∎ 班级：计科1503

学号：2015014357

北京化工大学

毕业设计（方案草稿）

**题目 领域问答系统的设计与实现**

**专 业 计算机科学与技术**

**学 生 黄奕达**

**指导教师 万静**

2019**年**4**月**12**日**

## 设计方案

领域问答系统的实现流程可以分为知识库构建，问题分类模型构建，命名实体识别，关系抽取，答案抽取。

1. 知识库构建
2. Python爬虫

Python爬虫爬取相应的信息之后，构建RDF数据，其中爬取的主要内容有疾病名称，药物，疾病简介，科室。这些内容以三元组的形式构成了知识库。

1. Jena与sparql

Jena是一个java语言的框架，可以用来处理RDF数据，sparql则是针对RDF数据的查询语言，其以带变量的RDF图的形式进行查询。

1）推荐系统数据Paper\_value.json: 包含paperAID、paperBID、以及推荐权值value。

2）图谱数据Kg.txt：包含三元组（h,r,t）即 实体-关系-实体，其中关系定义为同年出版、同一出版社出版和是否被引用关系。

3）Item\_eneity.txt: 将paperBID和 h进行一一对应。

1. 问题分类模型

### 概要设计

卡方统计量介绍：

其思想是通过判断该词在相应类别与其他类别出现与不包含的次数，进行运算，得到卡方值，其值大小就表示这这个词对于这个类别的重要性，也就是衡量这个词能否成为这个句子的关键字的因素。

通过计算：

N:训练数据集中句子总数

A:在该类别中包含该字的句子总数

B:在其他类别中包含该字的句子总数

C:在该类别不包含该字的句子总数

D:在其他类别不包含该字的句子总数

公式如下：

（4·1）

TF-IDF介绍

对选出的关键字计算权值，从而为每一个问题构成向量，其中TF是词频，IDF是逆文本频率指数。二者的乘积就是要求得到权值，TF-IDF。TF，也就是词频，其计算方法是计算这个词在相应的句子中的出现次数与这个句子中所有的词出现的次数之和的比值。也就是，对于句子中的词i和句子j，TF计算公式如下

（4·2）

就是IDF是逆文本频率指数。对于词i，N为句子总数，Ni为含有词i的句子总数，IDF计算公式如下：

(4·3)

TF-IDF的值便由TF与IDF相乘而获得。

SVM介绍：

SVM分类器的主要思想是，对于已经给出的，分好类别的一系列样本点，寻找超平面，使得距离这个超平面最近的点到这个超平面的距离最大，当超平面确定完成，分类器也完成了的训练。而当给出的数据是不可分时，SVM的核函数会帮助解决这一问题，通过将数据映射到高维空间，原始空间中的线性不可分问题就得以解决。也就是说，这时候的SVM是现在低维空间下完成计算，然后在高维空间中找到超平面。常用的核函数有多项式核函数，RBF核函数等。

语义依存树介绍：

分析句子各个语言单位之间的语义关联，并将语义关联以依存结构呈现。最终以树的结构将句子的各个部分的关系呈现出来。

深度学习算法介绍：

深度学习的概念由Hinton等人于2006年提出。基于深度置信网络(DBN)提出非监督贪心逐层训练算法，为解决深层结构相关的优化难题带来希望，主要有卷积神经网络（CNN），长短期记忆神经网络（LSTM）等。

Dijskra算法介绍：

该算法用于找到一个图中点到点之间的最短路径。设G=(V,E)是一个带权有向图，把图中顶点集合V分为两组，第一组为已求出最短路径的顶点集合（用S表示，初始时S中只有一个源点，以后每求得一条最短路径 , 就将加入到集合S中，直到全部顶点都加入到S中，算法就结束了）,

第二组为其余未确定最短路径的顶点集合（用U表示）,按最短路径的的递增次序依次把第二组中的顶点加入S中。在加入的过程中，总保持从源点v到S中各个顶点的最短路径长度不大于从源点v到U中任何路径的长度。

此外，每个顶点对应一个距离，S中的顶点的距离就是从v到此顶点的最短路径长度，U中的顶点的距离，是从v到此顶点只包括S中的顶点为中间顶点的当前路径的最短长度。

1. 输入输出结果设计

输入：

用户输入一个句子

输出1：

应当输出对应的答案，目前是问题的分类。

## 具体程序设计

1. Initprocess.java 初始化函数

作用：载入训练好的SVM模型，加载问题回答模板。

处理过程：

1. 将知识库kb.ttl，词向量载入程序
2. 将问题分类数据集载入程序，为SVM添加向量
3. 将问题最终的回答模板载入程序
4. QuestionAnalyzer.java 问题分析模块

作用：对于用户的输入进行分析，主要进行命名实体识别，关系抽取。

处理过程：

1. 将各个部分需要的参数进行装载
2. 调用data\_load.py函数
3. 调用train.py函数
4. Pretreatment.java 预处理模块

作用：对于问题分类模板进行分词，并且计算卡方统计量，选择关键字后再计算TF-IDF，最终生成向量并加入SVM。

处理过程：

1. 从两个问题分类文件中读取数据

2、分词并计算卡方统计量，选择关键字，并且计算TF-IDF，生成向量

3、将向量加入SVM

1. QAWordSegmenter.java 分词模块

作用：对于要进行的分词操作，在这里写入分词算法实现。

参数：输入的句子

处理过程：

1、接收数据并且与词向量表中的数据匹配

2、对于每一个存在于表中的词，将其作为一个分词单位保存在句中，以此实现分词