

《信息系统分析与设计》系统设计报告

论文题目：融合地理位置与领域知识的医疗健康信息推荐知识库

学 院：信息管理学院

专 业：信息管理与信息系统

年 级：2015级

小组成员：程威、胡启军、李紫宸、刘政昊、杨艳琦、蒙天宇、张莹

指导教师：陆泉

**目录**

[1. 设计可行性分析（蒙天宇） 1](#_Toc510984133)

[1.1 技术可行性 1](#_Toc510984134)

[1.2 经济可行性 1](#_Toc510984135)

[1.3 操作可行性 1](#_Toc510984136)

[2. 设计思路（胡启军） 2](#_Toc510984137)

[3. 设计流程 4](#_Toc510984138)

[3.1 数据获取（胡启军、蒙天宇） 4](#_Toc510984139)

[3.1.1 地理别名的获取 4](#_Toc510984140)

[3.1.2 医院数据获取 5](#_Toc510984141)

[3.2本体构建（杨艳琪、张莹） 7](#_Toc510984142)

[3.2.1 本体设计 7](#_Toc510984143)

[3.2.2 本体构建及实例 8](#_Toc510984144)

[3.3 命名实体识别（刘政昊、程威） 11](#_Toc510984145)

[3.3.1 简介 11](#_Toc510984146)

[3.3.2 难点分析 11](#_Toc510984147)

[3.3.3 识别流程的设计 12](#_Toc510984148)

[3.3.4 算法设计 14](#_Toc510984149)

[3.3.5 识别过程中的其他问题 16](#_Toc510984150)

[3.4 推理查询规则（胡启军） 17](#_Toc510984151)

[4.功能逻辑设计（杨艳琦、张莹） 17](#_Toc510984152)

[5.数据库设计（李紫宸） 19](#_Toc510984153)

[5.1.地点信息（t\_place） 19](#_Toc510984154)

[5.2.医院信息（t\_hospital） 22](#_Toc510984155)

[5.3.专家信息（t\_expert） 23](#_Toc510984156)

[5.4.地点别名信息（t\_alias） 23](#_Toc510984157)

[6.界面设计（程威、刘政昊） 24](#_Toc510984158)

# 1. 设计可行性分析（蒙天宇）

## 1.1 技术可行性

数据搜集主要使用Python爬虫，在已有的医院大全网站（99健康网、好大夫在线等）爬取医院信息，构建医院实体；爬取已有的地理行政区划信息，构建地名地理实体，建立地名之间的等级关系；利用百度地图API调用地理实体经纬度信息，利用爬虫获取。

命名实体识别可以学习使用现有的基于人工规则和字典与统计模型相结合的方法，进行分词与实体识别。

## 1.2 经济可行性

Python是免费、开源的软件，百度地图API也是免费、开源的应用，所以网页数据源均是开放资源。命名实体识别的方法与算法都是公开使用的，实体构建使用免费软件protégé进行构建。所有实施过程需要用的软件、应用、算法等都是免费、开源、开放使用的，可以方便地进行数据爬取与处理、实体识别与构建，无需付出经济代价。

## 1.3 操作可行性

知识库建设完毕后，可以通过命名实体识别识别用户的信息，判别出地名、机构名、人名（医生教授名）、药物名、疾病名等实体名称，转化为地名、经纬度、机构名，通过数据预处理，推理得出实体信息中隐射出的最可能的地理需求信息，将最近的匹配的医院、机构等信息反馈回去。整个过程中不需要用户主动输出检索词，而是通过用户在系统中的交互所产生的文本信息，自动进行实体识别，触发推荐反馈。这个功能属于一个推荐子系统，根据用户变化而自动变化，进行实时信息推荐，无需用户付出额外的行为消耗，发觉用户的潜在需求，帮助用户解决可能存在的问题。

# 2. 设计思路（胡启军）

本系统的目标是构建一个关联地理信息与医疗领域信息的知识库，系统设计也围绕着这一目标进行。

对于构建一个知识库而言，主要任务分为两个方面：一是数据的收集和整合，二是设计知识库的使用规则。对于融合地理信息和医疗领域知识的知识库而言，我们需要的数据主要有以下几个方面：1）全国地理实体数据；2）与地理实体对应的地名数据；3）全国医疗机构数据；4）大学研究所等医疗领域科研成果数据。以上数据整合后构建一个融合地理与医疗的整体知识库，这个知识库就是我们所需要的数据源。在此数据源的基础之上，本系统进行基于地理信息的医疗知识推荐。推荐的依据是通过命名实体识别提取的地理相关信息，推荐结果是知识库中存储的医疗领域知识。

因此，系统开发需要经历两个阶段：数据收集阶段和功能开发阶段。在数据收集阶段，任务是收集构建知识库所需的数据并建立知识库。在功能开发阶段，任务是根据构建的知识库设计推理查询机制并实现，从而达到使用知识库进行基于地理的医疗信息推荐。

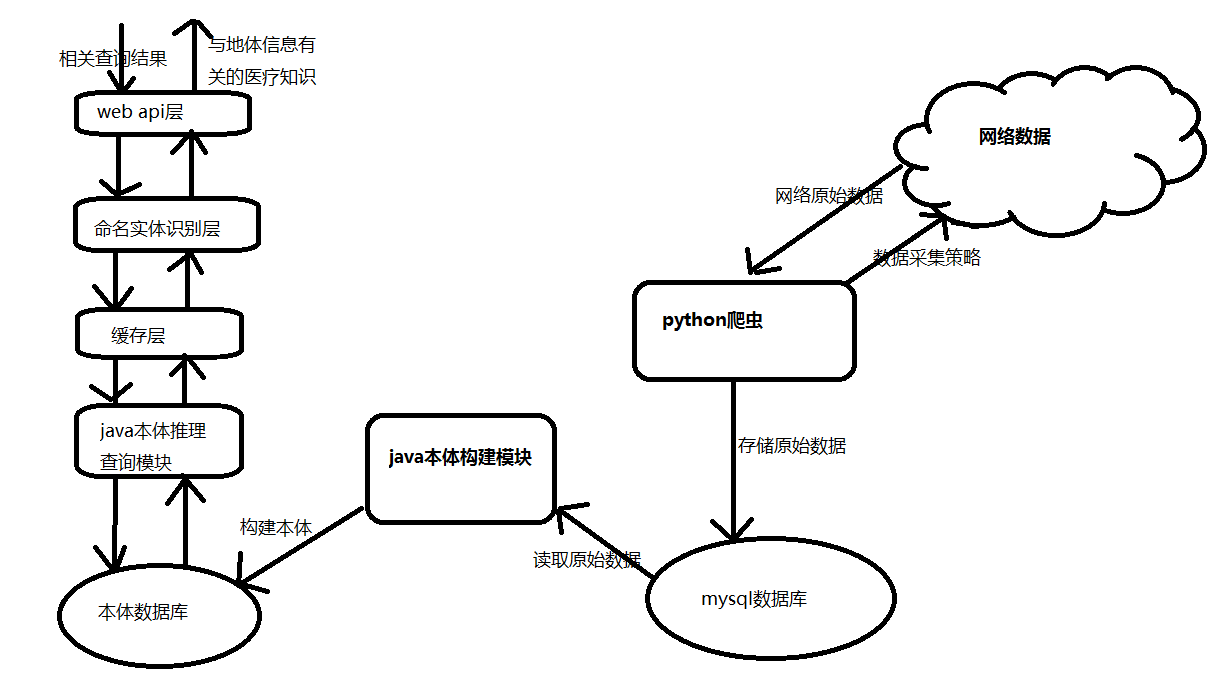
综合两个阶段来看，所需的开发环境如下：

|  |  |  |  |
| --- | --- | --- | --- |
| **用途** | **产品** | **版本** | **备注** |
| 服务器系统 | centos | 7.0 | 作为web app的服务器和数据库服务器 |
| 数据库 | mysql | 5.1 | 存储抓取的原始数据 |
| 爬虫开发语言 | python | 3.5 | 作为开发爬虫的程序设计语言 |
| 本体构建与操作语言 | java | 1.8 | 使用java构建本体并开发web接口服务 |
| Web容器 | tomcat | 7.1.3 | 作为java web应用的servlet容器 |

（表1 系统设计所需开发环境）

下面分别就两个阶段做深入阐述：

1. 数据收集阶段：该阶段的数据来源主要是网络资源和api调用资源。网络资源包括中国行政区划、中国医疗机构目录、地名百科等等。api资源使用包括地理位置经纬度的确定、各个行政区划人文景点等地理实体数据、大学研究论文数据等。数据采集使用python爬虫实现，对于简单的网页数据直接使用python内置的urllib实现，对于js动态生成的页面数据，无法直接爬取，使用selenium自动化测试框架模拟爬取。爬取结果暂存在mysql数据库中，以便之后构建本体知识库。本体的构建使用java语言实现，使用java的jdbc数据库连接接口和jena工具可以实现将数据库中的数据读取出来并构建成本体。
2. 功能开发阶段：功能开发使用java语言开发，web服务通过mvc模式实现，最终的系统运行在linux服务器环境中。其中，model层的功能是对于构建好的本体库的查询使用jena框架进行查询和推理，并返回推荐知识，view层的功能是展示返回的结果，controller层的功能是调度请求。考虑到开始提供服务之后，知识库内容变更较小，为了提高响应速度，减小数据库和cpu压力，拟采用内存缓存的形式对查询结果进行缓存，然后采取定时过期删除的策略防止缓存与知识库内容不符合。综上，整个系统的架构设计大致如下图所示：



（图1 系统架构设计图）

# 3. 设计流程

## 3.1 数据获取（胡启军、蒙天宇）

### 3.1.1 地理别名的获取

地理别名的获取来自百度百科，地名词条信息中的简称、别名一项的数据。



通过python中的Selenium自动化测试工具，实现模拟网页操作，从地理实体数据库中取出所记录的名称，模拟输入百度进行检索，模拟点击打开其对应的百度百科，爬取其信息中的简称、别名一栏的信息，构建地名别名库。

如遇到地名对应地名于百度百科中没有相关别名记录，则单独记录下这些地名，如果此类地名过多，则再去寻找其他可行的地名别名数据库；若不多，则可以人工查询标注。

若地名词条百度百科没有收录，则单独记录下这些地名，百度百科收录词较为全面，如有疏漏也应不多，这些地名可以人工检索。为避免没有百科记录无限循环，百度百科的链接通常出现了检索结果页面的前两页，若前两页没有出现百科链接，则不再翻页，认为该地名没有百科记录。

### 3.1.2 医院数据获取

医院及医疗机构数据主要作为与地理相关的信息返回，数据内容包括：医院相关医疗信息（重点科室、研究方向等）、医院地理信息（位置、交通等）、医院专家信息。这些信息源主要选取网络资源，网络上有很多按行政区域划分的中国医院名录。

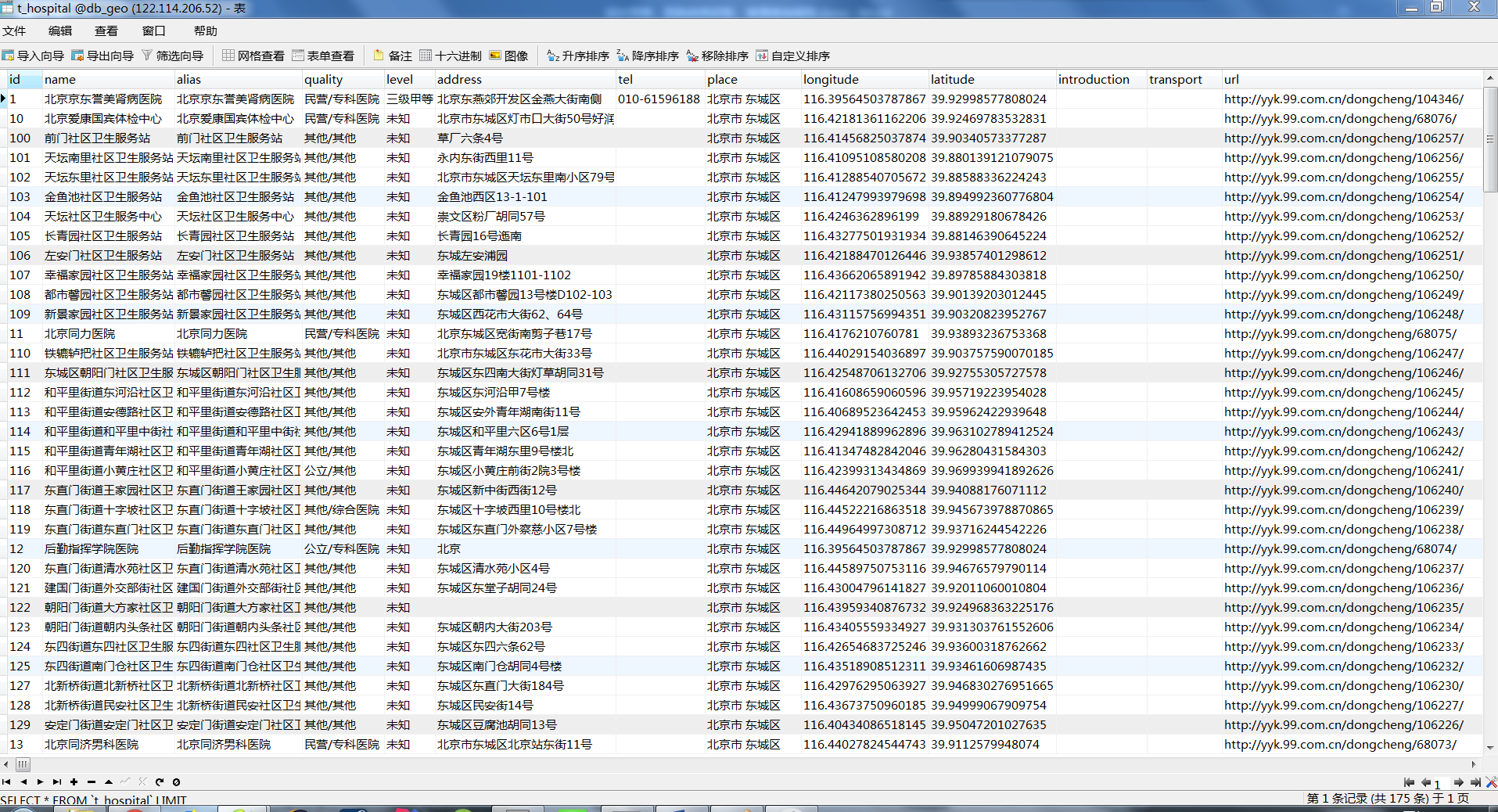
例如：<http://yyk.99.com.cn>



分析页面源码可知，该页面结构简单，抓取容易，可以直接使用爬虫获取相关数据，部分代码如下：



试验抓取部分数据存入数据库中，发现可以简单的实现：



## 3.2本体构建（杨艳琪、张莹）

### 3.2.1 本体设计

地理信息相关知识库的构建总会涉及到数据字典、地名词典、语义规则、领域专家知识、本体和元数据规范等信息。

本体则是地理知识库的核心，它详细描述了地理概念的内涵和层次结构。我们在对本体的概念、描述语言、推理以及它与地理空间信息的关系进行深入研究的基础上，结合地理位置与领域知识医疗信息库要求的特点，定义了place、hospital、expert三个本体，以下是本体构建的过程。

1995年,Gruber提出了构建本体模型的5条基本原则,也是我们构建本体的主要依据：

1.完全性原则，所给出概念术语的定义必须是完整的,并且能够充分表达和描述概念的内涵和外延。

2.明确性原则，本体模型需要用自然语言对其概念进行明确的形式化表达。

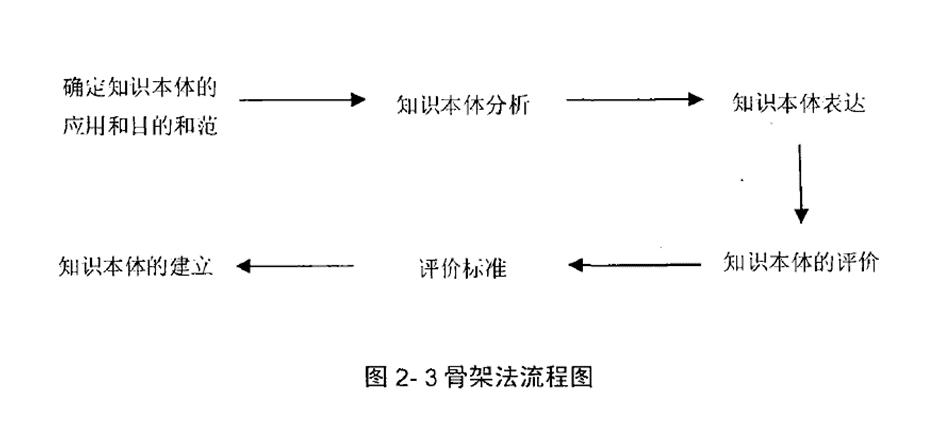
3.可扩展性原则向现有本体模型中添加新的概念时,不需要修改己有的内容,便于扩展。

4.最小承诺原则，针对本体模型的概念应提出尽可能少的约束条件。

5.一致性原则，由本体推理得来的结论必须与概念本身的含义是一致的。

在以上5条基本原则之外,构建本体还需要遵循以下三条原则：差别性原则、层次分散化原则、模块化原则。

明确了本体构建的原则，就要开始按照已有方法进行本体构建。本体构建的一般流程如下：



（图2 本体构建流程）

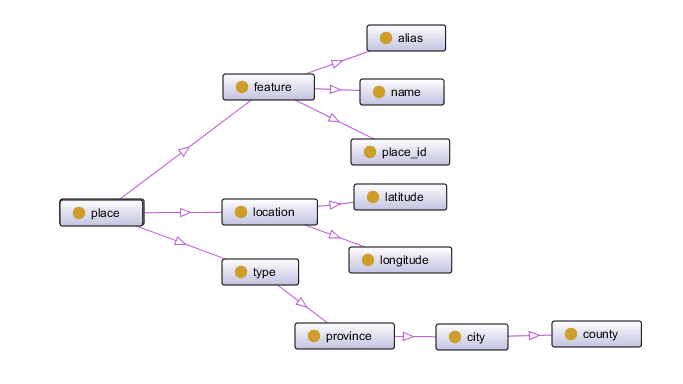
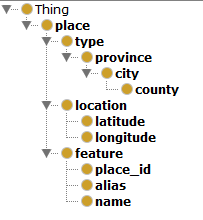
### 3.2.2 本体构建及实例

首先我们需要进行需求分析，确定本体模型的应用目的、用途和范围。在地理位置与领域知识医疗信息库中，本体的作用主要是关联相关概念，便于确定相关查询规则，使知识库中的信息能够以有序的形式存储。

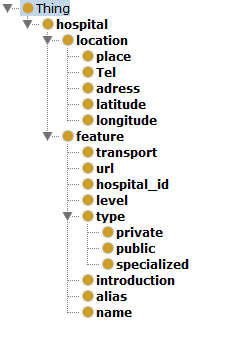
之后我们就可以开始构建本体。主要包括以下三个步骤本体捕获、本体编码和集成本体。本体捕获阶段的主要任务是收集领域内所有与研究目的相关的概念、关系以及属性,明确定义这些概念、属性和关系,然后识别出所有具有相同语义表达和关系描述的术语；本体编码主要是利用本体描述语言如OWL等,形式化的描述本体捕获阶段对术语概念化成果；集成本体是指建立现有本体和新建本体之间的映射关系,实现现有本体的扩展。

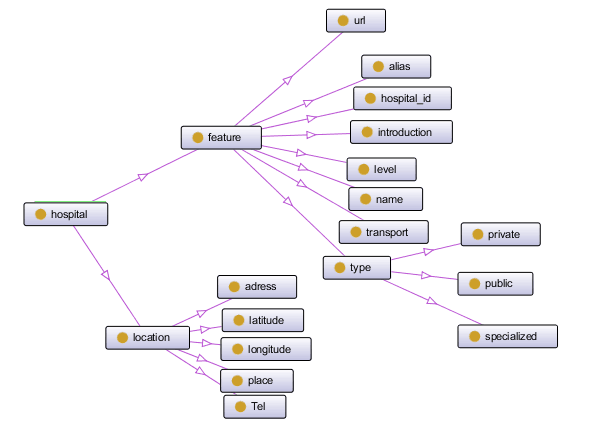
我们采用了protégé工具建立和编辑本体模型，该软件是由美国斯坦福大学基于语言研发设计的本体编辑工具,对外开放源代码,并且应用界面友好，可以用树状的层次目录组织本体模型中的概念和属性,用户可通过点击概念树中的相应结点来编辑概念和属性，还可以对概念之间的作用规则进行可视化的限定。

以下是我们构建的place、hospital、expert三个本体：

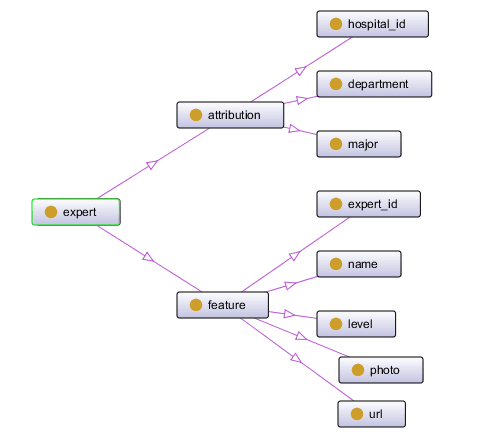
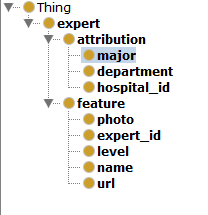


Place本体





Hospital 本体



Expert 本体

## 3.3 命名实体识别（刘政昊、程威）

### 3.3.1 简介

命名实体识别（NER）是指识别文本中具有特定意义的实体，主要包括人名、地名、组织机构名等，在本课题中主要**是对人名（专家名而非患者名）、地理名（具体的地理位置）、机构名（指医疗机构和医疗科学研究机构等）、疾病名（具体的病症）和药物名（用于治疗某一疾病的专门药物）进行识别**。

命名实体识别是将非结构化数据转为结构化数据的一个重要技术手段，是计算机正确理解文本信息的关键步骤，也是信息抽取、情感分析、问答系统等多个自然语言处理应用的基础任务，在本课题中，该工作的进行有利于后期进一步设置推理规则进行基于地理信息的推荐，因此命名实体识别在本课题中具有重要意义。

### 3.3.2 难点分析

由于中文语言自身的特殊性，中文命名实体的识别仍存在许多难点，其主要难点包括：

（1）**目前研究中的中文命名实体识别通常是基于单一模型的识别，这些模型虽然各有优点，但是对于本课题的研究来说有着很大的局限性。**如对人名的识别较为简单，采用单一的基于统计的模型即可得出较好的结果，对疾病名和药物名的实体识别则很大程度的依赖于医学词典，因此需要我们利用人工规则和其他模型算法相结合的方法进行识别，而对于地名尤其是机构名则需借助一些混合模型才能较为准确地识别。综上，在具体的命名实体识别算法设计时，应当考虑多种模型的复合或一些较新的混合模型的运用。

（2）**中文命名实体识别通常是基于词序列的识别，因此需要借助中文分词技术，**中文命名实体识别的效果往往依赖于中文分词的准确率，故选择合适的分词方法才能更好的进行识别工作。

（3）**待识别的文本中存在着大量的未登录词，即新的实体词，随着时间的推移，我们难以维护这些新词。**例如：在本研究中，一定区域范围内不断有新的医疗组织机构成立，而组织机构名的命名不存在固定的模式，即便是人直接判断该词是否是组织机构也存在一定困难，因此需要我们组建立的模型具有较强地上下文推理能力，才能够有效地识别新的实体词。

### 3.3.3 识别流程的设计

基于本课题研究的需要并结合对上述难点问题的考虑，我们组将打算按照如图1所示的流程进行命名实体识别。

我们组打算将结果集文档中涉及的地名、人名、机构名、药物名、疾病名进行命名的识别并将其统一转化为标准化地名、经纬度及相关疾病名。

其中，地名是指行政单位为区及区以上的地理位置名（其中别名问题另作考虑，详见1.5.1），识别后将所有地名按照之前构建的地理实体当中的“上一级”关系统一转化为区级行政单位的地名再进行以该地名为中心的医疗机构及相关信息的推荐。

人名在这里主要指专家名（患者信息可能不易采集或不真实，因此只识别专家名），识别出的专家名也要通过第六组或第七组所给出的专家所在的医疗机构的信息将其统一转化为医疗机构名，再将医疗机构名转化为经纬度便于直接根据百度地图API调用其周边的医疗机构来达到推荐的目的。

机构名主要包含医疗机构、大学及相关医疗研究所，识别出的这些机构名将统一转化为经纬度信息。

疾病名主要是指一些具体的病症，药物名则是治疗某种疾病的专治药物（这里也涉及到别名问题，参见1.5.1），识别出的药物名要先转化为对应的疾病名（第二组会给出疾病与药物对应关系的信息），然后疾病名要想对应到所在医疗机构，可以设计排序算法找到某疾病在哪所医疗机构治疗效果最佳从而将疾病名也转化为机构名进行处理。

结果集文档

命名实体识别

人名

（专家）

地名

（地理位置）

机构名

（医疗机构）

疾病名

药物名

对应第六组——基于网络医患交流数据的医患知识表示系统和第七组——心理健康网络咨询对话知识库

对应第一组——肿瘤术语及药物知识的关联数据基础与医学meta分析知识库第二组——用于合并的临床路径\_临床路径、药物不良相互作用知识库

保留

保留

查找所在机构转化

统一

疾病名

对医疗机构筛选排序

医疗机构名（区域面积较小）

地理位置名（区域面积较大）

经纬度

直接根据百度地图API调用该机构及周边医疗机构并根据地理知识库内信息给出这些机构的详细信息（包括机构简介等基本信息，也包括其他小组的如药物知识、医患关系的内容）

缩减范围至区级行政单位，在区内推荐相关医疗机构及其详细信息

### 3.3.4 算法设计

#### 3.3.4.1 概述

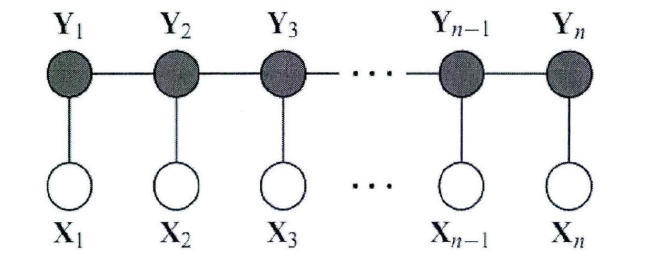
通过大量文献调研，发现命名实体识别的研究已较为成熟，早期的基于统计的诸多方法如：最大熵模型、隐马尔可夫模型、条件随机场、支持向量机等都是十分经典的算法模型，但是现实中的语言往往不满足统计学模型的条件独立假设。通过文献调研，发现**条件随机场模型（CRF）结合了最大熵模型和隐马尔可夫模型的优点，算作目前进行简单命名实体识别的一种主流方法，**但是其也存在一定的局限性，如：虽然可以通过特征选择较好的解决序列标注和划分序列问题，但是假设的当前状态只与前面的几个状态或者词有关，都假设与更前面的状态和词相互独立。此外，传统的统计模型只考虑了已经发生了的状态对当前状态的影响，而没有考虑到后面发生的状态也可能对当前的状态产生影响。

由于上述的统计模型的缺陷，在实际的应用中,自然语言处理并不完全是一个随机过程,单独使用基于统计的方法使状态搜索空间非常庞大，必须借助规则知识提前进行过滤修剪处理。目前几乎没有单纯使用统计模型而不使用规则知识的命名实体识别系统，因此结合实际需求，我们小组将使用条件随机场模型、基于人工规则和字典与统计模型相结合的方法来进行命名实体识别。

#### 3.3.4.2 条件随机场算法（CRF）

条件随机场(CRF)由Lafferty等人于2001年提出，结合了最大熵模型和隐马尔可夫模型的特点，是一种无向图模型，近年来在分词、词性标注和命名实体识别等序列标注任务中取得了很好的效果。本课题中将**采用CRF++开源工具对人名、地名、机构名进行实体识别**。

大量的命名实体识别研究表明，基于单个汉字提取上下文特征能够取得更为丰富的特征值，所以在本文中以字为单位进行文本切割，获取每个字上的上下文特征，然后采用条件随机场进行机器学习，最终实现对句子中包含的组织机构名、人名和地名的识别。



（**图3** 条件随机场模型结构，其中X为观察序列，Y为标注类别）

基于“字”的观察序列要对每个字进行标记，我们可以假设序列标注集合为L，**采用常用的三种标注符号B、I、N表示每个输入单元，分别代表实体开始、实体中间和其他**。对于人名、地名、机构名3种命名实体，我们可以定义一个7元组集合，即L={B-PER，I-PER，B-LOC，I-LOC，B-ORG，I-ORG，N}（PER表示人名，LOC表示地名，ORG表示组织机构名）[[1]](#footnote-1)

接下来要定义特征函数集合，用这个特征函数集合来为一个标注序列打分，并据此选出最佳的标注序列。定义好一组特征函数之后，还要给每个特征函数赋予相应的权重。特征函数F1（yi-1,yi,x,i）各占一定权重，是从训练语料中获取得到的，而且应该只能取0或1。假设观察函数用b（x,i）表示，特征函数的取值是由b(x,i)决定的。在满足某种条件的时候，特征函数就是观察值b(x,i).

在采用条件随机场进行命名实体识别的研究中，特征函数一般可分为上下文特征、词性特征等。本课题下，可以将特征函数写成{yt-1=L1,yt=L2,Xw=C}，其中t-1和t表示前一时刻和当前时刻，那么yt 就表示当前标记。L1，L2∈{ B-PER，I-PER，B-LOC，I-LOC，B-ORG，I-ORG，N }，W表示考察的位置，C表示W位置的内容。

#### 3.3.4.3 基于人工规则和字典与统计模型的混合算法

基于人工规则和字典的方法准确率高，因此可以作为CRF的补充修正模型，此外，对于疾病名和药物名的识别可通过医学分词词典进行直接识别。

### 3.3.5 识别过程中的其他问题

#### 3.3.5.1 别名问题的处理

对于此问题的处理，考虑到在百度百科中很多地名都有英文名、别名、简称等信息，因此打算直接使用这些信息进行别名处理（详见数据获取部分的地理别名获取）。

#### 3.3.5.2 实体排歧问题的思考与设计

通常, 实体消歧可以利用两方面的知识:一个是上下文信息, 如实体周围出现的词语;另一个是世界知识,如实体的分类体系、实体的关联等等。根据实体概念集合是否已知,命名实体消歧的任务又有所不同。在给定实体概念集合的情况下, 命名实体消歧的任务在于构建分类算法;在实体概念集合未知的情况下, 除了构建分类算法之外,还需要首先检测出底层的实体概念集合,也就是判断两个实体指称项是否指向相同的实体概念。

目前来说，命名实体消歧主要存在其算法在大规模语料上应用的问题和算法自适应性的问题，前者由于文档集合的规模并不大，因此聚类算法的时间复杂度并不算太大，但后者基于聚类的命名实体消歧方法的性能则会取决于实体指称项之间的相似度度量算法以及聚类算法的终止条件。

在综合考虑之后，我们组打算采用**基于文本向量空间的聚类方法**。即：给定一个实体指称项r 及其上下文, r 的歧义在一定程度上可被上下文所消解, 两个指向相同实体概念的指称项的上下文在一定程度上具有相似性。一个实体指称项可以表示成在其上下文中出现的词和命名实体的文本向量。基于实体的文本向量表示, 两个实体指称项的距离被表示成两个文本向量的距离。最终,基于实体指称项之间的距离,使用聚类方法来确定哪些实体指称项最终指向相同的实体概念。

## 3.4 推理查询规则（胡启军）

对于构建好的本体库，查询的输入为地理实体、经纬度和疾病名称组成的查询请求，查询结果为相关医疗机构信息、大学研究信息组成的响应。

因此本体数据库的查询过程为：

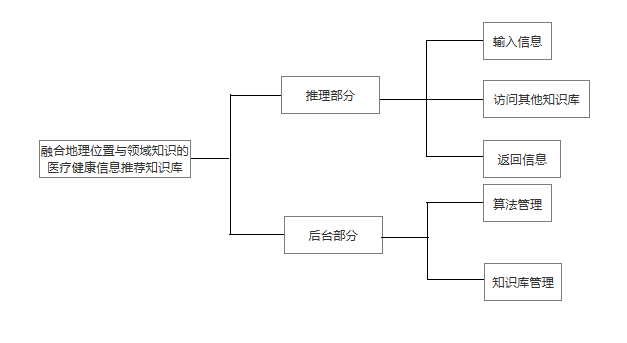
1. 对于查询请求的每一个经纬度对参数，查询附近一段距离（配置参数）内的医疗机构和大学研究院等地理实体，将这些地理实体按照距离查询点的距离排序形成集合，组成的集合记为S1
2. 对于查询的每一个地理实体名称，找到对应的地理实体对象，若地理实体为poi点，则取该点的经纬度，和①一样，查询附近医疗机构和大学并排序，将查询结果加入S1。否则，将所有行政区划地理实体组成集合T1。
3. 对T2中的行政区划集合进行整理：
4. 对于具有从属关系的行政区划，先查询范围较小的行政区划中的医疗机构等信息，若查询的信息量达到预设的标准（可配置参数），则忽略范围较大的行政区划。否则，对于较大的行政区划执行查询操作，直到没有更高级别的行政区划。
5. 对于不具有从属关系的行政区划，分别查询每个行政区划内的医疗机构、大学研究等数据，并对所有查询出的医疗知识取并集，形成S2。
6. 对S1和S2取并集，形成S3。
7. 按照查询请求中的疾病信息，与S3中每一组医疗机构和大学研究信息计算相似程度并按照一定权重（可配置参数）与之前的排序权重相加，使用最终权重排序。
8. 若S3元素数量大于预设的的查询结果数量N，则按照权重排序取前N条结果，否则返回全部排序结果。

# 4.功能逻辑设计（杨艳琦、张莹）

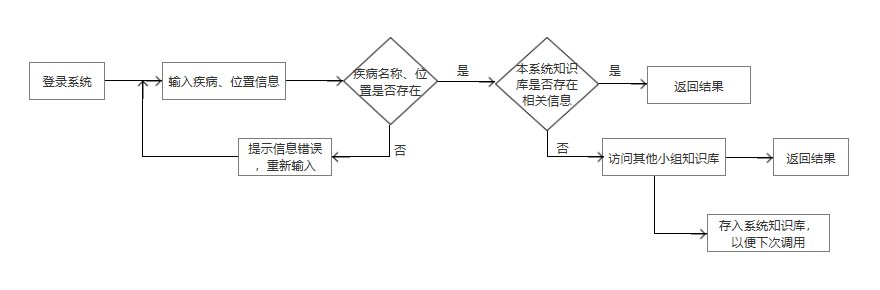
本系统主要功能有两项：一是对于普通用户，系统可以根据地理位置及疾病名称推荐该位置相近的医院信息、医疗相关研究，二是专业研究人员可通过输入某些疾病名称来返回各地的与该疾病有关的信息，达到收集信息的目的。

系统主要包括两个部分，一个是推理部分，包括用户输入信息（输入信息的正确性检验）、访问其他知识库、输出结果三个过程，用户输入自己想要得到的信息，系统通过访问其他小组的知识库，解析出相关的疾病相关信息、研究论文等，返回给用户，是一个完整获取信息的过程。

第二个是系统后台部分，包括算法管理、知识库管理，前者包括访问其他知识库的算法程序，在什么条件下调用该程序；后者是将从其他小组知识库中解析的信息加上用户搜索的关键词根据构建的本体存入本系统数据库，使得下一次再有相同搜索时不再访问其他知识库，只需要从本系统知识库中调用。



流程图如下：



用户打开系统，显示系统主界面

用户输入疾病、位置（可通过定位）等信息

系统检查该疾病名称、位置名称是否存在，存在则进入下一步，不存在则提示用户信息输入有误

判断用户需要的信息是否已存在于本系统知识库，若已存在则直接返回结果，不存在则访问其他知识库

将访问其他知识库所得的信息与用户输入的信息相关联，存入本系统知识库，以便再次调用

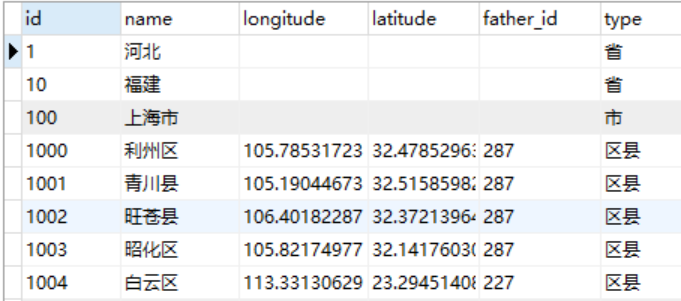
# 5.数据库设计（李紫宸）

数据分四个数据表进行存储，分别存储地点信息、专家信息、医院信息以及别名信息，从而为后期功能提供数据支持。各个数据表具体字段设计以及数据获取来源如下。

## 5.1.地点信息（t\_place）

|  |  |  |
| --- | --- | --- |
| 字段名 | 含义 | 备注 |
| id | 数据项编码 | 主码，数字编号，各不重复 |
| name | 地名及地点实体名称 | 省市区县名，以及具体地点实体名称，如“武汉大学中南医院” |
| longitude | 地点的经度信息 |  |
| latitude | 地点的纬度信息 |  |
| father\_id | 地点的上级区划 | 此字段用于确定地点的区划层级 |
| type | 地区的区划级别或地点实体类型 | 省、市、区县；医院、餐厅等等 |

数据库存储示例：



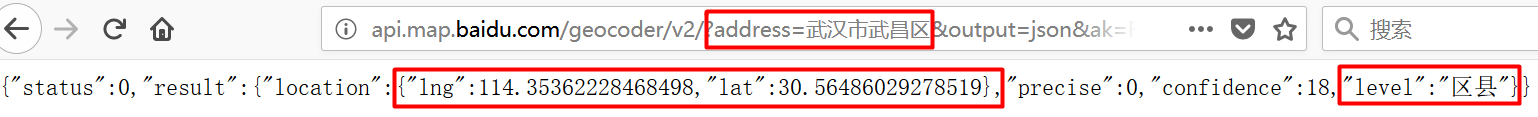
数据来源说明：

1. 行政区划名及级别

各级行政区划分为三级：省、市、区县，数据来源为中国行政区划网（www）。对该网站应用python爬虫，对各省、自治区的下属市、区县进行爬取，并按层级结构储存到数据库中。各地点根据行政区划级别，标为“省”、“市”、“区县”，上下级对应关系用father\_id字段进行描述。

②地点的经纬度信息

将区划名称作为address，调用百度api获取经纬信息（[http://api.map.baidu.com/geocoder/v2/?address=\*\*\*\*](http://api.map.baidu.com/geocoder/v2/?address=****)）。获取经纬度信息后存入数据库。如下图所示，address字段为“武汉市武昌区”，获取经纬信息如下。



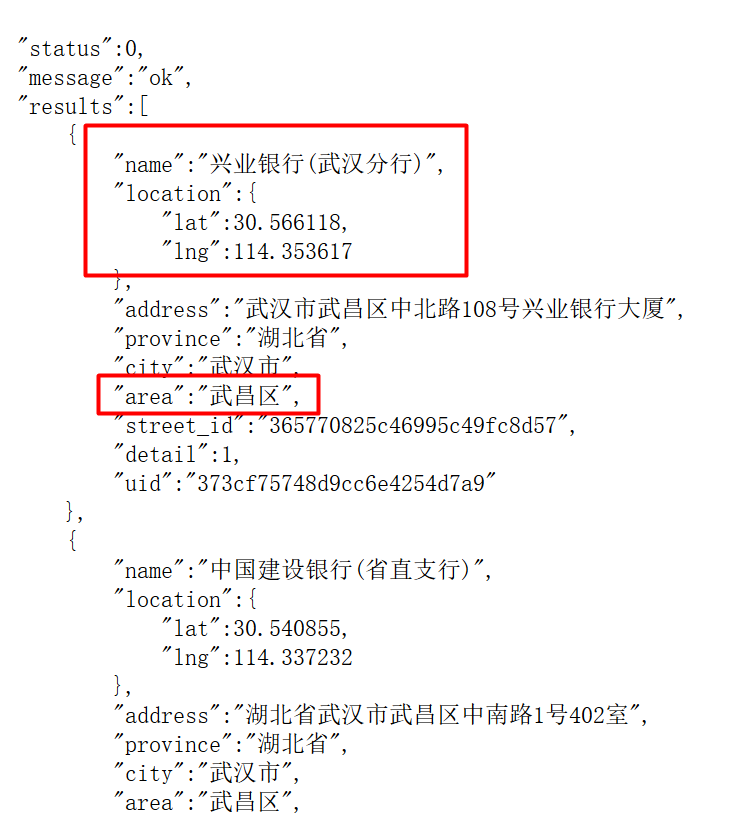
③地点实体信息

具体地点实体的信息通过调用百度WEB 服务API获得。百度地图Web服务API为开发者提供http/https接口，即开发者通过http/https形式发起检索请求，获取返回json或xml格式的检索数据。

query和tag使用百度LBS.云服务中提供的二级行业分类标签，结合region字段（区县名），获取某区县该类型的所有地点信息。



query=银行&tag=银行&region=武汉市武昌区，调用api后获取的信息如图所示，提取相关字段进行存储。



## 5.2.医院信息（t\_hospital）

|  |  |  |
| --- | --- | --- |
| 字段名 | 含义 | 备注 |
| Id | 医院编号 |  |
| Name | 医院名称 |  |
| Alias | 医院别名 |  |
| Quality | 医院性质 | 如“民营”、“公立”“专科医院” |
| Level | 医院等级 | 如“三级甲等” |
| Address | 医院地址 |  |
| Tel | 详细地址 |  |
| place | 所属区划 | 如“武汉市武昌区” |
| longitude | 经度信息 |  |
| Latitude | 纬度信息 |  |
| Introduction | 医院简介 |  |
| Transport | 交通方式 |  |
| url | 来源url |  |

数据库存储示例：



数据来源说明：

①医院信息

现已爬取北京市各医院作为示例，医院信息的各字段数据开源是99健康网，网址<http://yyk.99.com.cn。99>健康网的“医院大全”共收录医院46963家。某医院信息示例如下。，通过爬虫获取医院的详细信息，存入数据库。



②医院经纬度信息

同地点信息数据表中，调用百度api即可获取。

## 5.3.专家信息（t\_expert）

|  |  |  |
| --- | --- | --- |
| 字段名 | 含义 | 备注 |
| id | 专家编号 |  |
| hospital\_id | 所属医院编号 | 与专家所在的医院信息对应 |
| name | 专家姓名 |  |
| photo | 专家照片 |  |
| department | 所属科室 |  |
| major | 所属专业 |  |
| level | 职称 |  |
| url | 详情信息页面 |  |

数据库存储示例：



数据来源说明：

同上表，医生数据来源于99健康网，可按地区、按医院进行查询。将医生信息对应到所属的医院之中。

## 5.4.地点别名信息（t\_alias）

|  |  |  |
| --- | --- | --- |
| 字段名 | 含义 | 备注 |
| Id | 编号 |  |
| Alias | 地点别名 |  |
| type | 地点类型 | 如市、区县、银行、医院等 |
| Place\_id | 地点在地理信息表里的编号 | 与地点信息表对应地理实体的id，每个地理实体可对应多个别名。 |

# 6.界面设计（程威、刘政昊）



1. 基于规则和条件随机场的命名实体识别方法研究 程志刚 [↑](#footnote-ref-1)