,用户评论信息的添加等。最后用 POI 数据集合测试,实验结果表明,准确率、召回率等都有有效改进。

为实现 POI 数据的融合,本文首先对 POI 数据进行了简单校正,然后根据不同属性建立分类模型,具体实现如下:

#### POI 特征词的选择

POI 一般包括名称、地址、经纬度、类型等不同方面的字段信息,不同的字段代表一个 POI 实体的某个特点,其中名称(Title)是指一个实体对象的具体名称,不同 POI 的名称字段相似度可以通过字符串匹配算法实现融合;地址(Address)是指一个实体对象具体的地理信息位置,不同来源电子地图上,同一个 POI 点地址表述方式可能不同,可以通过地理编码实现地址的融合与校正;经纬度是 POI 点地理位置坐标。这三个字段信息都可以代表 POI 实体,可以用来作为实验数据特征值的选择。

#### POI 经纬度坐标统一

经纬度是一个实体对象的具体地理坐标,每一个实体对象在球面坐标系上都有自己的坐标,由于不同电子地图坐标系不同,加密也不同,带来的结果是各个电子地图之间表示同一地理实体的经纬度通常不同,这一问题可以通过电子地图接口实现坐标的统一。

#### 评论信息的添加融合

美团、大众点评网站等 Web 页面包括评论信息,这些信息具有重要的意义,评论<sup>[16]</sup>是由不同客户通过自己的亲身感受留下的,具有重要参考价值,这些信息可以看出一个店面的不同特点,例如环境等。将这些评论信息添加到地图数据中,可以扩大地图数据的内容,使地图数据具有更丰富的内容。

#### 评论信息主题词的抽取

评论信息一般是客户的主观评价,有正面评价,同时也有负面评价,每一段评论都可以通过分词<sup>[17]</sup>抽取一个主题词,概括这段评价到底是正面的还是负面的,并做一个统计,这样可以让其他用户看到是正面评价多还是负面评价多,方便用户可以做选择。



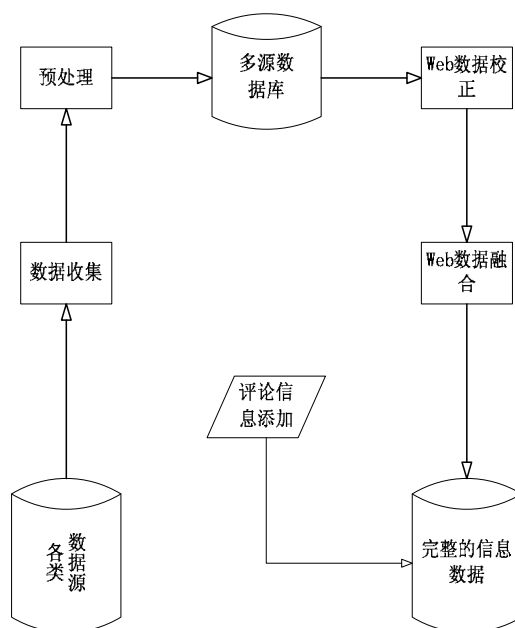


图 1-1 基于多数据源的信息融合流程

## 1.4 本文的组织结构

本论文的后续章节组织如下：

### 第二章 实验数据的获取

本章介绍了实验数据的获取信息，首先是详细页面数据的抽取，主要是大众点评网、美团网、口碑网等团购网站页面的数据，主要抽取店铺的名称、地址以及评论信息；其次是地图数据的获取，主要是将第一步获得的 Web 数据通过名称信息在百度，Google，Mapabc 等电子地图上模糊搜索，得到经纬度与地址等信息；第三步主要是实验数据集的生成，包括坐标转换等。

### 第三章 POI 的融合实现

本章主要介绍了可融合的实现，POI 融合是把多个数据源的数据整合成一条数据，这样数据融合包括地址的匹配融合，名称的校正融合，经纬度的融合，最后是评论信息的添加。

### 第四章 Topic Model 主题的抽取

本章首先介绍了 Topic Model 与 GibbsLDA++ 基本知识，其次是通过中文分词系统实现段落分词，得到分词数据，对分词数据进行实验，分析实验结果。

### 第五章 总结与展望

本章总结了 Web POI 可融合分类工作，Web 数据进行了融合工作，对采集

的评论信息做了主题抽取工作，然后讨论了主题抽取存在的不足之处，为下一步的继续研究打下基础。

## 2 实验数据的获取

### 2.1 POI 数据的介绍

POI, 即 Point of Interest, 中文译为兴趣点。一般情况下, POI 具备四个特征: 名称、经纬度、地址以及分类等。在地理信息系统中, POI 可以代表日常生活中常见的地理实体, 例如标志性建筑物、商店、火车站、中小学、体育馆、医院等等。本文实验中用到的 POI 数据主要使用了 POI 名称、地址以及经纬度等属性外, 还包括了邮编、电话号码、环境、营业时间、评论等其他更完善的特征信息。

ESRI ( Environment Ental Systems Research Institute ), 美国环境系统研究公司, 在全球各地都有办事处, 它是世界最大的地理信息系统技术提供商, ESRI 中国 ( 北京 ) 有限公司的公共地理信息服务平台是中国应用群体最多, 应用领域最广的, 因此本文采用该公司的 POI 分类系统。

本文中主要是对餐饮类的 POI 数据进行实验, 具体的行业分类及其分层与属性结构如下表所示, 由于本文使用的是餐饮类的, 就主要给出了餐饮类:

表 2-1 ESRI 公司给出的 POI 数据分类与代码

一级类		二级类	
一级代码	名称	二级代码	名称
01	餐饮	0101	快餐
		0102	西餐
		0103	清真
		0104	海鲜类饭店
		0105	烧烤类饭店
		0106	火锅类饭店
		0107	综合类饭店
		0108	特色饮食
		0109	咖啡茶馆
		0110	食品店
		0111	其他
02	购物		

03	住宿	
04	出行	
05	文体娱乐	
06	金融服务	
07	生活服务	

下面介绍 POI 数据结构的特点，如表 2-2 所示，从中可以可以看出一般应该具有的几个字段特点，并通过常见的限制设置了字段类型的长度，其大小可以进行更改。名称地址一般来说是必须的属性结构。

表 2-2 POI 数据图层属性结构：

字段名称	中文说明	字段类型（长度）	说明
*NAME	名称	TEXT(60)	
*TYPE	兴趣点一级分类	TEXT(20)	
*TYPE2	兴趣点二级分类	TEXT(20)	
ADDNAME	地址名称	TEXT(200)	详细地址
ADDCODE	地址编码	TEXT(30)	与地名地址库挂接
TELEPHONE	电话号码	TEXT(20)	
*PAC	所在行政区划代码	TEXT(20)	填至县区级，参考行政区划代码
DES	描述信息	TEXT(200)	该兴趣点文字描述信息
RelateID	关联 Table 表中 RelateID	LONG	作为外键，与关联表的数据进行关联。默认值为 0，当该 POI 存在对应的图片时，应赋唯一值。

## 2.2 详细页面数据的抽取

随着互联网技术的发展以及移动终端的发展，国内的团购网站技术不断完善，美团网，大众点评网，拉手网，滴答网，窝窝团购网，团 800，糯米网，58 团购等，本文我们选择了其中几个团购网站作为 POI 基础数据来源。另外，团购

网站包涵多方面类别（美食、购物、休闲娱乐、运动健身等），本文所用的 POI 数据为餐饮类，行业分类代码是 01。

作为最早一批团购网站，大众点评网于 2003 年 4 月在上海成立，它是第一个建立的独立第三方消费点评网站，可以为用户提供各种生活信息服务。目前，大众点评网在全国各大城市都有分支机构，拥有的点评数过百万。大众点评网不仅为用户提供各个店铺的全面信息，产品的点评内容以及团购优惠等服务，同时提供电子会员卡以及预约服务等信息。大众点评网一直致力于消费城市体验，在这里所有的评论都来源于之前用户，成功消费之后才可以进行点评，每个用户都可以对所消费的店铺里的产品进行评论，这样其他用户就可以根据已有评论的好坏做出自己的选择。

大众点评网站提供一个店铺的各种信息，包括店铺名称，地址，是否具有团购信息，支持交易类型，人均价格，人均打分，标签类型，特色特点，会员优惠，评价信息等。本文我们主要抽取店铺的名称、地址、评价等属性信息。

<b>粥全粥到(闽江路店)</b>   分店	<b>¥ 49</b>	8.1   7.8   7.8
地址: 市南区闽江路149号 85771568		 1293封点评
标签: 鲁菜 浮山所		
特色: 朋友聚餐 家庭聚会 随便吃吃		
团购: 粥全粥到!仅售42.5元,价值50元代金券1张!7店通用,无需预约,可累...		
会员卡: 凭会员卡可免费领取红枣汁一杯		
<hr/>		
<b>亨伯名家(崂山店)</b>   分店	<b>¥ 124</b>	8.8   8.8   8.7
地址: 崂山区香港东路87号建飞花园1期网点12号楼 88011080		 977封点评
标签: 韩国料理 麦岛大学区		
特色: 朋友聚餐 情侣约会 商务宴请		
团购: 亨伯名家(崂山店)!仅售128元,价值188元菌类火锅2-3人套餐!冬季来...		
订座: 本店支持在线订座,订座抽iPhone 5s!		
<hr/>		
<b>海岛渔村大酒店</b>	<b>¥ 94</b>	8.4   7.9   7.8
地址: 市南区云霄路40号 85973057		 2457封点评
标签: 海鲜 婚宴酒店 浮山所		
特色: 朋友聚餐 家庭聚会 可以刷卡		

图 2-1 大众点评网页面数据抽取

口碑网是阿里旗下的团购网站，同样为用户提供优质的生活服务，致力于成为全球最大的本地化生活社区，为用户提供方便快捷真实的生活服务信息。口碑网上面具有海量的商铺信息，消费信息，便捷的移动搜索服务，淘宝无线口碑卡

等应用。搜索应用是用来为用户提供搜索功能,可以让用户根据自己需求轻松地找到自己的需求信息,配合地图应用,使用户更方便的找到店铺信息。点评应用,是用户根据自己消费经验以及产品信息了解,对所消费产品的点评。这种点评应用很好地满足了消费客户对店铺急待了解的渴望。大量点评的累积,可以让店铺各种信息更加透明,更加丰富,让客户对店铺有一个全面感性的认识。

口碑网不同于大众点评的一个重要特点就是在主页面上面就有店铺图片介绍,同时对服务,口味,环境,性价比等不同类型进行顾客打分,最高 5 分,最低 1 分,等级分明。并对顾客的点评有一个概率统计,对好评率有一个良好的记录,方便其他顾客参考。



图 2-2 口碑网页面数据抽取

## 2.3 地图数据的获取

### 2.3.1 谷歌地图 API 与地图模糊搜索

Google 地图 API <sup>[18]</sup>是 Google 公司为电子地图开发推出的免费应用式编程接口,使用 Simple Object Access Protocol(SOAP,简单对象访问协议),用户注册之

后只需要使用网页编程语言 ( JavaScript、Perl、JSP、PHP、ColdFusion、Python 等 ) 调用 Google 地图<sup>[19]</sup>提供的 API 接口 , 可以根据 Google 搜索结果 , 开发自己的服务 , 进行文本数据挖掘等工作。Google 地图是一种典型的在线电子地图提供商 , 这种地图具有以下两种优势 : 一是用户可以自由使用 Google 的强大的后台服务 , 实现方法只需要调用 Google 的 API 即可 ; 二是用户注册之后可以根据自己的需要 , 将自定义的标签添加到 Google 的电子地图网页 , 开发自己的网络服务。

Google 地图目前开放的 API 已经超过几十种 , 这些 API 接口可以和 AJAX、JavaScript、XML、JSON 等技术相结合。它提供了地图标注、地图加载、实时搜索、街景视图、位置定位等功能。因为本文仅使用了地图搜索与显示、地图标注两个功能 , 所以本文先介绍两个 , 另外的地图功能的介绍可以通过 Google 地图的使用文档来参考。使用 Google 地图 API 实现的智能搜索结果举例如下 :



图 2-3 利用 Google 地图 API 根据名称自动搜索

上图表示的是通过名称属性在 Google 地图上面模糊搜索结果。从图中可以看出 , 通过 Google 地图可以在青岛市搜索到 “ 小倩倩馄饨 ” 13 家。上图中其中 search button 键可以用来单个搜索 , 在左端第一行空白文本框输入店铺名称 , 就会在页面显示搜索结果。Read button 键是用来读取 txt 文本文件 , 这个文件中可以存入多个店铺名称 , 以回车键分行 , 可以自动的搜索每一个店铺对应的所有信息 , 并且存储到文件中。

下面介绍介绍 Google 标注功能 , 如下图显示 :





图 2-4 利用 Google 地图 API 实现地点标注

注册自己的 Google 账号,然后登陆,按照 Google 地图 API 接口的提示可以进行店铺标注,将商铺信息标注到 Google 地图上,这样可以实时更新店铺信息,当某个点的 POI 对象发生变化的时候可以直接更新到 Google 地图数据库中。

### 2.3.2 百度地图 API 与智能模糊搜索

百度地图 API<sup>[20]</sup>是百度公司的地图应用接口,它同样是免费的,基础元素是百度地图,用户只要拥有百度账号就可以使用。分类有 Web 服务 API、车联网 API、静态图 API、URI API、Android SDK IOS SDK 等多种工具包。开发者可以根据自己的需要选择移动端开发或者服务器端开发,进而选择合适的接口,百度地图 API 提供主要功能有:基本地图点的搜索(包括普通搜索、视野内搜索、周边搜索)、公交查询(包括公交路线查询与公交方案查询)、路线规划与驾车导航、位置定位、三维街景显示、逆地址编码、地图标注与 LBS 云服务等功能。本文使用的是 JavaScript 语言编写的 API 接口,实现是通过 JavaScript,用户通过它可以使用地图的各种功能,自己构建内容丰富、功能强大的地图应用;百度地图还有一个优势,它隐藏封装了底层逻辑,以一种易于用户理解的方式发布,方便用户。图 2-5 是百度地图自动智能搜索显示结果:



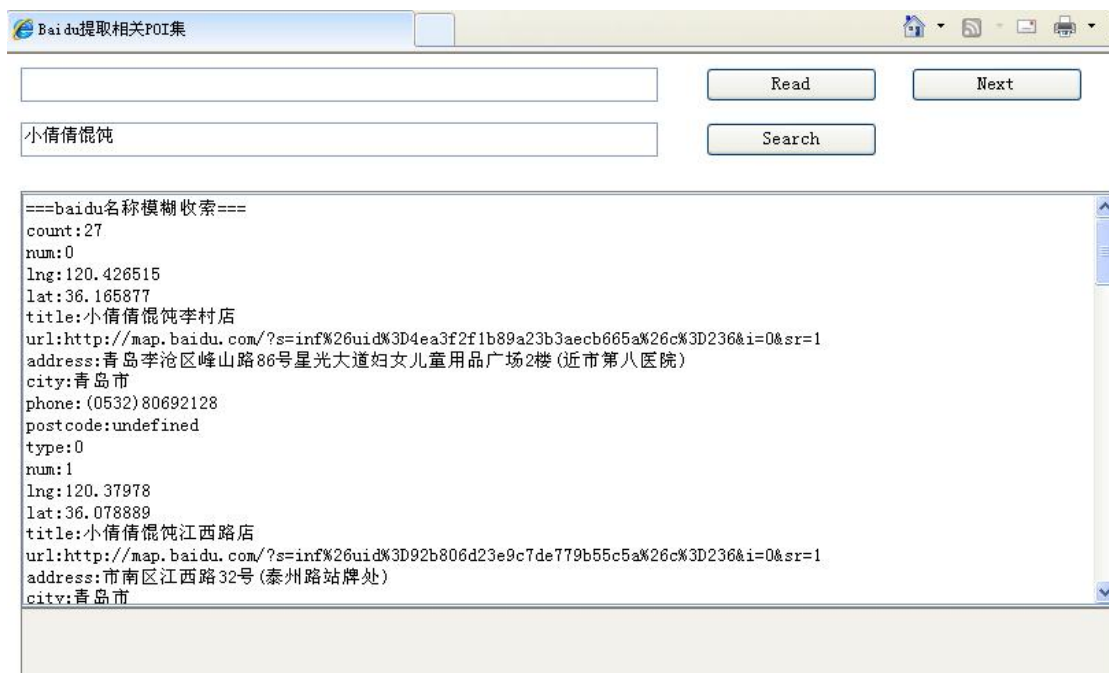


图 2-5 利用百度地图 API 根据名称自动搜索

上图表示的是通过名称属性在百度地图上面模糊搜索结果。可以看出，通过“小倩倩馄饨”这个名称可以在青岛市区搜索到 27 个结果。上图中其中 search button 键可以用来单个搜索，在左端第一行空白文本框输入店铺名称，就会在页面显示搜索结果。Read button 键是用来读取 txt 文本文件，这个文件中可以存入多个店铺名称，以回车键分行，可以自动的搜索每一个店铺对应的所有信息，并且存储到文件中。

### 2.3.3 Mapabc 地图 API 智能与模糊搜索

Mapabc<sup>[21]</sup>（北京图盟科技有限公司），是一家致力于地图服务的提供商，为用户提供基础地图服务，在有线与无线领域都有良好的表现。它致力于 Web 服务端地图服务、手机移动端地图服务和相关位置服务的等。Mapabc 主要提供以下服务：搜索服务、资源服务、导航服务、路径计算、地理编码、空间分析服务等，可构建各种地图服务运营平台，提供覆盖全国范围的地图搜索、位置定位、公交换乘、驾车路线导航以及多个城市的交通信息实时查询功能等，同时她与移动运营商合作，为用户提供手机定位以及无线地图服务，它的优势在于及时地应对大数据的访问。

Mapabc API 有移动版与网页版等不同版本。其中的地图 API 采用 Ajax 地图

显示，里面封装了 JavaScript 实现的 API，这样用户就可以在 HTML 中构建自己的 Ajax 地图应用；Mapabc 的地图搜索实现是基于 JavaScript 的计算接口，搜索结果能够显示在 Ajax 地图上面，但是用户想要使用则需要申请一个 Mapabc API 密钥，然后才可以开发自己的服务。类似地，下面给出通过 Mapabc 地图由名称字段智能搜索的示例结果：

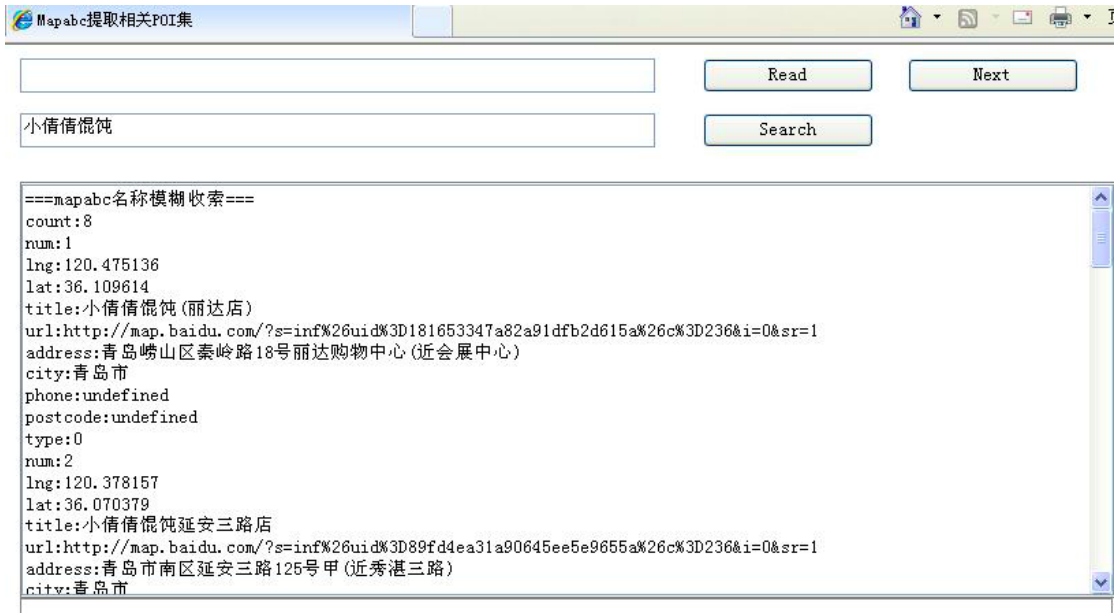


图 2-6 利用 Mapabc 地图 API 根据名称自动搜索

上图表示的是通过名称属性在 Mapabc 地图上面模糊搜索结果。可以看出，通过“小倩倩馄饨”这个名称可以在青岛市区搜索到 8 个结果。其中功能键的作用类似 Google 与百度，此处就不在复述了。

## 2.4 生成实验数据集

本文实验数据的初始来源是口碑网和大众点评网，通过网页爬虫在两个网站抽取了 3263 个 POI 点的数据信息，然后调用 Google、百度、Mapabc 三者的 API，通过名称字段完成搜索，得到 20145 条数据，去掉重复数据，选取其中 8000 条数据作为本文实验数据。

### 2.4.1 地理坐标统一问题

本文 POI 源的扩大实现是来源于百度、经纬度、Mapabc 等不同电子地图，这三个电子地图采用不同的坐标系，导致同一实体对象的经纬度坐标不同，这就是空间地理坐标不统一的问题。为了降低坐标问题对 POI 融合带来的影响，本文

选择进行一个坐标的统一转换。

下面先介绍下各个地图网站的基本情况，具体详情见下表：

表 2-3 国内常用地图的基本情况

地图 功能	Mapabc	百度地图	Google 地图	备注
地图接口				基本的地图操作，包含测距功能
搜索功能				模糊搜索功能，含周边查询
GPS 坐标标注	(收费)			通过 GPS 的 ID, 解析坐标并在地图上显示
地址解析和逆地址解析		(有限制)	(有限制)	地址转换成经纬度坐标信息和坐标信息转换成城市等地址信息
环境要求	IE 7+、FireFox 3+、Flash Player 10+	IE 6.0+、Firefox 3.0+、Opera 9.0+、Safari 3.0+、Chrome	IE 6.0+、Firefox 2.0+、Safari 3.1+	对于操作系统的要求不再给出，同时有些其他的要求譬如 Flash Player 的要求可能未给出
接口语言	JS、AS3	JS	JS、.NET、AS	由于各个提供者的描述标准不同，可能描述有偏差
更新周期	1-2 次/年	1-2 次/年	热点地区更新较快，其他地区较慢，1-2 次/年	地图更新取决于基础地图数据供应商的更新速度
“ ” 表示提供此服务，“ - ” 表示没有此服务，“ ” 表示情况不明				

从表 2-3 中我们可以看出 Google 等地图的更新周期、接口语言等基本信息，考察了各个地图的功能及环境条件，本文选则了 Google、百度、Mapab 三个，具体理由如下：

(1) 从表中可以看出，它们的 API 接口都是免费的，用户可以根据自己的需要选择功能调用。

(2) 它们的模糊搜索功能智能化较高，用户搜索体验较好，而这一功能是我们实现数据搜索必须经常使用的。

(3) 三者都有自己的数据更新周期，并且这一周期较为稳定，方便用户数据的采集与维护。

(4) 三者对环境要求相似，都能够加载在 IE 浏览器中使用，接口语言可以使用 JavaScript，在能够支持浏览器文档的 WebBrowser 控件类下进行集成开发，最终实现自动的数据采集<sup>[22]</sup>与抽取。

#### 2.4.2 地图纠偏

众所周知，我国法律规定，所有正式发行的地图类产品都必须进行强制性的加偏处理。也就是说，一个地点的真实坐标按照加密算法<sup>[23]</sup>进行加密处理，人为的加偏之后变成虚假的坐标，而这个加密方法并不是线性的，因此各个地点的偏移情况也会不同。包括 Google 地图在内的各个地图运营商都对自家的网络电子地图都进行了加偏处理。有些地图的运营商在第一次加偏的条件下进行了二次处理，这导致同一个点在不同的网络电子地图上经纬度坐标的不同。经纬度坐标的不同实际上为 POI 数据的融合带来了很大的困难，因此要继续进行 POI 数据融合必须进行坐标的统一，对加偏的坐标进行纠偏<sup>[24]</sup>工作。

本文主要使用了百度、Google、Mapabc 等网络电子地图，这几个电子地图的网络提供商都有开放的 API，提供坐标转换<sup>[25]</sup>的接口。因此，我们直接使用 API 接口，实现坐标的转换。具体实现的坐标转换如下：

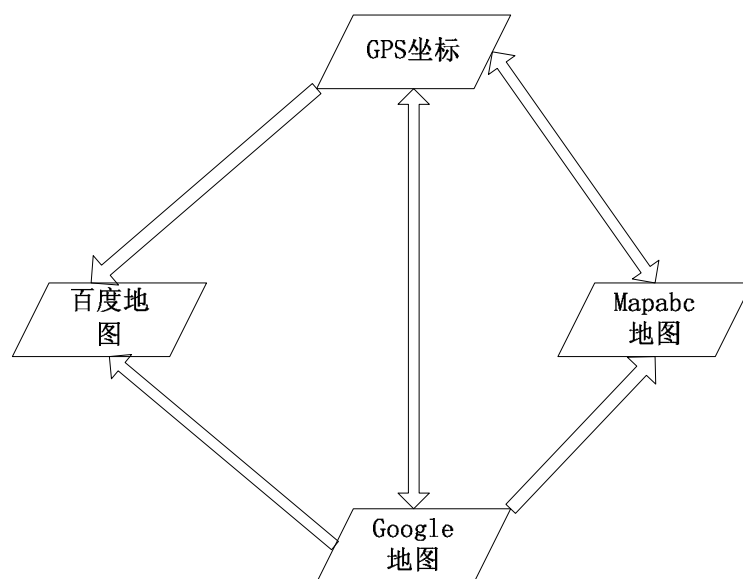


图 2-7 不同地图之间的坐标转换

上图显示了不同地图之间坐标转换的实现，其中双向箭头表示两者之间可以

互相转换。本文主要使用了百度，Google，mapabc 之间的地图转换，Google 地图到百度单组坐标转换可以使用以下接口：

```
http://api.map.baidu.com/ag/coord/convert?from=2&to=4  
&x=longitude&y=latitude
```

其中 longitude 代表 Google 电子地图上的经度，latitude 代表了纬度，数据返回值格式 Base64 码的，然后通过 Base64 编码解码工具对其进行解码操作。例如对谷歌地图坐标 (120.327904, 36.05323) 做处理进行转换：

在 IE 浏览器地址栏中书写下面这条网址：

```
http://api.map.baidu.com/ag/coord/convert?from=2&to=4  
&x=120.327904&y=36.05323
```

回车发现，接口返回值为：

```
{"error":0,"x":"MTIwLjM0NTY3MTIwNzMx","y":"MzYuMDkwOTEyMzk4ODYx"}
```

解码取小数点后六位)得到一组坐标值：(120.334422, 36.059192)

所以谷歌地图上坐标为 (120.327904, 36.05323) 的点坐标转换后在百度地图显示的值是(120.334422, 36.059192)。

这个转换接口适合单独一个坐标的转换，若是多个坐标再使用此接口则会速度慢且增加工作量，因此本文实验数据量较大，因此地图 API 的坐标转换函数是不错的选择 BMap.Convertor.translate(ggPoint, 0, translateOptions)；其中回调函数是 translateOptions()，坐标转换函数的使用方法可以参考 API 开发文档，转换结果示例如下图所示：

文件件名

dianping.txt

readfile

谷歌坐标转百度坐标结果:

translate

经度

120.345671

纬度

36.090912

id

1

转换后纬度

120.33924

转换后纬度

36.08475

1, 120.183695, 35.974783  
2, 120.18235, 35.971635  
3, 120.349144, 36.063828  
4, 120.363476, 36.088481  
5, 120.328015, 36.07824  
6, 120.325382, 36.07184  
7, 120.33706, 36.084365  
8, 120.336899, 36.084313  
9, 120.327934, 36.078322  
10, 120.376832, 36.093092  
11, 120.410184, 36.071204  
12, 120.397385, 36.072857  
13, 120.185231, 35.975133  
14, 120.360089, 36.089971  
15, 120.33924, 36.08475

图 2-8 同一地点 Google 坐标转换到百度坐标

上图是通过编码实现坐标转换的实验图，readfile 这个 Button 键是用来读取已经存储的坐标数据的文件，translate 是具体的坐标转换实现，按照顺序读取坐标之后，转换后的数据在右侧栏显示并且自动保存到文档。从上图中我们可以清楚地看出，同一个点的坐标在 Google 和百度电子地图上有不到一个分度的差值，而这种差值导致同一组经纬度在不同的地图上可能代表着不同的点，坐标转换的目的就是为了消除这种差异。为了完成经纬度数据融合实验，就需要解决坐标纠偏问题。

### 2.4.3 代理服务的介绍与使用

代理服务器<sup>[26, 27]</sup>, 用户由它可以获得额外的网络服务, 英文为 Proxy Server, 可以这样说, 它是具有缓冲功能的中转机构站, 客户端通过它获得服务器上的网络服务, 使用代理服务器客户端这边就可以突破自身 IP 的限制, 获得更多的网络服务。由于 Google、百度等地图 API 调用地址解析以及查询功能过程中, API 自动对用户调用次数进行了限制, 24 小时内使用过多的次数请求或者过快的使用频率都会被服务器限制, 这关系到本文数据采集能否正常进行, 因为在数据采集过程中要频繁的调用 API, 例如 3000 条数据可能会调用 API 次数过万, 为了使本文的数据采集工作顺利进行, 本文使用了代理服务器。

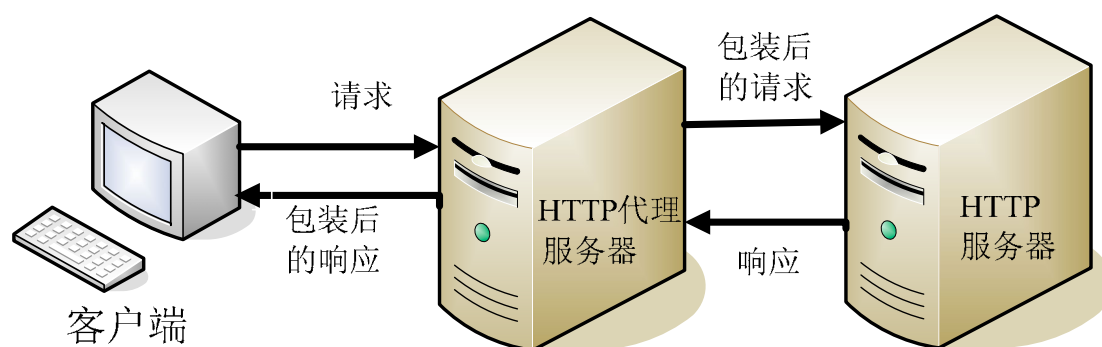


图 2-9 HTTP 代理服务器工作原理及过程

代理的基本工作原理是转发机制, 代理服务器位于远程服务器和本地客户端 (浏览器) 之间, 本文中所使用的是 HTTP 代理服务器, 顾名思义, 提供的是 HTTP 代理, 另外还有 socks 代理, VPN 代理。具体通信过程如下所述: 本地客户端 (Client) 与 HTTP 代理服务器之间的连接建立, 本地向代理发送包含 URL 地址的数据 Request; HTTP 代理接收此 Request, 读取此 URL 地址, 然后与 HTTP 远程服务器建立一个连接, 把此 Request 传递给 HTTP 服务器; 接收到网络上面的响应后代理服务器将数据下载到本地; HTTP 代理服务器将下载的数据发送回至本地。

通过使用代理服务器, 我们在对远程服务器进行访问的时候隐藏本地客户端的 IP 地址等具体信息, 这样就可以消除 API 接口使用过程中对 IP 数量的限制。例如, Google API 每天使用上限盈利性网站为 25000 次, 电子地图启动调用后,

Google 后台自动计算地图加载次数；加载地图后，用户对地图的基本操作例如平移，缩放不会增加次数，但是其他操作会增加，频繁的调用 API 超出限制会被禁用，因此我们使用代理隐藏 IP，实现 24 小时内的多次频繁的调用地图。

#### 2.4.4 生成实验数据集

本文首先是通过网页爬虫从团购网站采集基础信息，然后将其通过名称字段扩大到百度、Google、Mapabc 地图上，接着进行人工标注，以便和以后机器方法做对比，然后进行经纬度坐标的统一，最终生成实验数据集。接下来深入地按步骤介绍实验数据集的生成：

- 1、通过美团网，大众点评网，口碑网，糯米网等网络团购网站抽取 POI 数据，主要包括名称、地址、电话、点评、等级等信息，得到最初原始的 POI 集-source。

- 2、用百度地图、谷歌地图、Mapabc 地图提供的智能搜索功能对上一步得到的数据集进行搜索，搜索字段是用店铺名称，这样就在不同的电子地图上面得到扩大，原始数据集的信息更加丰富，增加了经纬度等地理信息；然后通过上文提到的坐标转换，将三个电子地图坐标统一到百度地图，得到 Web 数据集。

- 3、人工标注，对数据源信息进行人工标注，找出对应项，方便和以后机器实验得到的数据集进行对比。

- 4、准备开始实验。

综上所述，POI 实验数据集生成的过程可以归纳为：利用团购网站爬虫得到 POI 数据集，利用谷歌地图、百度地图、Mapabc 地图提供的搜索功能，搜索整个市区，扩大数据源；然后用数据库存储智能搜索得到的 POI 数据集，本实验是建立了具有合适的数据存储表格的 MySQL 数据库；再对数据源扩大得到的地图数据做数据去重以及经纬度坐标统一。图 2-8 是本文具体的 POI 数据集生成流程图：



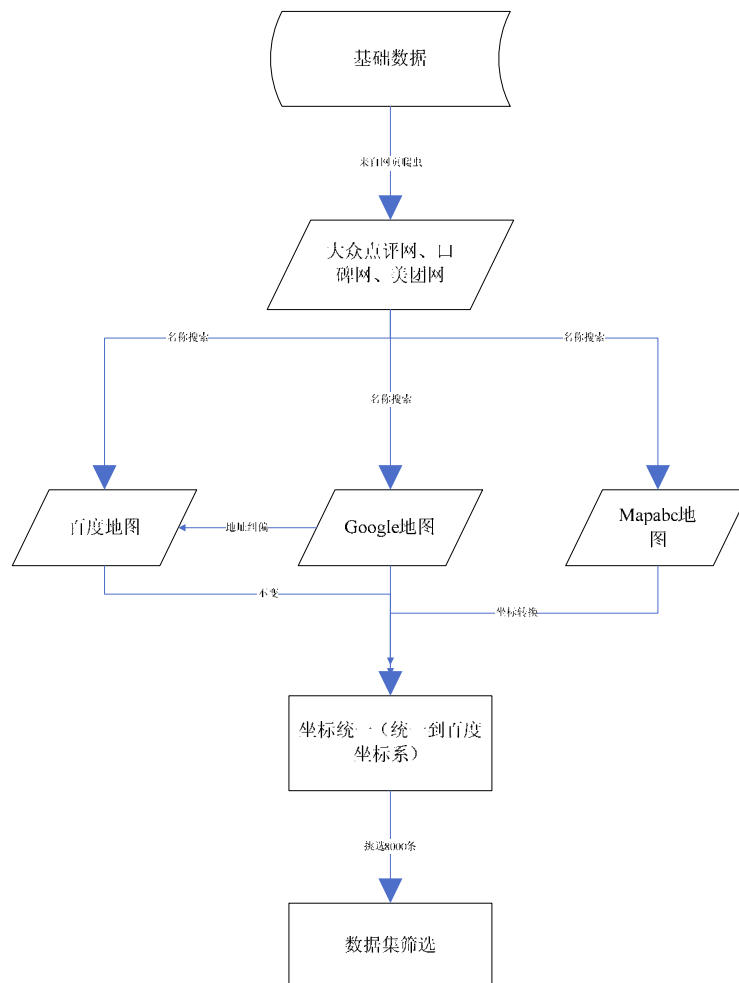


图 2-10 POI 数据集产生过程图

接下来给出原始 source 数据集通过经纬度坐标转换得到的结果：

序号	title	address	fit	tel	lon	lat	t_lon	t_lat	s_source
1	红荔村肠粉王(众孚店)	福田区福民路众孚新村3栋1号	1	83838880	114.051222	22.522514	114.057772	22.528204	58
2	香港新发烧腊茶餐厅(凤凰)	罗湖区凤凰路10号凤山大厦地	1	25417722	114.138832	22.547340	114.145282	22.553527	58
3	香港新文华茶餐厅	南山区文心五路33号海岸城西座120-122	1	86308558	113.937142	22.518304	113.943692	22.523963	58
4	永和快餐大王(新洲二街店)	福田区新洲二街73号南溪新苑1	1	88375471	114.046322	22.521326	114.052882	22.527064	58
5	汉堡王(KKMALL店)	罗湖区深南东路5016号京基·百纳空间购	1	82025622	114.106312	22.541093	114.112742	22.547348	58
6	艾薯(深大店)	南山区南海大道3688号深南大道深圳大	1	86310506	113.932822	22.539545	113.939362	22.545224	58
7	太兴餐厅(东海坊店)	福田区香林路东海花园二期东海坊步行	1	83076799	114.027692	22.538478	114.034172	22.544516	58
8	成都牛王庙面馆(报业店)	福田区景田东路1号景苑大厦1楼6	1	83069308	114.046832	22.542163	114.053372	22.547896	58
9	香港新发烧腊茶餐厅(春风)	罗湖区春风路2047号凯悦华庭1楼103-10	1	82140078	114.126392	22.540039	114.132812	22.546359	58
10	合味芳柳州螺蛳粉(华强总店)		0	82700300	114.012672	22.545575	114.019092	22.551843	58

图 2-11 POI 数据原始数据截图

Title 表示是名称字符段，address 表示地址，fit 表示人工标注（其中 1 表示匹配，0 表示不匹配），tel 表示电话，lon 表示原始纬度，lat 表示原始经度，t\_lon 表示转换后的纬度，t\_lat 表示转换后的经度，s\_source 表示数据来源（58 表示数据来源于 58 同城团购，还有糯米，美团，大众点评等）。

## 2.5 本章小结

本章首先对 POI 三个属性（名称、地址、经纬度）做了简单的介绍，并分析了数据来源网站大众点评网与口碑网等的优势，然后详细介绍了数据采集的全过程。首先是页面的抽取，然后扩大到数据源，接着进行了坐标的统一，并对坐标统一与纠偏做了一个简单的评测，最后实现了坐标的加载等工作。

### 3 POI 融合实现

更新、维护 POI 数据的过程中，我们一般会考虑 POI 数据的最直观的属性，名称字段属性与地址字段属性，字段描述相近的两个字符串表示同一个实体的概率比字符串相似度小的大，从这点上来说，字符串相似度大小可以直观的显示两个 POI 实体是否是同一实体对象。POI 名称字段融合实现主要是靠中文字符串实现的。本文 POI 融合实现主要指的是名称的别名化、地址标准化、经纬度的融合。

#### 3.1 中文名称相似性

POI 拥有名称(Title)、地址(Address)、经纬度坐标(Coordinate)等属性。其中名称字段是一个 POI 基本的属性，本文所抽取的 POI 中名称字段是一个店铺的名称，一般代表一个店铺的基本信息。

POI 中名称字段一般是比较短小的且无明显规律的字符串，例如“红舵坊码头火锅”、“好奇西式甜点”，这些字符串除了代表一个店铺的名称，无其他明显意义，缺乏一定的语义上特征。在数据挖掘、信息搜索、文本校对等领域，这种短小的字符串相似度有一定研究意义。判断两个中文字符串是否匹配的方法有很多种，普遍被使用的方法如下：Levenstein 距离算法、Jaro 距离算法和 Jaccard 相似方法。

根据现在已有的研究发现,在计算两个字符串相似度<sup>[28]</sup>的时候，常用的方法大多是使用单个汉字的，这忽略了汉语中词语对相似度的影响，而现实中我们可以了解到，汉语交流中很多词汇放在一起意义可能会发生变化，例如“这点店的蛋糕不好”，若是单个汉字匹配，“不”与“好”会分开，而分开之后显而易见，表达的意义就相反了，因此本文在计算中文字符串相似度的时候选择匹配的最小单位设定为词。

常见的中文字符串匹配算法有以下几种：Levenstein 距离<sup>[29]</sup>算法是一种用于计算字符串编辑距离的算法，它指的是将字符串  $s$  通过添加、删除、插入、替换等操作变成另一字符串  $t$  的最少操作频次，可以用来衡量字符串  $s$  与字符串  $t$  之间的相似性，用来检查整个文本的相似性。Levenstein 距离越大，字符串  $s$  与字符串

$t$  之间相似度越低即两者越接近不相似。

$$\text{公式定义为：} \quad \text{Edit}(S1, S2) = 1 - \frac{\text{distance}}{\text{MaxLen}} \quad (\text{式 3-1})$$

Edit 的值代表了两个字符串的相似度，Edit 值代表了两者间的相似性，Edit 越大则表示两字符串相似度大，最小为 0，表示两者完全不相似，最大值为 1 表示两者完全匹配。

Jaro 距离<sup>[30]</sup>算法同样可以用来计算字符串  $s$  与字符串  $t$  两者间的相似度，Jaro 距离定义如下：

$$d_j(s1, s2) = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3} \left( \frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{m} \right) & \text{otherwise} \end{cases} \quad (\text{式 3-2})$$

式中  $d_j$  是最后距离得分，代表着匹配度， $m$  则是匹配字符数，换位数是用  $t$  来表示的，它的值即为顺序不同的匹配字符的一半，

Jaro 距离算法里面有一个同样重要的公式，即匹配窗口计算公式，定义如下：

$$MW = \left( \frac{\text{Max}(|S1|, |S2|)}{2} \right) - 1 \quad (\text{式 3-3})$$

式中  $S1$  与  $S2$  代表两个等待判断相似性的字符串。当两个字符的距离不大于式 3-3 的最后结果(匹配窗口)即认为是匹配的。在计算  $S1$  与  $S2$  的相似度值时，若  $S1$  与  $S2$  拥有一样的字符  $x$ ，且  $x$  的距离小于等于匹配窗口  $MW$  的值时表示两个字符串  $S1$  与  $S2$  匹配。

Jaccard 系数<sup>[31]</sup>，又叫做 Jaccard 相似性系数，可以用来比较两个字符串的相似性，它是字符串差异性与分散性的一个统计概率。

字符串  $S1$ 、 $S2$ ，它们间的 Jaccard 系数定义公式如下：

$$\text{Jaccard} = \frac{|S1 \cap S2|}{|S1 \cup S2|} \quad (\text{式 3-4})$$

若  $S1$  与  $S2$  都为空，则定义 Jaccard 系数为 1。很明显，Jaccard 系数也是介于 0 和 1 之间，系数越大说明相似度越大。

### 3.2 地理信息相似性

广义上来说空间地理信息与非空间地理信息构成了地理信息的概念。狭义

上,空间地理信息就是指的是关于空间地理分布的信息。不管哪种定义下,POI 中的名称(Title)字符串都属于非空间信息,而 POI 中的经纬度则被看作是空间地理的代表。一个确定 POI 点的经纬度信息是一定的,但是由于不同的电子地图加偏方法的不导致 POI 经纬度标注的不一样。而一个 POI 的中文地址一般是确定的,因此我们通过中文地址字符串匹配方法来确定一个 POI 的地理信息的相似度。另外,有些地理 POI 点可能有多种描述的情况,例如处于交叉路口的点,有别名的点(举例说明,中国海洋大学崂山校区与中国海洋大学新校区就是同一个 POI 实体),这时通过地址匹配方法得到的结果可能会有偏差,这样本文通过结合经纬度等空间地理信息进行了全面的校正。

### 3.2.1 中文地址的相似度

中文地址<sup>[32]</sup>是一段有效字符串,具有一定的格式,包含国家、省市、县市、乡镇、街道、门牌号码等信息,常用自然语言来描述,它是空间位置信息的一个典型属性。中文地址的格式不是完全标准统一的,对于中国海洋大学崂山校区,在百度地图上地址为山东省青岛市崂山区松岭路 238 号,在 Google 地图上显示的一处地址为山东省青岛市崂山区 s214 与松岭路交汇处。对于这种情况,我们不再只使用字符串匹配方法,而是使用了地址分词,将地址按照行政区等级划分,计算等级划分之后各个字段的匹配度。

本文将行政区的省级、地级市、县、乡镇等字段作为特征词,基于它们对地址进行了分词,得到各个级别的元素,然后根据规则,考虑所有给出的元素,得到地址相似度。具体实现过程:首先根据中国行政区划表构造出词典,这样使地址中的省(自治区、直辖市)、市、县、乡更加规范,如“市南区鱼山路5号”,分词结果为“山东(省)青岛(市)市南区(区)鱼山(路)5(号)”,不仅划分出不同等级元素,其省略的等级元素也会附加上,这样等级元素就完整了。乡镇级行政区域下级的部分例如村、街道、建筑物等,分词过程中因为没有完整的划分方法,所以对这部分进行特征字分词。最终分词结果标准格式为“XX(省)XX(市)XX(县)XX(乡/镇)XX(路)XX(号)XX(其它)”,分词实现数据如下 3-1 举例:

原始数据	切分数据							其他
	省	地市	县/区	乡镇	村	路	号	
崂山区仙霞岭路17号金领世家南区14号楼1单元702室	山东省	青岛市	崂山区			仙霞岭路	17号	金领世家南区14号楼1单元702室
李沧区中崂路1036号甲	山东省	青岛市	李沧区			中崂路		
市南区彰化路2号海景花园大酒店新南楼B1	山东省	青岛市	市南区			彰化路	2号	海景花园大酒店新南楼B1
市南区东海中路30号银海大世界院内	山东省	青岛市	市南区			东海中路	30号	银海大世界院内
崂山区秦岭路18号丽达购物中心	山东省	青岛市	崂山区			秦岭路	18号	丽达购物中心
市北区山东路138号(家乐福北)	山东省	青岛市	市北区			山东路	138号	(家乐福北)
李沧区夏庄路1号伟东乐客城B1楼(近维客广场)	山东省	青岛市	李沧区			夏庄路		伟东乐客城B1楼(近维客广场)
崂山区仙霞岭路27号乐天玛特内	山东省	青岛市	崂山区			仙霞岭路	27号	乐天玛特内
城阳区正阳路136号家佳源3楼(青威路/正阳路路口)	山东省	青岛市	城阳区			正阳路	136号	家佳源3楼(青威路/正阳路路口)
崂山区仙霞岭路	山东省	青岛市	崂山区			仙霞岭路		

图 3-1 基于特征词/关键字的分词

从上图我们可以看出，原始数据是从县区级别开始，地址切分之后我们将省市等级补充完整，并将其余地址按照等级切分。我们采集的数据来源于大众点评网，从图中我们可以看出这些数据集中于青岛市各市区，因此在地址等级上面缺乏乡镇和村两个等级，但这并不影响我们做地址字符串的匹配。下文中将会对缺失某一等级地址的情况具体分析。

具体来说，对待两个待匹配中文地址，要计算它们的相似度，首先要对他们进行地址切分，将其分成标准等级格式的地址形式，第一级是省级、第二级则为市级，第三级则是县级，第四级为乡镇。分词过程中，对缺少的行政区字段进行添加，使地址更加完整与规范。一般来说，该四级中乡镇匹配，那么一般县级也就匹配，那么地市级一般会匹配，直到省级，反之未必成立。对于乡镇级别以下的字段，若比较地址 S1 与地址 S2，S1 地址中含有低于乡镇级别的字符段，而 S2 不包含，那么就不需要计算 S1 地址与 S2 地址的相似度，我们设定此类情形下相似度值为 -1，意思是该字段忽略不计。我们用 SIM、SIM1、SIM2 分别代表总的相似度、省市乡镇相似度、村路号建筑物相似度，那么总的相似度 SIM 是通过计算 SIM1 与 SIM2 来实现的，图 3-2、图 3-3 分别是 SIM1、SIM2 的计算流程图。

SIM 的具体计算过程如下：

初始化，设置 SIM、SIM1、SIM2 都等于 -1。

若乡镇级行政区字段可以匹配，那么令 SIM1 等于 1，转向 ；不匹配的话，则计算县级行政区字段的相似度，匹配则令 SIM1 等于 0.8，转向 ；类似地，计算地市级、省（自治区、直辖市、特别行政区）的字符段相似度，市级设置

SIM1 等于 0.4, 省级设置 SIM1 为 0.3 ; 否则 SIM1 仍等于 -1。

匹配路级别的字段的相似度  $t$  : 当  $t$  的值大于设定的阈值 0.8 时设置行政区中的路级字符串的相似度记为  $s_1$  ;  $t$  小于 0.8 则令  $s_1 = \frac{1}{2}t$  ; 否则令  $s_1$  等于 -1 。

类似步骤 , 将号码 ( 建筑级 ) 字段的相似度记作  $s_2$ 。

同样依照上面所说的中文字符串匹配算法 , 统计其余字符串的相似度记作  $s_3$ 。

设置 SIM2 中字段的权值 ,  $s_1$  的权值是  $x_1$  ,  $s_2$  的权值是  $x_2$  ,  $s_3$  权值则为  $x_3$  如果  $s_1$ 、 $s_2$ 、 $s_3$  三者中有一个的值为 -1 , 那么设定该字段对应权值就为 0 。乡镇以下低级字段的相似度计算公式如下 :

$$SIM2 = \begin{cases} -1 & , x_1 | x_2 | x_3 = 0 \\ \frac{\sum_{i=1}^3 x_i * s_i}{\sum_{i=1}^3 x_i} & , \text{其它} \end{cases} \quad (\text{式 3-5})$$

同样 SIM1、SIM2 字段对应的权值  $y_1, y_2$  ; 当 SIM1、SIM2 二者中有一个为 -1 , 那么相应的权值就为 0。两个字符串总的匹配度计算公式如下 :

$$SIM = \begin{cases} 0 & , y_1 \text{ and } y_2 = 0 \\ \frac{y_1 * SIM1 + y_2 * SIM2}{y_1 + y_2} & , \text{其它} \end{cases} \quad (\text{式 3-6})$$

如果 SIM1、SIM2 值都为 -1 , 那么设定 SIM 为 0。

通过实验结果比较分析 , 设定的阈值对实验结果有一定影响 , 同样的两个字符串在不同阈值条件下计算出的相似度会有差别。经过多次实验 , 我们选择了合适的阈值 , 经过阈值选择 , 我们设定  $x_1, x_2, x_3$  的值分别为 4、3、3。对于两个字符串地址 , 若是低于乡镇级的字符串匹配 , 例如村街道匹配 , 那么一般来说这两个地址是匹配的。若是考虑乡镇级别匹配 , 而低于乡镇级别的未知 , 或者只有一方已知的情况 , 那么两个地址就不一定匹配 , 县区 , 市级 , 省类似。由此可以看出 , 乡镇级别后 5 个字段的相似度对于总的相似度影响要大些 , 那么阈值的影响作用会更大些 , 经过多次实验 , 设置  $y_1, y_2$  的权值分别等于 1、3 , 效果更好些。

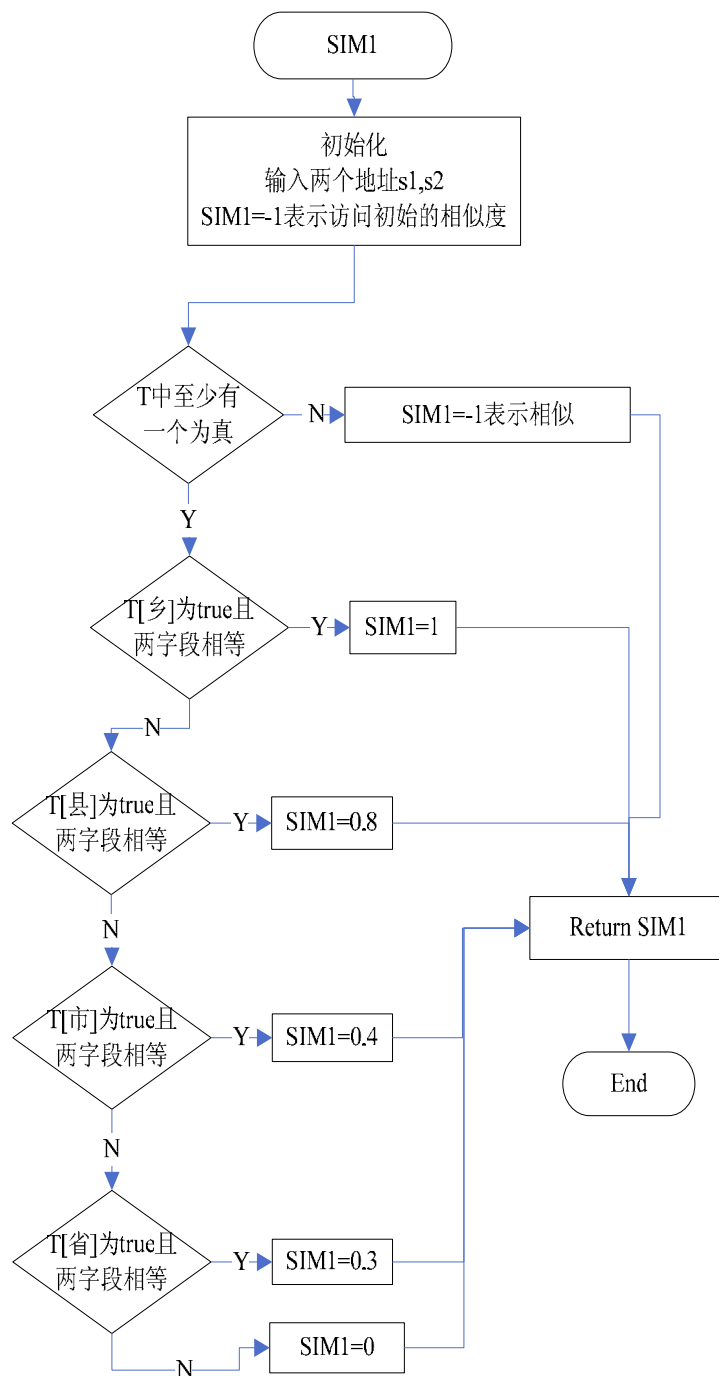


图 3-2 基于省、市、县、乡级别的相似度  $SIM1$  的计算流程图



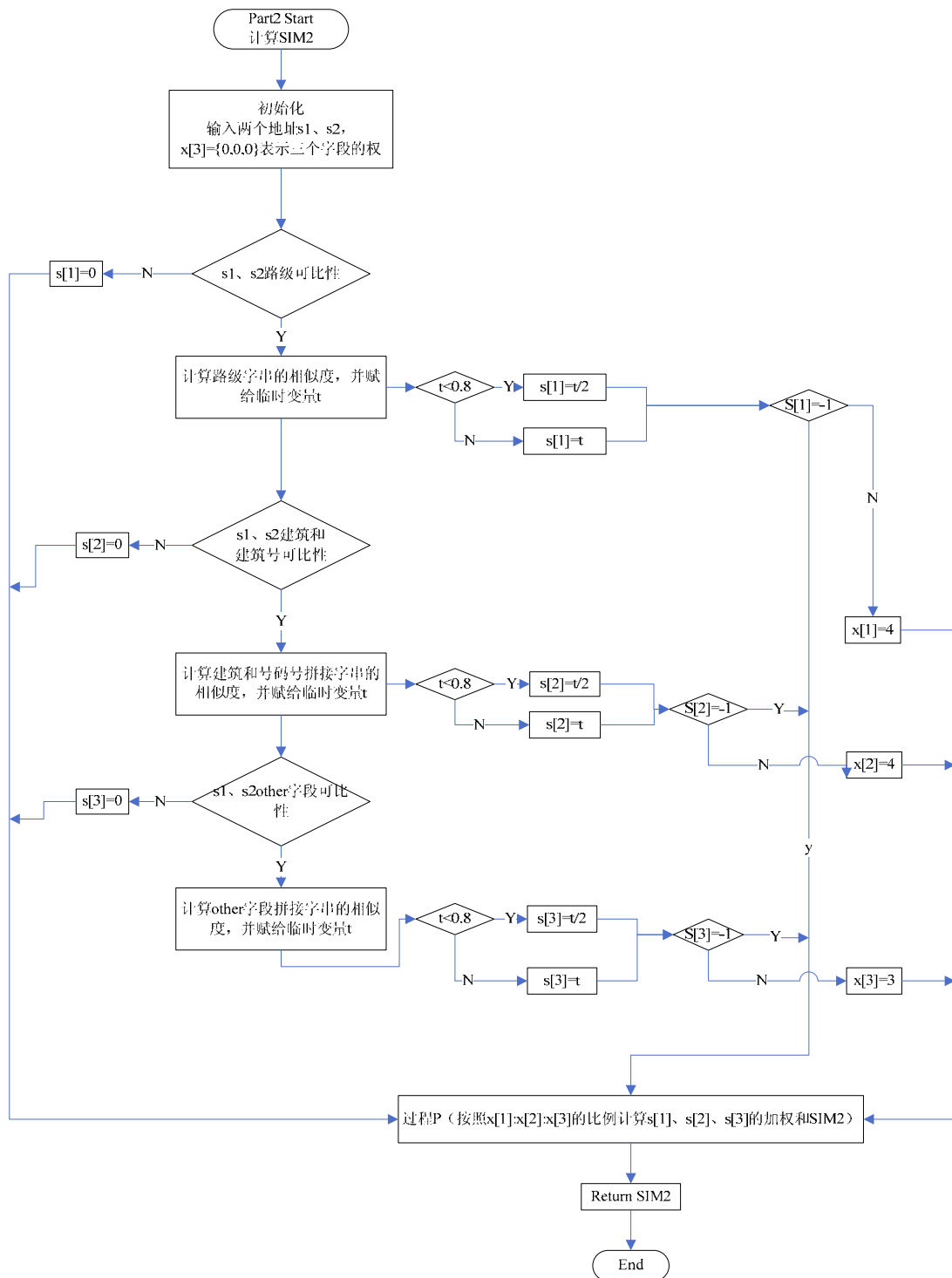


图 3-3 基于街道及号、建筑及号码的相似度  $SIM2$  的计算的流程图

### 3.2.2 空间地理信息相似度

空间地理相似性<sup>[33]</sup>是进行 POI 校正、融合、分类的关键, 空间地理相似性在 Web 空间数据中起着重要作用, 在地理信息融合与检索中有着深刻研究。空间地理数据具有经纬度、位置、维数等特征, 其中经纬度最能代表一个 POI 点的

属性特征。

经纬度是地理坐标系统的一个组成，它被用来定义地球上的空间球面，可以明确地标注地球上的每一个点的位置。POI 中的地址字段是通过自然语言来描述一个 POI 点的位置，经纬度是用二维数组来描述 POI 点的具体位置。计算两个 POI 点的经纬度相似性，最简便和有成效的方法使统计两点之间的球面距离，即这两点在球面上的最短距离<sup>[34]</sup>。该地理坐标相似度定义为：

$$\text{Sim}_{\text{lat-lon}} = \frac{1}{\text{dis tan ce}(S1,S2)} \quad (\text{式 3-7})$$

其中  $\text{dis tan ce}(S1,S2)$  是匹配的两个 POI 点  $S1$ 、 $S2$  的球面距离。当  $\text{Sim}_{\text{lat-lon}}$  的值大于设定的阈值，就可以判断  $S1$  与  $S2$  匹配。

### 3.3 各字段融合实现

对于上文进行的 POI 数据校正，得到一个可以匹配的校正数据集，然后对于校正集里面的数据进行融合处理。主要包括 3 个部分：名称字段属性相似度融合，地址相似度融合，评论信息添加等。本文中统计实验结果采用的方法是标注地计算准确率、召回率以及 F1 值，将它们的值作为实验结果的衡量。名称字段的匹配采用不同的字符串匹配算法，统计两个名称字段的相似度，然后根据设定的阈值，将可以融合的名称字段选择出来，做进一步的融合，包括名称的别名化；地址相似度的实现是将两个来源不同的地址进行标准化处理；经纬度的融合实现，则是对纠偏校正统一到一个标准的经纬度做融合处理。下文中将具体介绍各字段融合实现的具体方法。

#### 3.3.1 名称相似度融合实现——名称别名化

在电子地图中一个 POI 点可能具有不同的表示方法，例如青岛市松岭路中国海洋大学，在 A 电子地图上面显示结果为海大崂山校区，在 B 电子地图上则为中国海洋大学，在 C 电子地图可能显示为中国海洋大学崂山校区。这样同样一个地方，不同的表示方法，为了尽量避免这一现象的出现。我们对实验数据名称字段进行了融合实现，主要是名称字段的别名化。具体实现如下：取出百度，Google，Mapabc 得到的数据集，对数据集的名称字段进行字符串匹配，若字符

串匹配度为 1，则不做处理。

The Diner	31	36.071	120.391	东海西路35号	1	The Diner(东海西路店)	21	36.0706	120.392	市南区东海西路35号太平洋中心(近五四广场)
三合园水饺	82	36.0735	120.408	漳州二路39号	1	三合园水饺(燕儿岛店)	78	36.0726	120.409	市南区漳州二路39号(近麦凯乐)
黄土地食府	21	36.0739	120.385	山东路8号	1	黄土地食府	4	36.0732	120.385	山东路8号(潜艇学院南侧)
绣罗韩国料理	87	36.0742	120.409	古田路12号	1	绣罗韩国料理(泉州路店)	42	36.0735	120.413	市南区古田路12号(香港花园广场西北侧上行)
海明威大饭店	4	36.075	120.385	山东路7号	1	海明威大饭店山东路店	2	36.0743	120.385	山东路7号(闽江路与山东路交叉口)
大昌海鲜酒楼	3	36.0759	120.388	闽江路30号	1	大昌海鲜酒楼	28	36.0754	120.388	市南区闽江路30号(闽江路与南通路交叉口附近)
翠峰苑火锅	38	36.0774	120.399	云霄路82号	1	翠峰苑火锅青岛云霄路店	180	36.0767	120.399	云霄路82号(银城花园附近)
船歌鱼水饺	102	36.0776	120.405	闽江二路57号	1	船歌鱼水饺	41	36.0769	120.406	青岛市市南区闽江二路57号
老转村山东菜	125	36.0779	120.395	南京路43号	1	老转村山东菜馆CHINA公社	205	36.0785	120.402	市南区闽江三路8号1388文化街内(近翠峰苑)
彤德莱火锅青	130	36.0785	120.4	闽江路141号	1	彤德莱火锅青岛闽江店	185	36.0776	120.401	市南区闽江路141号(闽江四路和闽江路交叉口)
唐家长院子风	124	36.0802	120.402	闽江三路18号	1	唐家长院子风味酒楼(闽江店)	206	36.0792	120.402	市南区闽江三路18号(闽江三路与江西路交叉口)
甜糖湾渔港大	109	36.0842	120.413	大尧三路1号	1	甜糖湾渔港大酒店(市南店)	55	36.0833	120.413	大尧三路1号(欣宇花园北侧)

图 3-4 名称字符段属性匹配结果（百度与 Google 电子地图）

图 3-4 显示了百度与 Google 电子地图名称字符段匹配显示结果，从图中我们可以看出同一个店铺，在百度和 Google 电子地图上面的名称可能会有些许差别，而我们要做的就是将名称字符段进行融合。具体实现是通过对名称字段的字符串匹配，设置阈值，统计正确率。

表 3-1 不同阈值下三种算法正确率统计

方法	准确率	召回率	F 值
Jaccard	0.78	0.83	0.815
Levenstein	0.92	0.91	0.915
Jaro	0.85	0.81	0.82

表 3-1 显示了不同字符串方法的准确率与召回率，阈值等情况。从表中可以看出，三种方法的准确率都超过了 75%。本文首先进行了一个简单校正工作，如第二章介绍可知，不同的电子来源若是都有同一个 POI 点，且其基本信息是相同的，那么我们就可以认为该点是同一个点，这就是一个简单的校正工作，可以看出在校正之后的数据里面，进行名称融合的准确率是比较高的。

### 3.3.2 地址相似度融合实现—地址规范化

不同来源的地址信息描述方式不同，例如对于中国海洋大学崂山校区的描述，有的电子地图直接是松岭路 238 号，而有的地图是山东省青岛市崂山区松岭

路 238 号 ,这就需要对地址信息进行规范 ,然后对切分之后同一等级的地址进行字符串匹配。下图显示的是不同电子地图的地址切分之后的结果。

百度						Google					
省	市级	县/区	乡/镇	街道/路/里	号	省	市级	县/区	乡/镇	街道/路/里	号
山东	青岛	崂山		松岭路	238号					松岭路	238号
山东	枣庄	滕州	西岗			山东		滕州	柴	煤矿	
	青岛			彰化路	2号			市北区		彰化路	2号
省	市级	县/区	乡/镇	街道/路/里	号						

图 3-5 地址属性的字符串匹配前结果

从上图我们可以看出对于同一个 POI 点 ,左右两边电子地图的表示不是完全一样的 ,对于不同字段有可能有缺失 ,我们做的主要目的是补充 ,使地址数据信息更完全。例如上面 3 条数据 ,我们可以做这样的补充 ,第一条将左边省市县添加 ,后面等级不变。第二与三条可以则是将两边的等级地址内容相互补充。得到如下结果 :

地址匹配结果举例						
	省	市级	县/区	乡/镇	街道/路/里	号
	山东	青岛	崂山		松岭路	238号
	山东	枣庄	滕州	西岗	煤矿	
		青岛	市北区		彰化路	2号

图 3-6 地址属性的字符串匹配后的结果

### 3.3.3 评论信息的添加

本文所抽取的都是 Web 上面的商铺的信息 ,商铺主要指的是团购美食店铺 ,这类团购的一个特点是店铺地址详细 ,店铺名称易于理解 ,客户流量比较大 ,客户反馈比较多 ,这样的评论信息就比较丰富 ,具有良好的参考价值。一般一个店铺的评论信息都超过千条 ,这些评论信息是不同客户根据自己的真实体验所提交的。从某种方面来说 ,这些评论信息是全面客观的 ,因此如果对这些信息进行有效的利用 ,对店铺来说可以更加反馈改善体验 ,对客户来说可以更好地做出选择。

本文所做的就是将点评网站的评论信息 ,添加的数据库中 ,反馈到百度 ,Google ,Mapabc 等电子地图提供的数据库源中 ,使之信息更加全面 ,为客户提供

更加丰富的信息，更好地进行服务。

### 3.4 本章小结

本章主要系统地介绍了 POI 数据的融合实现过程，主要是从名称、地址、评论信息添加等方面实现。具体来说，名称融合主要通过中文字符串匹配算法实现，名称融合也是一个别名化的过程，这样具有不同名称介绍的同一个 POI 点的信息更加全面化；地址融合主要是通过地址等级切分，字符串判定等步骤进行融合；评论信息的添加则是对网页抽取的评论信息对应添加到地图网站得到的信息的数据库中，最终丰富了 Web POI 数据库内容。

## 4 基于 GibbsLDA++主题抽取原理与应用

### 4.1 主题抽取介绍

主题抽取 ( Topic Model )<sup>[35-37]</sup>指的是一种统计模型,它可以被用来从大量文档的中发现抽象的主题。在机器学习领域<sup>[38]</sup>和自然语言处理以及统计语言领域, Topic Model 能够表达数据的潜在的语义,它可以用来发现出现在一个文档集合的抽象的“主题”。换句话说,因为一个文件是关于某个特定主题,人们所期望的特定词<sup>[39]</sup>或多或少频繁出现在文件中。Topic 主题是每个文档中词语信息的一个平衡的信息的统计。主题模型建立,是在自然语言处理的背景下实现的,他们在其他领域,如生物信息学的应用,也有其他应用。而本文中主要是用来对评论信息进行主题的抽取。

通过上文我们了解到 POI 评论信息一般是一些短小的段落,这些小段落的内容代表了顾客的个人评价,具有一定的代表意义。因此对这些大量的评论信息进行主题抽取,得到有效的评论信息,就意义重大。例如我们上文中所用到的评论信息,主要是对一个店铺的客户评价,因此类似“好”“很好”“不好”“好吃”“可口”“美味”等词会经常出现。

一段评论信息通常有 2-3 个主题,而评论中的中的特定词汇则会体现主题的具体描述。在机器学习与自然语言处理中,评论信息可以看做是一段有效文本。文本主题抽取的实际含义就是统计一段文本中词语的概率分布。

如下图所示,它显示了 Topic 内的每个词的词频, Topic1 中 money 4 次, bond 2 次...那么任意给一段评论信息我们都可以统计其中词汇的出现次数,进而计算该词语的概率即  $P(w=w_i | t=t_j)$ ,这一概率显示了一个文档中各个词汇的概率,通过概率比较来提取一个文档中的主题含义。

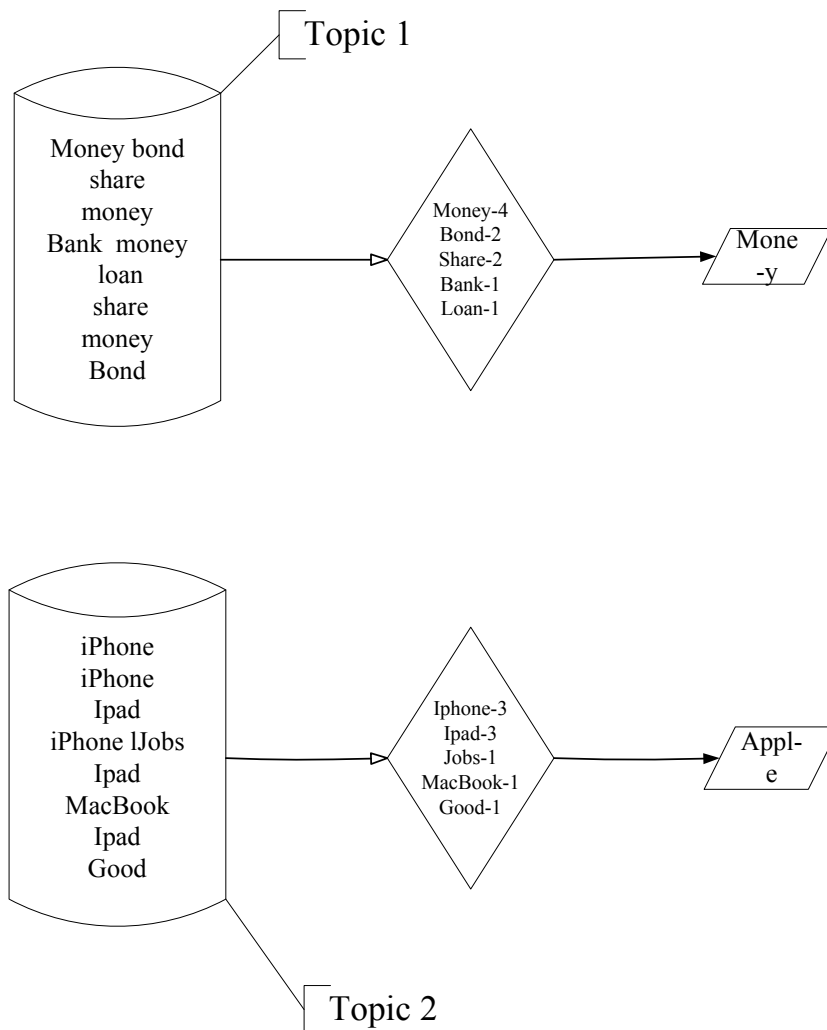


图 4-1 Topic model 主题模型介绍

## 4.2 Gibbs 抽样与 GibbsLDA++

### 4.2.1 Gibbs 抽样

Gibbs 抽样<sup>[40, 41]</sup>是被广泛应用的一种 Metropolis Hastings<sup>[42]</sup> 采样方法,它通过建立 Markov 链,对未知的变量进行估计,当 Markov 链达到稳态分布的时候就是所要求的。

Gibbs 抽样具体实现过程如下：用  $U$ 、 $V$  两个英文字母来代表随机变量；概率密布函数则用  $[U]$ 、 $[V]$  来表示， $[U|V]$ 、 $[V|U]$  用来代表两者之间的条件分布函数。一般来说，若是已经确定条件分布  $[U|V]$ 、 $[V|U]$ ，无论  $[U, V]$  是否确定，服从  $[U, V]$  的随机点列  $\{(U, V)_m\} = \{(U_1, V_1), (U_2, V_2), \dots, (U_n, V_n), \dots\}$  都可以

通过下面的过程呈现，这个点列的“边缘数列”  
 $\{(U)_n\} = \{U_1, U_2, \dots, U_n, \dots\}$ ,  $\{(V)_n\} = \{V_1, V_2, \dots, V_n, \dots\}$  则服从  $[U, V]$  的边缘分布  $[U], [V]$ ，这个边缘数列具有良好的收敛性：

$$\lim_{u \rightarrow \infty} \frac{1}{u} \sum_{i=m+1}^m + ng(u_i, v_i) = E[g(u, v)] \quad \text{式(4-1)}$$

实现过程如下：

Step1 从  $U$  的可能取值中选择一个点  $U_1$ ，根据条件分布  $[V | U = U_1]$  得到随机数  $V_1$ ，那么  $\{(U, V)_n\}$  中的第一个随机数对  $(U_1, V_1)$  产生；

Step2 依照条件分布  $[U | V = V_1]$  得到  $U_2$ ，再依照条件分布  $[V | U = U_2]$  得到  $V_2$ ，那么  $\{(U, V)_n\}$  中的第二个随机数对  $(U_2, V_2)$  产生；

Step3 迭代以上两个步骤  $n$  次，最终得到随机点列  $\{(U, V)_n\}$ 。

#### 4.2.2 GibbsLDA++

GibbsLDA++是由 Xuan-Hieu Phan 和 Cam-Tu Nguyen 发布的，是一个使用 C / C++ 工具实现 Latent Dirichlet Allocation<sup>[43]</sup> (LDA)，它借用 Gibbs 抽样技术进行参数估计和推断。它的非常快，开发者的目的是为了分析大规模的数据包括大型文本/Web 文档数据等的主题结构<sup>[44]</sup>，LDA 最早由 David Blei 提出，它还可以由 java，matlab 等语言实现。本文使用的是 C++ 版本。

GibbsLDA++在很多领域都有非常好的应用：信息检索和搜索（分析大文本集合了更多的智能信息检索语义/潜在主题/概念结构）；文件分类或聚类，基于内容的图像聚类，文档自动文摘、文本和 Web 挖掘社会大众，目标识别和计算机视觉等。

LDA 实现主题抽取过程如下图 4-2 所示，首先是文本数据，然后对其进行分词处理，对应词表与 LDA 索引文件，建立简单主题模型，最后进行模型主题抽取，还可以进行新文本数据参数估计与主题抽取。



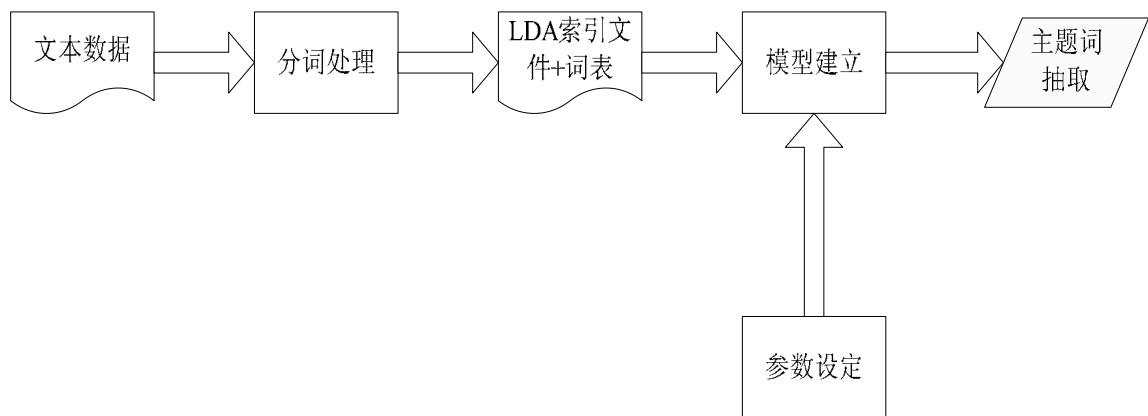


图 4-2 基于 LDA 主题模型抽取流程图

具体使用过程如下（本文使用的是 GibbsLDA++-0.2 版本）：

在 Linux 系统下，进行如下基本操作。

基本解压命令：`$ gunzip GibbsLDA++.tar.gz`

`$ tar -xf GibbsLDA++.tar`

Compiled 编译环节：

`$ make clean`

`$ make all`

Command line: 输入操作命令，下面这几条语句是整个实现的核心。

`$ lda -est -alpha <double>] -beta <double>] -ntopics <int>] -niters <int>] -savestep <int>]  
-twords <int>] -dfile <string>`

`$ lda -estc -dir<string> -model<string> [iters<Int>] [savestep<Int>] [tword<im>]`

`$ lda -inf -dir<string> -model<string> [-niters<int>][ -twords<int>]-dfile[<string>]`

参数解释如下：

-est 表示进行模型估计；-estc 继续估计；-inf 表示用训练好的构建的模型估计新文本。

-alpha -beta

“其中  $\alpha$  ,默认值是 topic 数目除以 100。为了使每一个文档的主题接近同一个,可以调大这个参数;调大  $\beta$  可以让每个文档的主题近可能的集中在某几个词语上面,换句话说,可以让这些词语的百分比集中到一个主题词那儿。

-ntopics

代表输入到语料库的主题数, LDA 模型原理是用大量的训练数据产生主题模型,若是尽可能的覆盖所有主题,则 inf 的结果会更准确。所以训练的数据越大越好。

-niters

迭代次数,指的是程序重复迭代运行的数目, 1000 次以上才能达到收敛的效果,一般来说迭代次数多,效果好同样效率会降低。

-savestep 具体抽样迭代次数计算,重复多少轮保存到硬盘:一般来说如果实验语料较大的话,需要勤保存,默认值为 200,此参数对模型计算结果基本没有影响。

-twords

统计词汇概率从高到期来确定文本主题词。

-dfile

训练文件。输入文档中的第一行表示总共文本个数,第 2 到第 N 行中代表了每一个文本,一行是一个。具体实验数据是分词之后去掉停用词等之后的具有实际意义的词汇。

GibbsLDA++的输出包含以下文件:

<model\_name> 每次迭代产生的模型的名称,例如第一次迭代名字为 model-1,第 100 次为 model-100,特别强调最后一次的迭代保存名称为 model-final。

<model\_final>.others 参数设定保存在这个文件夹里。

<model\_final>.phi 这个文件包含了文章主题的描述,每一行是一个主题,每一行是一个词的统计

<model\_final>.theta 每一个潜在主题词的概率统计

<model\_final>.tassign 每一个潜在主题词的分布情况包含在这个文件夹下。

<model\_final>.twords 每一个潜在主题词的前置词

### 4.3 文本分词预处理

从中文文本中词与词之间没有显著的界限的特点来说,做中文主题抽取就要实施一定的中文预处理<sup>[45]</sup>,另外 GibbsLDA++所需要的输入文档需要是分割之后的词,因为它实际上是对文本文档中所出现词汇的一个概率统计,因此首先需要将文本分词。

本文使用的文本数据是团购网站抽取的评论信息,目的是为了从评论信息分析出用户的意见,因此选取的文本一般能够描述主题特征的一些评论词汇,具体实现是通过以下步骤:

第一,词性:首先是名词,名词是描述性词,用来表述人、事物、地点或者抽象事物。在主题抽取的时候能够代表一定的意义,对主题表述有深刻影响;动词是表示各种动作的词语;必须要注意的是在分词时注意过滤掉停用词,可以减少计算的复杂度以及带来的消极影响,并且节省存储空间和提高效率。

第二,词频:一般来说用户在对一个店铺的产品进行评价时,会反复使用与主题密切相关的词汇,例如在评论一家川菜的口感时,“辣”“很辣”“非常辣”,用户可能使用三个不同表述方法,用来表述这家菜馆确实很“辣”。因此,从某种意义来说,词汇的出现频率,在某个层面上反应了该词汇是否与主题相关,也就是说频繁出现的词汇在一定程度上代表着主题含义所在。

第三,词语位置:从常规中文描述来看,用户一般将第一感觉放在一段评论信息的开始位置,因此可以这样说开头位置的词汇出现概率在一定程度上代表了主题含义,例如同样是对一家川菜馆的评论,A用户可能评论开始就用形容词“不好”,那么通过开始评论我们就可以了解到,A用户对此不满意,他的主题想表达的就是“菜馆不好”,而B用户可能一开始评论为“味道不错”,那么从这里我们可以看到,B用户是基本满意的。通过上述的例子,可以表明,评论开头处的信息很重要。

中文文本的主题词抽取一般有四个步骤,分别是:中文分词、删除停用词、计算词汇的统计概率、主题词的确定。这样分词之后再删减去停用词,,名词、动词等具有描述意义的词汇会继续存在,本章主要解决的问题就是评论信息分词之后进一步的主题抽取工作。

因此,在对评论信息进行主题抽取之前需要对其采取中文分词,具体主要指的是通过已经存在的分词词典、或者自己构建新的特征词词典、或者语法分析等,将中文字符串切割成一个一个词语,形成词语集合的过程。从技术专业的角度看,常见的分词,有基于词典(即字符匹配)、基于统计(即字符串统计概率)、基于语法与规则(即语义语法分析)三种。具体来说有以下五种:机械匹配法分词法从字面意义看就是机械地一个一个匹配,其基本原理是设置标准词库,词库一般涵盖所有可能出现的词汇,然后根据词库对文本信息匹配分词:具体实现的算法是按照顺序依次扫描每个字符,切分之后与字典中的词汇进行对比,当字符串与词库中的词语匹配上之后结束。特征词库法是对上述方法的提高与改进,特征词法需要先构造特征词库,一般包涵虚词、叠词、前后缀等,然后用词库里的词切分已知字符串,获得 N 个子串,子串的处理参照机械匹配法;约束矩阵法的优点在于它可以分析消除句子的歧义,这个算法的实现是首先构建一个可以判断相邻词语逻辑语法关系的语义约束矩阵,把它作为规则去判断词语。语法分析法是对约束矩阵法的一个改良,它构建一个汉语的语法规则,进一步提出了全局约束的特点,并改进了中局部约束的不足。理解切分法是一种注重语言整体性的方法,它改善了语法分析法,使用它时需要先自己构建一个定义的规则的语义库。

根据上面几种不同的分词算法,研究员们设计开发出了几种不同的分词系统。本文使用的是中国科学院计算技术研究所研发的 NLPIR<sup>[46, 47]</sup>系统,这个汉语词法分析系统是基于多层隐马尔可夫模型的。该汉语分词系统提供中文分词、词性标注以及新词识别等功能。据了解,中科院的这套中文分词系统的分词正确率可以达到 96.75%,分词速率可以达 544ib/s,词性标注的基本和分词速率一致,另外其基于角色标注的新词识别召回率高达 85%,综上所述,该分词系统可以有效地实现中文分词,在使用该系统的基础上,本文又考虑了分词之后停用词的影响,通过批处理方式删减掉一些停用词,降低其带来的负面因子影响。

本文使用的是 NLPIR 2013 版 java 版(另外有 C++版、C#版等版本,可以参考编程使用)。原评论如下:

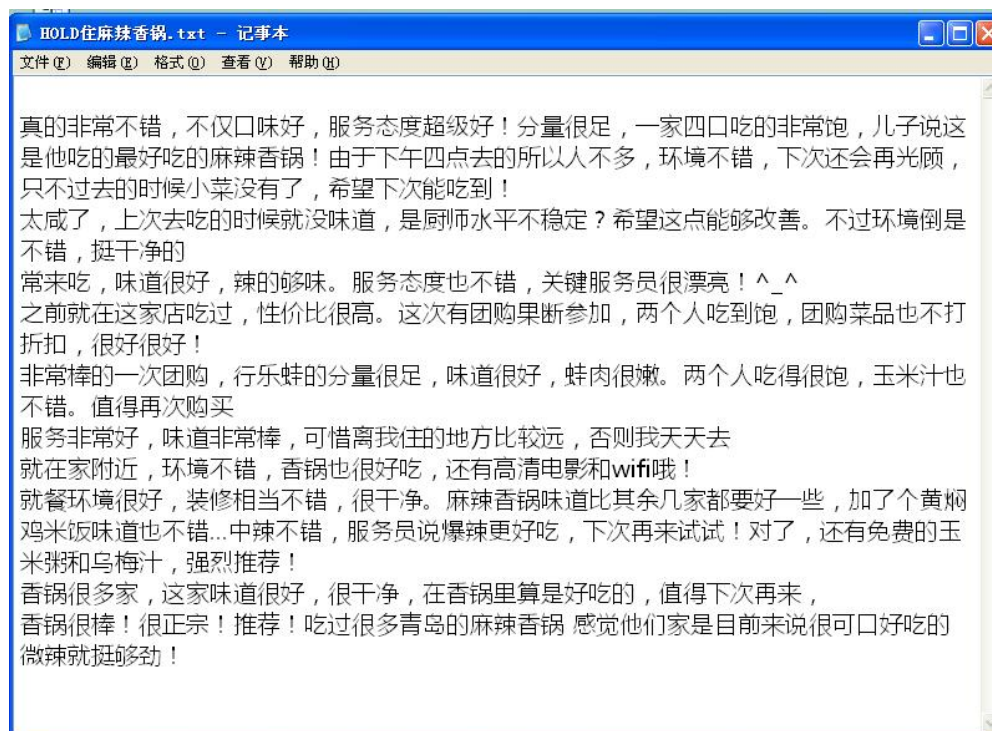


图 4-3 基于分词前的评论信息

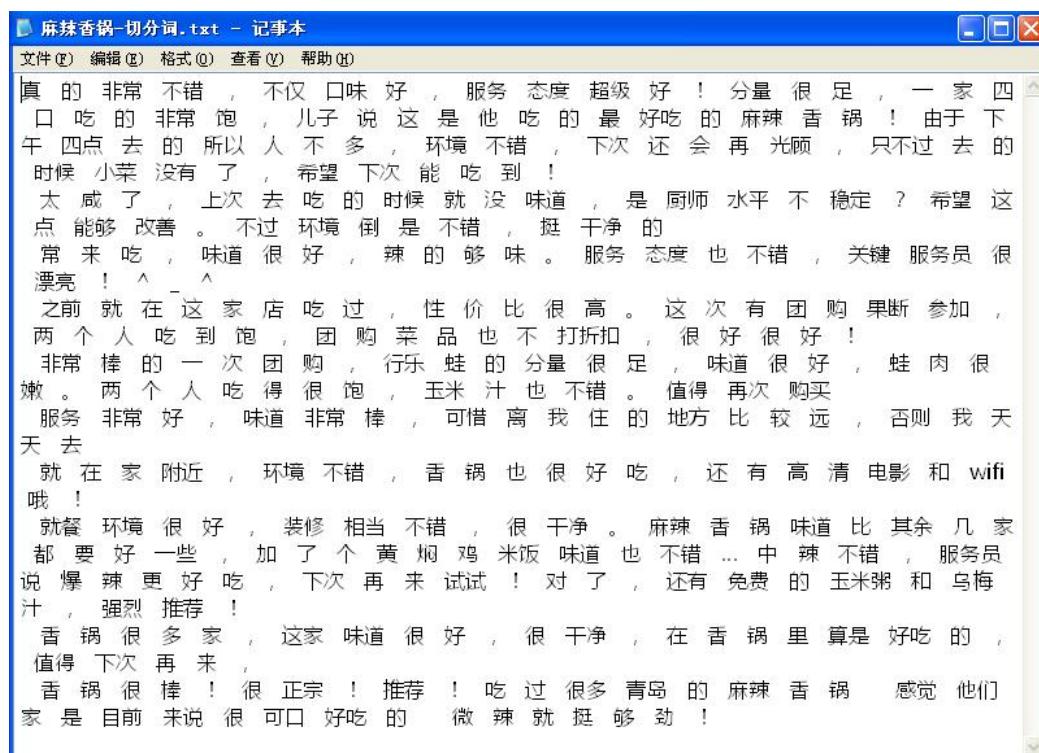


图 4-4 基于 NLPIR 分词实现评论信息的分词

图 4-3 是分词前的评论信息，图 4-4 是基于 NLPIR 分词系统实现的分词，从图中我们可以看出，文本有效地被分成词组。同时，我们也主要到另一个问题，分词中含有许多信息量低的停用词。停用词没有明确的意义，一般来说文本中停

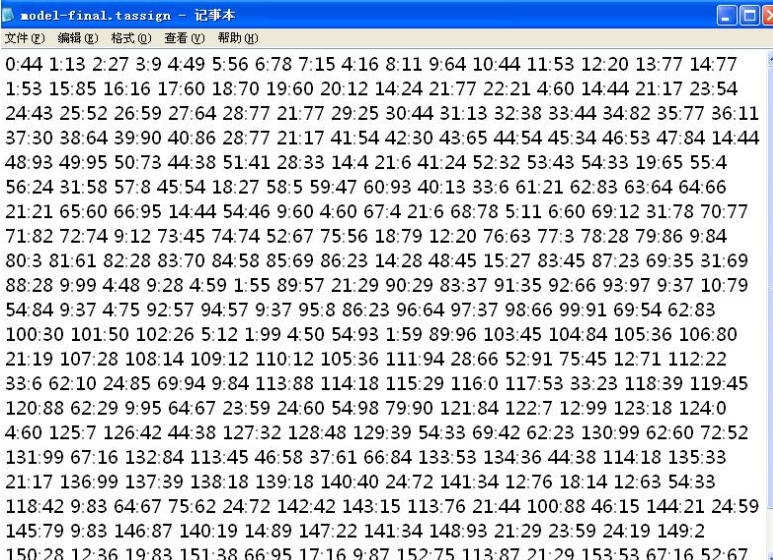
用词比较普遍，没有明显功能，对主题抽取与文本标识没有什么作用，比如 "at""which""on""whose""that" 等英文单词，“由于”、“哪里”、“地”、“啊哈”等中文字符在信息检索与主题抽取中没有什么明显含义，但是他们的存在会带来一定的计算误差。实验中我们选取了 100 家店铺的评论信息，选取其中评论信息有 19 封、140 封、996 封的三个阶段做出统计，如下表 4-1

表 4-1 评价词汇统计

总评价	总词汇	停用词	去重之后	有效词概率
A	1222	236	373	30.52%
B	10500	2488	3594	34.22%
C	107568	15632	27634	25.69%

从上表中我们可以看出，融合评论信息的时候，词汇的重复率还是比较高的，从图 4-4 我们也可以看出评论信息中评价为正面信息的，几乎每条都会有一个“好”字。而重复率高的词汇对主题的影响因子就相对大些，在做分词处理后，停用词我们使用的是常见中文停用词库，批处理去掉了多余停用词，尽可能的消除停用词<sup>[48]</sup>带来的噪声影响。

## 4.4 实验结果与分析



0:44 1:13 2:27 3:9 4:49 5:56 6:78 7:15 4:16 8:11 9:64 10:44 11:53 12:20 13:77 14:77  
1:53 15:85 16:16 17:60 18:70 19:60 20:12 14:24 21:77 22:21 4:60 14:44 21:17 23:54  
24:43 25:52 26:59 27:64 28:77 21:77 29:25 30:44 31:13 32:38 33:44 34:82 35:77 36:11  
37:30 38:64 39:90 40:86 28:77 21:17 41:54 42:30 43:65 44:54 45:34 46:53 47:84 14:44  
48:93 49:95 50:73 44:38 51:41 28:33 14:4 21:6 41:24 52:32 53:43 54:33 19:65 55:4  
56:24 31:58 57:8 45:54 18:27 58:5 59:47 60:93 40:13 33:6 61:21 62:83 63:64 64:66  
21:21 65:60 66:95 14:44 54:46 9:60 4:60 67:4 21:6 68:78 5:11 6:60 69:12 31:78 70:77  
71:82 72:74 9:12 73:45 74:74 52:67 75:56 18:79 12:20 76:63 77:3 78:28 79:86 9:84  
80:3 81:61 82:28 83:70 84:58 85:69 86:23 14:28 48:45 15:27 83:45 87:23 69:35 31:69  
88:28 9:99 4:48 9:28 4:59 1:55 89:57 21:29 90:29 83:37 91:35 92:66 93:97 9:37 10:79  
54:84 9:37 4:75 92:57 94:57 9:37 95:8 86:23 96:64 97:37 98:66 99:91 69:54 62:83  
100:30 101:50 102:26 5:12 1:99 4:50 54:93 1:59 89:96 103:45 104:84 105:36 106:80  
21:19 107:28 108:14 109:12 110:12 105:36 111:94 28:66 52:91 75:45 12:71 112:22  
33:6 62:10 24:85 69:94 9:84 113:88 114:18 115:29 116:0 117:53 33:23 118:39 119:45  
120:88 62:29 9:95 64:67 23:59 24:60 54:98 79:90 121:84 122:7 12:99 123:18 124:0  
4:60 125:7 126:42 44:38 127:32 128:48 129:39 54:33 69:42 62:23 130:99 62:60 72:52  
131:99 67:16 132:84 113:45 46:58 37:61 66:84 133:53 134:36 44:38 114:18 135:33  
21:17 136:99 137:39 138:18 139:18 140:40 24:72 141:34 12:76 18:14 12:63 54:33  
118:42 9:83 64:67 75:62 24:72 142:42 143:15 113:76 21:44 100:88 46:15 144:21 24:59  
145:79 9:83 146:87 140:19 14:89 147:22 141:34 148:93 21:29 23:59 24:19 149:2  
150:28 12:36 19:83 151:38 66:95 17:16 9:87 152:75 113:87 21:29 153:53 67:16 52:67

图 4-5 主题模型建立数据显示

model-final.tassign 中的每一行代表着每一个输入文档。代表了该文档中每个词语与文档匹配的主题。



例如第一行数据 0:44 1:13 2:27 3:9 4:49 5:56 6:78 7:15 4:16 8:11 9:64 10:44 输入文档有在 wordmap 中标号为 0, 1, 2 等的词语组成, 这些词语分别被分配给了主题 44, 13, 27 等。而在实际主题抽取中因为评论信息较为短小, 我们一般设置 10 个以下的主题。

model-final. twords			model-final. twords		
1	Topic 0th:		27	Topic 2th:	
2	电影	0.062857	28	非常	0.006061
3	口味	0.005714	29	棒	0.066667
4	好	0.005714	30	口味	0.006061
5	服务	0.005714	31	好	0.006061
6	态度	0.005714	32	服务	0.006061
7	超级	0.005714	33	态度	0.006061
8	分量	0.005714	34	超级	0.006061
9	很	0.005714	35	分量	0.006061
10	足	0.005714	36	吃	0.006061
11	吃	0.005714	37	饱	0.006061
12	饱	0.005714	38	说	0.006061
13	说	0.005714			

图 4-6 主题模型建立

model-final.twords 用来描述文档的主题词, 在实验中我们设置了 3 个主题, 输出分别为 topic 0,topic1 , topic2 , 如上图所示显示的是 topic 0 与 topic 2 两个主题中词语的概率。

由以上实验结果可以看出, GibbsLDA++对有些评论信息文件尚不能够有很好的识别, 产生有效的主题模型, 分析一下因素会导致这种问题的产生:

- (1) 本身部分评论信息较短, 有可能是 4-5 字, 就没有什么重要意义。
- (2) 评论信息中包含有较多的停用词, 例如是, 和, 地, 的, 这些词语本身不具备有效的意义, 但是同时占据有文档内容, 做统计时候, 只能消除这些简单停用词, 有复杂的停用词依然有可能影响实验结果。

因此, 综上所述, GibbsLDA++对于评论信息建立的模型还有待改进, 尤其

是消除停用词带来的消极影响。

## 4.5 本章小结

本章主要介绍了 Topic Model 的简单知识以及应用，然后引入了 Gibbs 抽样的相关概念，然后详细介绍了用该抽样函数进行主题抽取的 GibbsLDA++ 的使用配置过程以及具体参数的意义介绍，统计了不同文本下主题模型的构建内容，对从 web 页面抽取的评论信息，构建了简单的主题模型，并为后续主题模型的抽取打下了基础。



## 5 总结与展望

### 5.1 总结

本文深入而系统地研究了 POI 数据融合技术和评论信息文本挖掘及主题抽取领域，最终实现了在少量人工干预的情况下的 POI 数据的自动抽取与融合功能。下面总结下本文的主要研究内容以及实验主要工作：

本文实现了 POI 数据的校正工作，基于空间属性方法与非空间属性方法统一的方法，对现有的单空间或者单非空间方法提出改进。

由于不同的电子网络地图加偏造成的经纬度偏差影响了实验结果，本文提出了有效地解决方案，即坐标的纠偏与统一，将不同来源 POI 数据的经纬度坐标统一到一起。

本文实验过程中，需要频繁使用网络访问电子，访问由于受到地图本身限制原因影响实验的实施，本文通过使用 http 代理服务器隐藏 IP 等信息解决了这个问题。

实现了名称别名化、地址标准化、经纬度统一化等三种不同的融合。实验结果表明，对 POI 信息的融合使其信息变得更加全面准确完善。

对 POI 评论信息进行了主题抽取，建立了主题模型，有效地使用 GibbsLDA++，使主题模型的抽取变得更加准确有效。

### 5.2 展望

#### 5.2.1 存在问题

本人采用了空间位置属性方法与非空间位置相结合的方法，有效地使用了 POI 的名称、地址、经纬度属性，并对这三个属性的特征进行了融合匹配。大大提高了融合集的质量，使空间 POI 数据的可用率更加准确。同时，本文有效地使用了主题模型，对 POI 评论信息进行了主题抽取，构建了不同的主题模型，

此外，本文还存在不足之处，各种参数的设定以及最近阈值的选择是比繁琐而较复杂的过程，并且对于不同的方法，阈值是不同的，有效的进行阈值的选择

也是一个不可忽略的因素。另外，在主题抽取模块，刚刚开始进行，使用 GibbsLDA++主要是对主题进行了一个简单的抽取，模型的构建刚刚起步。

### 5.2.2 未来工作安排规划

继续推进主题抽取研究工作，建立一套完整主题抽取的方法，构建主题模型，进一步加深对 GibbsLDA++的使用，有效提高文档的词语识别率与准确率，逐步加大 Topic model 的数量，有效地使用，使主题构建更加顺利。

## 参考文献

- [1] Faouzi N E, Leung H, Kurian A. Data fusion in intelligent transportation systems: Progress and challenges—A survey[J]. Information Fusion, 2011,12(1):4-10.
- [2] Gu Ben. 信息过载问题及其研究[J]. 中国图书馆学报, 2000(5):42-45.
- [3] Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases[J]. AI magazine, 1996,17(3):37.
- [4] Krösche J, Boll S. The xPOI concept[M]//Location-and Context-Awareness. Springer, 2005:113-119.
- [5] 张玲. POI 的分类标准研究[J]. 测绘通报, 2012(10):82-84.
- [6] 刘东琴. 地理实体数据库构建研究[博士学位论文]. 青岛: 山东科技大学, 2010.
- [7] Liu B, Grossman R, Zhai Y. Mining data records in Web pages: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 200:603-606
- [8] 戴冬冬. 基于地址匹配方法的 POI 数据更新研究[J]. 电脑知识与技术: 学术交流, 2010,6(1):1-2.
- [9] Beeri C, Kanza Y, Safra E, et al. Object fusion in geographic information systems: Proceedings of the Thirtieth international conference on Very large data bases-Volume 30, 2004:816-827
- [10] Beeri C, Doytsher Y, Kanza Y, et al. Finding corresponding objects when integrating several geo-spatial datasets: Proceedings of the 13th annual ACM international workshop on Geographic information systems, 2005:87-96
- [11] Safra E, Kanza Y, Sagiv Y, et al. Integrating data from maps on the world-wide web[M]//Web and Wireless Geographical Information Systems. Springer, 2006:180-191.
- [12] 高新院. 基于空间位置信息的多源POI数据融合问题的研究[硕士学位论文]. 青岛: 中国海洋大学, 2013.
- [13] 李瑞姗. 基于自然语言处理的多源 POI 数据融合的研究[硕士学位论文]. 青岛: 中国海洋大学, 2013.
- [14] 张永新. 面向 Web 数据集成的数据融合问题研究[博士学位论文]. 山东大学, 2012.
- [15] Elwood S, Goodchild M F, Sui D Z. Researching volunteered geographic information: Spatial data, geographic research, and new social practice[J]. Annals of the Association of American Geographers, 2012,102(3):571-590.
- [16] Wang L, Wei B, Yuan J. Document Clustering Based on Probabilistic Topic Model[J]. Acta Electronica Sinica, 2012,11:33.
- [17] Peters J, Matusov E, Meyer C, et al. Text segmentation and label assignment with user

- interaction by means of topic specific language models and topic-specific label statistics[Z]. Google Patents, 2012.
- [18] 江宽, 龚小鹏. Google API 开发详解: Google Maps 与 Google Earth 双剑合璧[M]. 电子工业出版社, 2008.
- [19] Allauddin M, Azam F. Service Crawling using Google Custom Search API.[J]. International Journal of Computer Applications, 2011,34.
- [20] LIANG G, LI H. Designing and Developing Virtual Campus Based on Baidu Map API [J][J]. Journal of Langfang Teachers College (Natural Science Edition), 2012,6:16.
- [21] Wang Y, Zhu X. A coal-bed methane WebGIS based on MapABC Maps API and DWR: Computing, Control and Industrial Engineering (CCIE), 2011 IEEE 2nd International Conference on, 2011(2):167-170.
- [22] Buttler D, Liu L, Pu C. A fully automated object extraction system for the World Wide Web: Distributed Computing Systems, 2001. 21st International Conference on., 2001:367-370.
- [23] Mohammad Seyedzadeh S, Mirzakuchaki S. A fast color image encryption algorithm based on coupled two-dimensional piecewise chaotic map[J]. Signal Processing, 2012,92(5):1202-1215.
- [24] Baier P, Weinschrott H, Durr F, et al. MapCorrect: automatic correction and validation of road maps using public sensing: Local Computer Networks (LCN), 2011 IEEE 36th Conference on, 2011:58-66.
- [25] 刘恩信. 厦门市二调数据成果 1980 西安坐标转换[J]. 测绘与空间地理信息, 2009,32(3):198-201.
- [26] Northrup A. NT Network Plumbing: Routers, Proxies, and Web Services[M]. IDG Books Worldwide, Inc., 1998.
- [27] Boldyreva A, Palacio A, Warinschi B. Secure proxy signature schemes for delegation of signing rights[J]. Journal of Cryptology, 2012,25(1):57-115.
- [28] 王静帆, 邬晓钧, 夏云庆, 等. 中文信息检索系统的模糊匹配算法研究和实现[J]. 中文信息学报, 2007,21(6):59-64.
- [29] Girres J F, Touya G. Quality assessment of the French OpenStreetMap dataset[J]. Transactions in GIS, 2010,14(4):435-459.
- [30] Rajabzadeh M, Tabibian S, Akbari A, et al. Improved dynamic match phone lattice search using Viterbi scores and Jaro Winkler distance for keyword spotting system: Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International Symposium on, 2012:423-427.
- [31] Hamers L, Hemeryck Y, Herweyers G, et al. Similarity measures in scientometric research: the Jaccard index versus Salton's cosine formula[J]. Information Processing & Management, 1989,25(3):315-318.
- [32] 程昌秀, 于滨. 一种基于规则的模糊中文地址分词匹配方法[J]. 地理与地理信息科学,

2011,27(3):26-29.

- [33] Xu J, Wu F, Qian H, et al. Settlement matching algorithm using spatial similarity relations as constraints[J]. Geomatics and Information Science of Wuhan University, 2013,38(4):484-488.
- [34] Ye F, Shi X, Wang S, et al. Spherical interpolation over graphic processing units: Proceedings of the ACM SIGSPATIAL Second International Workshop on High Performance and Distributed Geographic Information Systems, 2011[C]. ACM.
- [35] Li H, Yamanishi K. Topic analysis using a finite mixture model[J]. Information processing & management, 2003,39(4):521-541.
- [36] Steyvers M, Griffiths T. Probabilistic topic models[J]. Handbook of latent semantic analysis, 2007,427(7):424-440.
- [37] Nallapati R, Cohen W W. Link-PLSA-LDA: A New Unsupervised Model for Topics and Influence of Blogs.: ICWSM, 2008[C].
- [38] Su J, Zhang B, Xu X. Advances in machine learning based text categorization.[J]. Ruan Jian Xue Bao(Journal of Software), 2006,17(9):1848-1859.
- [39] 张晨逸, 孙建伶, 丁轶群. 基于 MB-LDA 模型的微博主题挖掘[J]. 计算机研究与发展, 2011,48(10):1795-1802.
- [40] Porteous I, Newman D, Ihler A, et al. Fast collapsed gibbs sampling for latent dirichlet allocation: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008:569-577.
- [41] 张放, 鲁华祥. 利用条件概率和 Gibbs 抽样技术为分布估计算法构造通用概率模型[J]. 控制理论与应用, 2013,30(3).
- [42] Chib S, Greenberg E. Understanding the metropolis-hastings algorithm[J]. The American Statistician, 1995,49(4):327-335.
- [43] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. the Journal of machine Learning research, 2003,3:993-1022.
- [44] Ahkter J K, Soria S. Sentiment analysis: Facebook status messages[J]. Unpublished master's thesis, Stanford, CA, 2010.
- [45] 叶娜. 面向信息抽取的文本预处理和规则自动学习技术研究[硕士学位论文]. 沈阳: 东北大学硕士学位论文, 2004.
- [46] 张华平, 刘群. 汉语词法分析系统 ICTCLAS [EB/OL][D]., 2010.
- [47] Zhou L, Zhang D. NLPPIR: A theoretical framework for applying natural language processing to information retrieval[J]. Journal of the American Society for Information Science and Technology, 2003,54(2):115-123.
- [48] Butgereit L, Botha R A. Stop words for “ Dr Math ” : IST-Africa Conference Proceedings, 2011, 2011:1-9.

## 致谢

随着论文的工作即将完成，我的硕士研究生生活也接近了尾声。一转眼两年的硕士研究生生活即将结束，回想这两年的点点滴滴，心中有万分不舍，两年的时光里，经历了许多，也收获了许多。而我的所有收获都与每一位给以我指导、关心、帮助的老师、家人、同学、朋友都是分不开的。

在此，感谢尊敬的导师张巍，谢谢他精心的指导与关怀，让我获益很多。感谢他在学术的道路上给我指明方向，教会了我进行科研的方法。导师认真的治学态度以及学术上敏锐的洞察力，对我的科研工作给予了巨大指导和帮助；同时谢谢导师在生活上面给以的帮助与鼓励，让我能够顺利地完成研究工作。

感谢实验室的同学们，谢谢他们在我遇到问题的时给以的巨大帮助。

感谢身边所有的同学与朋友，感谢你们对我的各种帮助。

感谢我的家人，谢谢他们对我的养育之恩，谢谢他们为我付出的一切。

最后，对所有帮助我的老师、同学、朋友和家人表示崇高的敬意和由衷的感谢！

## 个人简历

1988 年 1 月 25 日出生于山东省济宁市金乡县。

2008 年 9 月考入青岛科技大学信息科学技术学院信息工程专业，2012 年 7 月本科毕业并获得工学学士学位。

2012 年 9 月考入中国海洋大学信息科学与工程学院软件工程专业攻读工学硕士学位。



中国海洋大学  
OCEAN UNIVERSITY OF CHINA

# 硕士学位论文

