

密级：_____

浙江大学

硕士学位论文



论文题目 基于中草药语义网的
自动问答系统的研究与实现

作者姓名 钱宏泽

指导教师 张引 副教授

学科(专业) 计算机应用技术

所在学院 计算机科学与技术

提交日期 2016-01-10

A Dissertation Submitted to Zhejiang
University for the Degree of
Master of Engineering



TITLE: Research and Implementation of
Question Answering System based on
Traditional Chinese Medicine Semantic Web

Author: Hongze Qian

Supervisor: Associate Prof. Yin Zhang

Subject: Computer Application

College: Computer Science and Technology

Submitted Date: 2016-01-10

摘要

无论何时，人们都希望在第一时间得到问题的答案。为此，自动问答一直是人工智能和自然语言处理领域中一个倍受关注并具有广泛发展前景的研究方向。随着硬件的强有力支持，以及互联网和人工智能技术的发展，也产生了像 IBM Watson 这样优秀的问答系统。但目前绝大多数的问答系统是面向开放领域的，数据相对可靠度不高，而且回答的问题也多属人工构造，对实际问题的回答往往效果不佳。因此，建立受限领域的问答系统不失为一个更加现实的方向。

中草药作为传统文化的瑰宝，千百年来为中华民族的繁衍生息做出了巨大的贡献，受到广泛关注，特别是青蒿素诺贝尔奖的获得，更是掀起了新的研究和学习热潮。论文基于大量专业中草药数据，重点研究和实现中草药领域的自动问答系统，具有十分重大的意义。

本文主要工作如下：

- 1) 研究语义网的相关理论，设计中草药领域顶层本体，并且通过对数据库的重构，实现信息自动抽取工具，最终构建成拥有 300 多万三元组的中草药语义网。
- 2) 研究自动问答的关键技术，分析问题特点，提出一种层次过滤式的 SPARQL 查询语句生成策略，由上而下分别基于模式匹配、基于领域知识和基于机器学习。采用不同的组合特征进行对比实验，论证了相关方法的可行性。
- 3) 根据以上研究，设计并实现了基于中草药语义网的自动问答系统。通过对实际问题的测试，验证本文相关方法在受限领域问答的适用性以及本系统的高效性和实用性。

关键词： 自动问答，语义网，受限领域，RDF，SPARQL，中草药

Abstract

People want to get the answer to the question in the first time. For this reason, Question Answering has always been a promising research area of artificial intelligence and natural language processing. With the support of hardware as well as the development of internet and AI technology, excellent QA systems such as IBM Watson were built. But most of them are open-domain QA, the data of which is not reliable, and the questions it answers are always elaborately constructed. As a consequence, building a QA in specific-domain has more realistic significance.

As a treasure of traditional culture, Chinese herbal medicine contributes a lot to the breeding of the Chinese nation, and interests more and more people. The Nobel Prize for Artemisinin makes it a great upsurge in learning and studying Chinese herbal medicine at present. This paper is based on huge amounts of data in specific domain, and focuses in research and implementation of the QA system for Chinese herbal medicine, which has a significant meaning.

Our work is summarized as follows:

1) We study the theory of semantic web first, then design the top level ontology in traditional Chinese medicine. After reconstructing the database, we implemented a tool which can be used to extract information to build semantic web automatically. Finally we obtained about 3 million triples.

2) After research on QA, we propose an approach of parsing natural language to SPARQL, which is based on patterning matching, domain knowledge and machine learning. By combining different features as an input, the results of the experiments show this approach works well.

3) Based on the research above, we design and implement the QA based on TCM semantic web. The test of questions in reality shows that the methodology is workable in specific domain QA and the system we developed is efficient and practical.

Keywords: QA, Semantic Web, Domain Specific, RDF, SPARQL, Traditional Chinese Medicine

目录

摘要.....	i
Abstract	ii
图目录.....	IV
表目录.....	V
第 1 章 绪论	1
1.1 课题背景和意义	1
1.2 本文主要工作	3
1.3 本文组织结构	3
1.4 本章小结	4
第 2 章 问答系统相关技术研究	5
2.1 国内外发展状况	5
2.2 问答关键的组成和技术	7
2.2.1 问题解析	7
2.2.2 信息检索	11
2.2.3 答案抽取	12
2.3 相关工具	13
2.4 本章小结	14
第 3 章 语义网相关理论研究	15
3.1 语义网概念	15
3.1.1 本体	16
3.1.2 资源描述框架	17
3.2 国内外构建状况	20
3.3 相关工具	21
3.4 本章小结	21
第 4 章 中草药语义网的设计与构建	22
4.1 数据基础	22
4.1.1 关系型数据库	22
4.1.2 非结构化数据	23

4.1.3 中国中医药学主题词表	23
4.2 语义网总体结构	24
4.3 语义网详细设计	24
4.4 构建方案	26
4.5 构建成果	28
4.6 本章小结	30
第 5 章 中草药自动问答系统的关键技术	31
5.1 问答策略	31
5.1.1 问题分析	31
5.1.2 查询三元组模式	32
5.1.3 查询生成策略	33
5.2 预处理相关	34
5.2.1 领域词典构建	34
5.2.2 问题语料	35
5.2.3 多分词校正	38
5.2.4 领域词汇识别	39
5.3 特征设计	39
5.3.1 Doc2Vec	39
5.3.2 词袋模型	40
5.3.3 疑问词	40
5.3.4 核心关键词	41
5.3.5 主谓宾特征	42
5.3.6 领域知识特征	42
5.3.7 限制特征对	42
5.4 基于模式匹配的方式	42
5.4.1 问题类型	43
5.4.2 模板设计	43
5.4.3 自定义模板及匹配算法	44
5.4.4 方法评价	45
5.5 基于领域知识的方式	45
5.5.1 问题类型	45

5.5.2 领域分类器	45
5.5.3 问句生成	46
5.5.4 方法评价	46
5.6 基于机器学习的方式	47
5.6.1 有效问题分类	47
5.6.2 问题领域分类	51
5.6.3 问题属性分类	55
5.6.4 方法评价	56
5.7 本章小结	56
第 6 章 基于语义网的自动问答的实现	57
6.1 系统架构	57
6.2 系统功能分析	58
6.3 系统实现	59
6.3.1 数据格式	60
6.3.2 支撑模块	61
6.3.3 应用模块	62
6.4 系统展示	64
6.4.1 基于模式匹配的Q1类问题	64
6.4.2 基于模式匹配的Q2类问题	64
6.4.3 基于领域知识的Q1类问题	65
6.4.4 基于机器学习的Q2类问题	66
6.5 本章小结	66
第 7 章 总结与展望	67
7.1 总结	67
7.2 展望	68
参考文献	69
攻读硕士学位期间主要的研究成果	72
致谢	73

图目录

图 2.1 问答系统架构	7
图 2.2 问题的形式化表示	10
图 3.1 Linked Data	15
图 3.2 语义网层次结构	16
图 3.3 RDF 图形化表示	18
图 4.1 中医药学主题词表	23
图 4.2 语义网概念关系（局部）	24
图 4.3 URI 与 URL 转换	25
图 5.1 层次过滤式生成策略.....	33
图 5.2 领域词典词条样例	35
图 5.3 问句分词样例	39
图 5.4 问题词向量表示	40
图 5.5 句法依存树	41
图 5.6 doc2ve 特征分类结果对比	48
图 5.7 词数特征分类结果对比.....	49
图 5.8 有效问句分类 TF-IDF 特征分类结果对比	50
图 5.9 单一特征对比实验	52
图 5.10 组合特征对比实验.....	53
图 5.11 领域特征对比实验.....	54
图 6.1 接口测试界面	57
图 6.2 系统架构图.....	58
图 6.3 功能流程图	59
图 6.4 SPARQL 生成流程图.....	63
图 6.5 基于模式匹配的Q1类问题	64
图 6.6 基于模式匹配的Q2类问题	64
图 6.7 基于领域知识的Q1类问题例一	65
图 6.8 基于领域知识的Q1类问题例二	65
图 6.9 基于机器学习的Q2类问题例	66

表目录

表 3.1 RDF 模式	19
表 3.2 SPARQL 例句	20
表 4.1 URI 对比	25
表 4.2 功效细粒度拆分	27
表 4.3 基于配置的信息抽取算法	28
表 4.4 中草药语义网数据统计	29
表 4.5 中草药语义网样例文件	30
表 5.1 基本查询三元组模式	32
表 5.2 复合查询三元组模式	32
表 5.3 问题是否有效标注样例	36
表 5.4 问题实体领域标注样例	37
表 5.5 标注结果统计	38
表 5.6 疑问词提取算法	41
表 5.7 模式匹配对比	43
表 5.8 自定义模板	44
表 5.9 模板匹配算法	44
表 5.10 SPARQL 生成	45
表 5.11 有效问句分类 doc2ve 特征分类结果对比	47
表 5.12 有效问句分类词数特征分类结果对比	48
表 5.13 有效问句分类 TF-IDF 特征分类结果对比	49
表 5.14 单一特征对比实验	51
表 5.15 组合特征对比实验	52
表 5.16 领域特征对比实验	54
表 6.1 问题模板样例	61

第1章 绪论

1.1 课题背景和意义

互联网的迅猛发展和广泛普及,使人们可以比以往任何时候都方便地从网络上获取大量的信息,但是如何从大量信息中筛选出自己需要的信息,或对自己有用的信息,却一直没有得到很好的解决。现有的检索系统,无论是受限领域的检索还是互联网搜索引擎,一般都是基于关键字检索,这样的检索有几个方面的不足:首先,检索返回的结果往往是众多和答案关系或近或远的文本或网页,还需要用户从中进一步筛选,这样就造成了用户的极大不便;其次,用户的检索往往是比较复杂的,以几个关键词的逻辑组合来表示检索需求本身就不能很好的表达清楚真实检索意图,从而也就没法直接检索出令用户满意的答案;另外,从最根本上讲,以关键字为基础的索引匹配算法尽管简单易行,毕竟停留在语言的表层,而没有触及语义,因此检索效果很难进一步提高。

而自动问答或问答^[1](Question Answering, QA)系统允许用户以自然语言形式进行提问并直接返回精确答案,它的设计理念、运行机制及其期望结果完全有别于现有的关键字检索。目前,问答系统是人工智能和自然语言处理领域中一个倍受关注的研究方向。从面向的领域划分,问答系统可以分为开放领域问答和受限领域问答。

在国际文本检索会议(Text Retrieval Conference, TREC)以及跨语言评价论坛(Cross Language Evaluation Forum, CLEF)等机构的推动下,基于大规模文本的开放领域问答系统已经取得了很大进展,出现了NUS^[2]、BBN^[3]、Columbia^[4]等参加TREC评测的定义型问答系统,同时也产生了一系列的评价指标^[5],国内出现了哈工大基于常问问题集的问答系统^[6]等。但是,开放问答依赖网络资源,准确度不高,存在过时、矛盾、错误的信息。相比而言,受限领域问答系统有着其独有的优势:

- 1) 因受限领域问答可以应用领域知识,提高问题解析和答案抽取的准确率。
- 2) 某受限领域问答成熟的解决方案可以更容易地推广应用到如智能商务、

公共管理等其他受限领域。

在受限领域，相较于英日德的应用型问答系统，国内中科院实现了红楼梦人物关系问答系统^[7]，但其尚属研究阶段，远未达到应用阶段。中文领域的自动问答系统多数是基于经常问到问题（Frequently Asked Questions, FAQ），即通过提取问题特征进行相似度计算来返回排序后的答案，其 FAQ 的构建也大多源自网络。对中文特别是受限领域问答的研究而言，具有更大的挑战性，难点就在于中文自然语言处理存在着以下困境：

- 1) 特征提取特别依赖自然语言处理技术，而目前汉语的自然语言处理技术无论从分词还是词性标注、语法树解析上都存在很大的问题。
- 2) 中文语料库尤其是领域相关语料库及其匮乏，并且由于中文表现形式的多样性和复杂性，特征之间关联更强。

作为国家工程科技思想库基础建设之一，中国工程科技知识中心^[8](China Knowledge Centre for Engineering Sciences and Technology, CKCEST)建设，是一个以服务与应用为目标的建设和研究综合项目。中国工程科技知识中心是以汇集和加工我国工程科技领域海量数据为主要建设内容，以深度数据分析和智能获取知识为主要技术手段，以搜索、利用和辅助创新为主要服务内容的工程科技知识整合平台，并为国家工程科技思想库提供基础信息保障，为国家工程科技进步与创新提供信息支撑。针对医药卫生学部下的中医药领域，2012年启动了中草药专业知识服务系统^[9]子课题的建设。该子课题的建设目标是汇聚打通海量异构的中草药传统和现代研究成果的数据资源，通过大数据分析挖掘和知识建模表示技术，将资源分解、重组、提炼和综合，挖掘蕴藏在其中的知识，构建不断演化的中草药知识语义网，并针对中医药的核心要素中草药进行全方位、多角度的知识揭示与深入探究，为给广大用户提供更深层次的专业知识服务。作为中草药专业知识服务系统的组成部分，专业知识问答^[10]是一个互动型社区，旨在为广大用户提供中草药问答交流的平台。自动问答系统则是其中的重要组成部分，其目标在于通过充分分析用户问题，理解其提问的意图，从底层已有知识库中迅速返回准确答案，以供用户参考。

中草药专业知识服务系统经过一段时间的建设，已经积累了大量的专业资源，且拥有在此基础上分析挖掘的新知识。因此，本论文以此项目为依托，基于可信度高的专业中草药数据，研究知识建模和表示，以及问答系统的关键技术，并实现一个满足领域需求的问答系统，具有十分重要的意义。同时也希望此方法能为其他受限领域问答系统的构建提供借鉴。

1.2 本文主要工作

本文主要应用知识建模、信息检索、机器学习以及自然语言处理等技术，研究并实现了基于中草药语义网的自动问答系统。本文主要工作如下：

- 1) 研究语义网特点及技术，设计中草药顶层本体，从重构的数据库中抽取信息自动构建中草药领域语义网，最终获得拥有 300 多万三元组的语义网。
- 2) 从语义网抽取关键信息，构造领域词典。同时使用爬虫抓取了原始问题作为语料，并按照一定的准则对其标注。
- 3) 研究自动问答的关键技术，并通过分析问题特点定义了两类问题，针对不同类型的问题提出一种层次过滤式的 SPARQL 查询生成策略，分别为基于模式匹配的查询生成、基于领域知识的查询生成和基于机器学习的查询生成。
- 4) 在基于机器学习生成方式中进行重点实验，分别采用基础特征、语法特征和领域知识特征进行对比，论证方法的可行性。
- 5) 根据设计和研究，使用 JFinal 开源 WEB 框架、Redis 内存数据库以及 Apache Jena^[11] 语义网应用框架，实现基于中草药语义网的自动问答系统，并通过实际问题的测试，证明了本文相关方法的在受限领域问答的适用性以及本系统的高效性和实用性。

1.3 本文组织结构

本文的内容按照以下结构进行组织：

第一章，首先阐述课题的背景，简要介绍中草药专业知识服务系统及其知识问答子系统的建设情况，说明研究并实现自动问答的现实意义，最后介绍本文的主

要工作与组织结构。

第二章，首先介绍自动问答系统的国内外发展状况，重点阐述问答系统的关键技术，接着介绍目前基于知识库的问答系统的研究情况，最后对问答系统建设中所使用的一些工具进行说明。

第三章，本章阐述语义网的相关概念、层次结构、RDF 和 SPARQL，随后介绍国内外著名的语义网，并研究相关构建方法，最后简要说明语义网的构建工具。

第四章，首先分析系统已有的数据状况，并根据其特点，设计中草药领域的概念及其关系，并简要论述设计思想。随后重点介绍为方便语义网自动生成而对抽取信息源重构的过程，包含实体消歧、细粒度拆分等步骤。最后统计语义网构建的结果，并展示样例数据。

第五章，首先分析潜在问句的特点，阐述从自然语言解析为 SPARQL 查询语言的过程，然后提出一种层次过滤式的 SPARQL 查询生成策略，由上而下分别基于模式匹配、基于领域知识和基于机器学习。接着介绍领域词典的生成、问题语料构建、分词优化、领域词汇识别。之后设计了列举特征及其提取方式。最后详细介绍了三种 SPARQL 生成方式，并对基于机器学习的方式进行对比实验分析，证明此方案的可行性。

第六章，介绍中草药专业知识服务自动智能问答的系统架构以及详细实现，并通过实际问题的测试，证明本文相关方法的在受限领域问答的适用性以及本系统的高效性和实用性。

第七章，对全文进行总结的同时阐述目前系统存在的不足，并对未来的工作提出展望。

1.4 本章小结

本章简要介绍中草药专业知识服务系统及其知识问答子系统的建设情况，说明研究并实现自动问答的现实意义，最后介绍本文的主要工作与组织结构。

第2章 问答系统相关技术研究

问答系统,又称为人机对话系统,是信息检索的一种高级形式,其目标是“以自然语言交互的方式,准确回答用户的问题”。由于快速准确获取信息的需求越来越高,自动问答一直是目前人工智能和自然语言处理领域中研究热点,有着广泛应用前景。

问答系统的分类尚且没有严格的定义。从面向的领域划分,问答系统可以分为特定领域问答和开放领域问答。从处理对象的数据格式划分,问答系统可以分为基于结构化数据、基于半结构化数据以及基于文本数据。从系统答案的来源划分,问答系统可分为互联网问答、FAQ 问答以及知识库问答等。

2.1 国内外发展状况

问答系统的历史可以追溯到上世纪 50 年代图灵在论文《Computing Machinery and Intelligence》中提出的“机器智能”的概念。此后问答系统的发展大致可以分为 4 个阶段。

第一阶段是上世纪 60 年代是基于模式匹配的专家库,如 LUNAR、MACSYMA、BaseBall 等。这类系统的特点是可以通过自然语言完成问答,但其自动获取知识的能力存在瓶颈,同时由于采用定制模板的方法,覆盖率较低且不易扩展。

第二阶段是 90 年代基于信息检索技术的问答系统,如 Textract、Webclopedia 以及 TREC 的 QA Track 中的评测系统等,其底层数据主要是非结构化的原始文档、网页等自由文本。这类系统的特点是无需要建立大规模的知识库,但相对而言数据准确性无法保证。

第三阶段是 2000 年左右基于网络的检索式问答系统,典型系统如 START、Encart、ASKJeeves 等,通过分析网页返回答案给用户。START^[12]是世界上第一个基于 WEB 的问答系统,它由 MIT 计算机科学与人工智能实验室联合开发,主创人为 Boris Katz。不同于信息检索系统仅提供一系列结果,该系统致力于提供给用户“最准确的答案”。目前系统能回答成千上万有关地理、电影、任务、词典

的问题。START 优先使用自有的两个数据库进行答案检索，若能回答则直接返回答案；否则抽取关键词，返回相关网页链接。不同于传统的搜索引擎，这类系统往往对用户的问题进行浅层语义分析，并且根据人工维护的模板库，返回给用户最贴近的答案。

第四阶段是 2010 年左右兴起的基于知识图谱的问答，其底层就是一个庞大的知识库。典型的系统有 IBM Wason 以及 Wolfram|Alpha 等。Wolfram|Alpha 是 Stephen Wolfram 开发出的新一代的计算知识引擎，类似谷歌搜索，但又有着极大的不同，它能根据问题直接给出答案。Wolfram|Alpha 以公众和获得授权的资源为数据基础，通过发掘建立起了一个异常庞大的经过组织的数据库，最后利用高级的自然语言算法进行处理。

基于知识库的问答是目前问答系统发展的趋势，在开放领域知识库如 YAGO^[13]、DBpedia^[14]、FreeBase^[15]、NELL^[16]等以及受限领域知识库上，都有许多相关研究。

Frank 等人^[17]提出了一个基于健壮语义分析的混合 NLP 系统架构，并且实现了受限领域结构化知识库的问答。此方法在无需过多的领域知识的情况下，问题解析过程也会产生高质量的量化原型问句，并且从原型问句生成的查询语句可以高效地计算知识库的最小生成树。Zhang 等人^[18]提出了一个线性整数规划 (Integer Linear Programming, ILP) 模型，把对齐构造与查询构造过程融合在一个步骤，来解决多知识库的联合查询问题。Fader 等人^[19]通过先把问题分解为子问题，再经过问题改写、查询生成等步骤，结合语料以及知识库中学习出的成百万的规则，来实现基于精心构造知识库和 WEB 抽取知识库的问答。

与国外的研究水平相比，国内起步较晚，主要原因是中文自然语言的表现形式更加灵活，中文自然语言处理工具并不能地解决通用问题，同时国外成熟技术也很难直接应用其上。其次是中文领域的语料库十分缺失，并缺少对应的评价机制。目前国内的陆汝钤等^[20]依照“Agent 和本体是常识知识库的两大支柱”的观点，建立了一个大型的常识知识库“盘古”，并在其上构建了 Autotalk System。中国科学院计算所的曹寸根等^[21]研发的 NKI (National Knowledge Infrastructure) 知

识问答系统，包括地理、人物等 16 学科领域共 23 个知识库，并支持自然语言形式的查询。百度等搜索引擎也开始基于知识图谱提供简单自然问句的回答。

而在受限领域，多是的社区问答或基于 FAQ 的问答。存在部分实验性质的基于知识库的问答，如基于美食本体、农业本体的问答，但是底层数据很少，也没有应用于实际。

2.2 问答关键的组成和技术

问答系统本质上是信息检索系统，只是它从问句中获取更多的信息，返回更加精确的答案。一般的问答系统都会包括如图 2.1 所示三部分：问题解析、信息检索以及答案抽取。

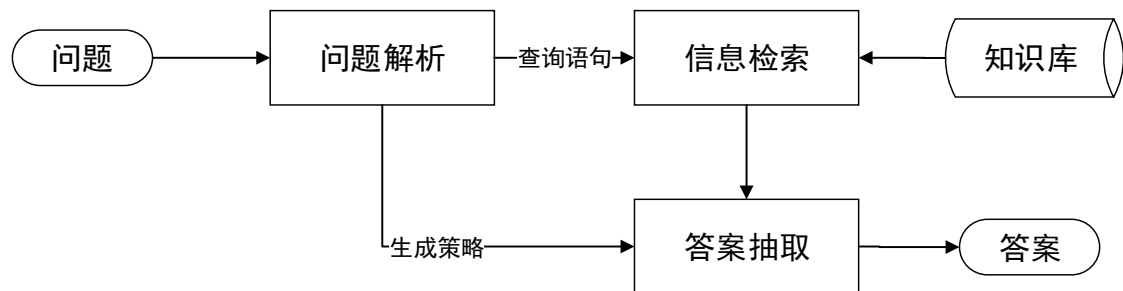


图 2.1 问答系统架构

2.2.1 问题解析

问题解析的目的是充分理解用户的查询意图，过程中主要通过自然语言工具对问题进行预处理，提取特征信息，确立问题关键词、类别等信息，并用于之后的信息检索和答案抽取过程。这部分主要包括分词、词法标注、句法分析、命名实体识别、问题分类、问题扩展等。

1) 分词

词是最小的能够独立活动的有意义的语言成分，中文分词是中文自然语言处理的基础技术，并且对之后处理结果有很大的影响。中英文分词存在很大的区别，英文单词之间是以空格作为自然分界符的，而中文是以字为基本的书写单位，词语之间没有明显的区分标记。分词中最常见的是基于规则的词典匹配的方法，当出现歧义分词时，也有最大切分（向前、向后、前后结合）、最少切分、全切分等

策略，但都存在一定不足。在受限领域的分词，都需要构造自身的领域词典，来提高分词的准确率。

2) 词性标注

词性（Part-of-Speech, POS）指根据词的特点划分词类，中文词性标注目前没有统一明确的规范。CTB（Penn Chinese TreeBank）是相对广泛使用的词性标注集，定义词性为 11 个大类和 33 个小类。有效的词性标注方法可分为基于规则的方法和基于统计的方法两类，且后者效果较好。

3) 句法分析

句法分析是指对自然语言的语法结构进行形式化定义，可以划分为短语结构语法与依存语法。依存句法的算法有 Yamada 算法和 Nivre 算法等。目前中文领域的句法分析都不理想。

4) 命名实体识别

命名实体识别的任务就是识别出待处理文本中三大类(实体类、时间类和数字类)、七小类(人名、机构名、地名、时间、日期、货币和百分比)命名实体。在中文领域，其任务还包括英文命名实体的识别。但在受限领域，命名实体也可定义为识别该领域的专有名词。

5) 问题分类

问题分类即根据问题的特征把问题归为预先定义好的一个类。由于问题也属于文本，所以可以借鉴文本分类的方法。但问题又属于短文本，所含的词汇信息以及上下文环境很少，为提高分类准确率常需要对问题做更深层次的分析。常用的方法是借助语义分析来扩展问题特征，也可以借助语义词典如 WordNet^[22]、HowNet^[23]等对问题上下位词、同义词进行扩充。准确的问题分类作用巨大。在进行信息检索时，可以缩小搜索域，更能命中目标信息；而在答案抽取阶段，也可以针对分类而应用不同的方法和策略。问题分类主要有基于模式规则匹配的方法和基于统计学习的方法，其中机器学习的方法占据主导。在一个问句中疑问词常带有十分重要的信息，所以可以定制针对疑问词的规则来判断问题，并且简单直观，如 where 表示位置，who 表示人物，when 表示时间，而 how 以及 which 则通

过后接的名词或形容词进行判别。基于模式匹配优点是无需语料库，也没有人工标注的错误率和工作量，同时还可以保证不错的效果。但由于中文表述形式的灵活性，许多问句甚至不含疑问词，所以规则方法适用性不强。而在统计学习方面，常应用机器学习分类算法，如 K 临近距离^[24]（K-Nearest Neighbor, KNN）、支持向量机^[25]（Support Vector Machine, SVM）、朴素贝叶斯^[26]（Naive Bayes, NB）、决策树^[27]（Decision Tree）等。Dell Zhang 等^[28]人采用类似语法分析树的树核，使用 SVM 对英文问题进行分类，准确率达到了 90%。但是此方法十分依赖句法分析的精度，所以在相对不成熟的中文领域应用较难。Kyoungman Bae 等人^[29]从类别的角度分析了单词分布，提出了一种基于语言模型的分类方法，可以克服短文本特征稀疏的问题。Xin Li 等人提出的层次分类法，利用 WordNet 选择词汇、语块、命名实体、中心词、相关词等作为特征，使用 Snow 分类器，在 UIUC 问题集 6 大类与 50 小类的分类准确率上分别到达了 91%和 82.4%。Zhiheng Huang 等人^[30]利用 SVM 和 EM，将焦点词和 WordNet 的上位词作为分类特征输入，最后分别达到了 89.2%和 89.0%的准确率。以上方法都是针对英文问题，在中文领域，文勋等人^[31]提出了一种特征提取的新方法，此方法依赖于句法分析结果，通过把主干词和疑问词及其附属成分作为 BN 的特征输入，在大类和小类的准确率分别是 86.62%和 71.92%。孙景广等人^[32]利用 HowNet 语义词典，把疑问句、句法结构、疑问意向词在知网中的首义原作为特征输入，采用 EM 分类器，在大类和小类的分类精度分别达到了 92.18%和 83.86%。牛彦清等人^[33]通过提取问题疑问词、关键核心词的主要义原、核心关键词的首义原、问句主谓宾的主要义原、命名实体、名词单复数等六种特征，采用 SVM 分类器对事实疑问句进行不同特征组合的分类对比试验取得了大类 91.78%和小类 79.86%的准确率，并且时间和空间复杂度相比之前有极大的下降。

6) 问题扩展

由于短文本在向量空间模型（Vector Space Model, VSM）中特征表示的极度稀疏性，为提高问题的信息量，需要对问题进行扩展，丰富原语句的信息表达，扩大信息检索的覆盖面。目前有两种主流的方式，一是通过搜索引擎等外部文本

扩展^[34],或者借助知识库如 WordNet 或 Wikipedia 等,挖掘词之间的内在联系^[35],如问题分类所述。

问题解析过程在基于知识库的问答系统中尤为重要,其主流方法有两类,一类是基于符号的表示方法,另一类基于深度学习的分布式表示方法。

基于符号的表示方法把问题问句表示为形式化的查询形式,如逻辑表达式、Lambda Calculus、DCS-TREE^[36]或 Fun-QL 等形式,之后再转化为对应的查询语言如 SQL、SPARQL、Prolog、FunQL 等,其过程如图 2.2 所示。

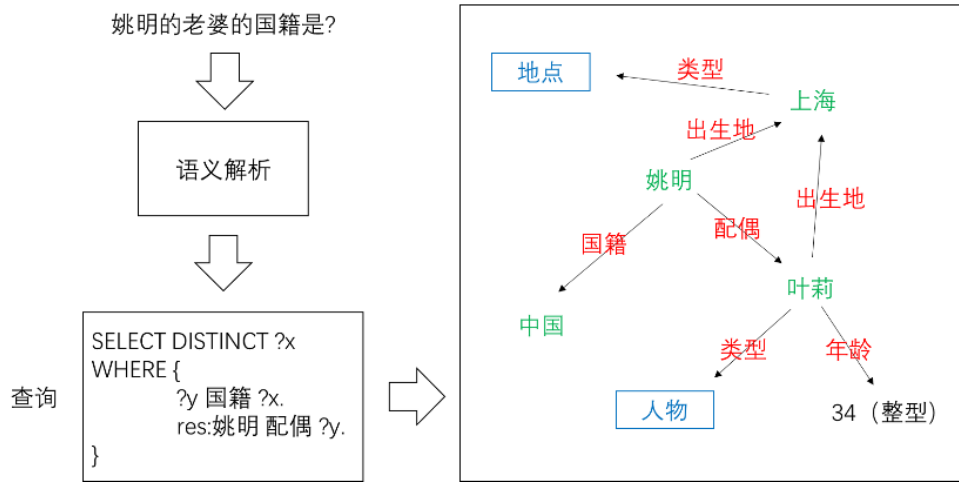


图 2.2 问题的形式化表示

问题解析已有许多成熟的研究,过程中需要解决两个关键问题,短语到资源的映射和文本歧义。Zettlemoyer 等人^[37]提出的组合范畴语法 (Combinatory Categorical Grammars, CCG) 采用词典和组合规则,把自然语言看成由词构成词组、词组构成语句的符号系统,自然语言的构造是一种计算推演过程。Wong 等人^[38]使用统计机器翻译的技术,构建了一个同步的上下文无关文法语义解析器,能从问句生成器正确的逻辑形式,但它的需要标注的语句以及其正确的逻辑结构。

由于深度学习在图像领域取得的巨大成就,其方法也开始应用于自动问答领域。深度学习的优势是在无 NLP 预处理以及人工特征设计的情况下,也能产生不错的效果。Yin 等人^[39]使用深度卷积神经网络匹配问句和谓词序列,将语义解析

过程分解为查询子图生成过程，并且通过分层的剪枝过程来缩小搜索空间，同时简化了语义匹配问题。Bordes 等人^[40]提出了将答案和问题映射到同一空间的方法，其优势是只需要少量预先标注的语料，并且不需要词法以及语法解析的过程。

2.2.2 信息检索

信息检索则以问题解析模块的结果作为输入，从底层知识库中返回一系列相关的排序文档。其目的是通过缩小候选答案域的范围，提高答案抽取的精度。检索常用的模型有布尔模型、向量空间模型以及概率模型。

1) 布尔模型

布尔模型是一种简单检索模型，基于集合论和布尔代数。其查询由联接符 AND、OR 和 NOT 构成，通过对每个关键词对应的倒排索引取交集、并集或补集，返回若干相关文档给用户。

2) 空间向量模型

向量空间模型是现在的文本检索系统以及网络搜索引擎的基础，它把文档以及用户的查询都表示成向量空间中的点，用它们之间夹角的余弦值作为相似性度量。若文档有 K 个词，则每篇文档和问句都可以表示为一个 K 维向量，如公式 2.1 所示：

$$doc_j = (w_{1j}, w_{2j}, w_{3j} \cdots w_{Kj}) \quad \text{公式(2.1)}$$

其中 w_{ij} 表示词项 i 在文档中 j 中所占的权重。其中 $w_{1,i}$ 可以是 One-hot 表示，也可以是文档单词计数，但是更常用的是加权的 TF-IDF^[41]。假设文档集合大小为 N ， df_i 为特征词 t_i 出现过的次数， f_{ij} 为特征词在某篇文档 doc_j 中的次数，则由公式 2.2-2.4 定义：

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j} \cdots f_{Kj}\}} \quad \text{公式(2.2)}$$

$$idf_{ij} = \log \frac{N}{df_{ij}} \quad \text{公式(2.3)}$$

$$w_{ij} = tf_{ij} \times idf_{ij} \quad \text{公式(2.4)}$$

同理问句也可表示为一个 K 维向量。则问句和文档的相似度可以使用余弦距离计算公式 2.5 获得：

$$\cos(doc_j, q) = \frac{\sum_{i=1}^K w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^K w_{ij}^2} \times \sqrt{\sum_{i=1}^K w_{iq}^2}} \quad \text{公式(2.5)}$$

计算结果的值越大，表明相似性越高，返回结果中排名越靠前。

3) 概率检索模型

概率检索模型通常利用关键词作为线索，通过统计得到每个关键词在相关的文档集中出现和不出现的概率以及其在与该查询不相关的文档集中出现和不出现的概率，最终根据这些概率值，计算问句和文档的相似度。为了方便计算，该模型通常关键词在相关文档集中的分布相互独立，即使这一假设与实际情况不完全一致。

2.2.3 答案抽取

答案抽取即从返回的文档列表抽取更加精确的答案。

1) 模式匹配^[42]

通过预定义的模式去匹配答案，该模式可以手工创建，也可以通过机器学习生成。例如问句“姚明出生于什么地方”，其答案通常为“姚明出身于上海”，则其答案模式可设置为“<人物>出生于<地点>”。Cui 等人^[43]提出一种基于软模式的处理方法，进行定义类问题的答案抽取。

2) 关系预定义

关系抽取的一种有效的实现方式是将句子转换为由主语、谓语、宾语构成的三元组形式，然后从中抽取答案。也可以建立问题到答案的中间逻辑表示，再通过句法树规则进行计算获得^[44]。

3) 语义相似度

语义相似度即对句子进行相似度评判，通常需要语义词典如 WordNet、HowNet 等作为辅助资源。目前也可以对百科语料进行 Word2vec 处理后，用其结果生成的词向量计算相似度。

对于两个词 w_1 和 w_2 ，定义其相似度为 $\text{sim}(w_1, w_2)$ ，其词距为 $\text{dis}(w_1, w_2)$ ， α 为一个可调参数，则有如公式 2.6 所示的转换关系：

$$\text{sim}(w_1, w_2) = \frac{\alpha}{\text{sim}(w_1, w_2) + \alpha} \quad \text{公式(2.6)}$$

定义两个句子a和b，其中a包含a₁、a₂、a₃、……、a_m共m个词，b包含b₁、b₂、b₃、……、b_n共n个词，定义词间相似度为sim(a_i, b_j)，则任意两个词相似度可由公式2.7-2.9计算获得：

$$\text{sim}(a, b) = \left[\frac{\sum_{i=1}^m sa_i}{m} + \frac{\sum_{j=1}^n sb_j}{n} \right] / 2 \quad \text{公式(2.7)}$$

$$sa_i = \max(\text{sim}(a_i, b_1), \text{sim}(a_i, b_2), \text{sim}(a_i, b_3), \dots, \text{sim}(a_i, b_n)) \quad \text{公式(2.8)}$$

$$sb_j = \max(\text{sim}(a_1, b_j), \text{sim}(a_2, b_j), \text{sim}(a_3, b_j), \dots, \text{sim}(a_m, b_j)) \quad \text{公式(2.9)}$$

计算所得结果越大，说明相似度越高。

2.3 相关工具

目前已有许多技术成熟的开源工具，合理应用可以简化许多处理任务，从而把研究重心放在关键点的解决。

1) 中文自然语言处理工具

FNLPL^[45]是复旦大学为中文自然语言处理而开发的工具包，提供分词、词性标注、语法树解析等功能，同时也包含为实现这些功能的机器学习算法和数据集。FNLPL采用Java编写，使得它可以轻松运行在各种不同的平台之上，不仅可以通通过命令行调用，同时也可以十分便捷的集成到各种Java项目之中。

HanLP^[46]是另一个Java中文自然语言处理工具包，由一系列模型与算法组成，其目标是普及自然语言处理在生产环境中的应用。HanLP不仅提供分词的基本功能，而且提供词法分析、句法分析、语义理解等其他的功能。官方模型训练自2014人民日报语料库，同时也允许用户使用内置的工具训练自己的模型。

2) Weka^[47]

Weka (Waikato Environment for Knowledge Analysis) 作为一个公开的数据挖掘工作平台，集合了大量能承担数据挖掘任务的机器学习算法，包括对数据进行预处理，分类，回归、聚类、关联规则以及在新的交互式界面上的可视化。对于分类领域，其提供了Navie Bayes、Decision Tree、Logistic等多种算法，同时也可

以通过导入林智仁^[48]的 Libsvm 包实现 SVM 算法。此外 Weka 也提供 Java 调用接口供系统集成。

3) Word2vec & Doc2Vec

Word2vec 是 Google 的一款开源工具，用于将词表征为实数值向量，其采用的模型有 CBOW 和 Skip-Gram 两种。由其训练获得的词向量不可直观解读，但是依然带有语义信息，例如官方例子： $\text{vector}(\text{'Rome'}) = \text{vector}(\text{'Paris'}) - \text{vector}(\text{'France'}) + \text{vector}(\text{'Italy'})$ 。通过 Word2vec 可以把词之间的运算简化为向量空间中的向量运算，如计算向量空间上的相似度，来表示文本语义上的相似度。Doc2Vec 的提出建立在 Word2vec，除增加一个段落向量外，方法几乎等同于 Word2vec。其作用是把文档映射到一个向量空间。

2.4 本章小结

本章首先介绍自动问答系统的国内外发展状况，重点阐述问答系统的关键技术。接着，本章介绍目前基于知识库的问答系统的研究情况，最后说明问答系统建设中所使用的一些工具。

第3章 语义网相关理论研究

语义网^{[49][50]} (Semantic Web) 是自然语言理解及认知科学领域研究中的一个重要概念,是现有万维网的扩展与延伸,由 Tim Berners Lee 提出,目的是使计算机更能解读万维网,并通过链接网络上分布着的海量数据,构建一个更为庞大的语义网,如图 3.1 所示。

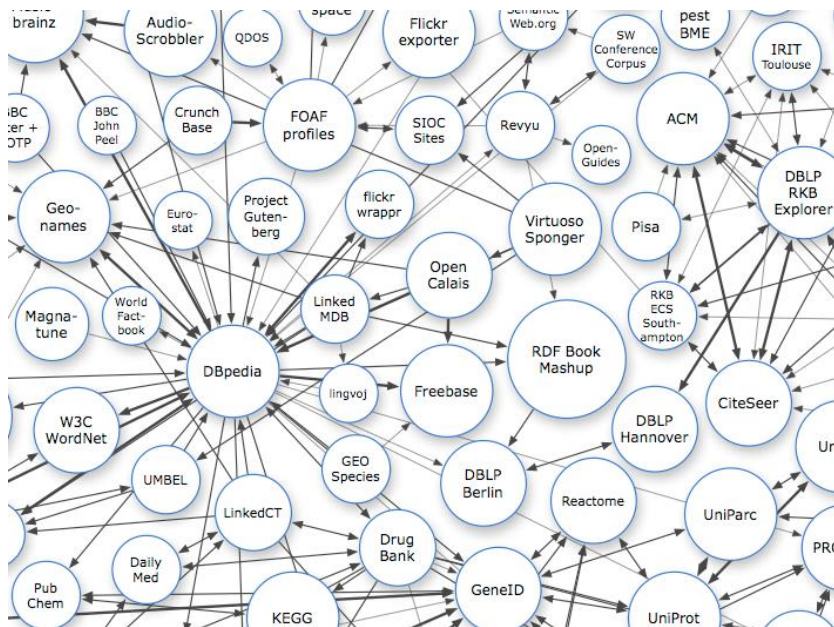


图 3.1 Linked Data

语义网的设计原则可概括如下:

- 1) 使得结构化和半结构化的数据以标准化的格式在万维网上可用。
- 2) 不仅制造万维网上的数据集,同时还创建可解读的个体数据元素及其关系。
- 3) 使用形式化模型来描述这些数据的隐含语义,使得这些隐含语义能够被机器处理。

3.1 语义网概念

语义网的组成包括元数据、资源描述框架以及本体,其中元数据是具有“语

义”的描述数据的数据，资源描述框架则是存放具体资源的格式或结构，而本体用于处理相同概念不同形式下的属性和关系。此外，语义网还具有一定的层次结构，其层次的构建必须遵循两个基本原则：向下兼容性和向上部分理解。向下兼容性指某一层次的代理应该解释和使用更低层次的信息，如 OWL 语义应该可以理解 RDF 和 RDFS 中的信息。向上理解部分信息指某一层次的代理至少应该可以利用部分上层的信息，如 RDF 和 RDFS 可以解释部分的 OWL 知识。但是即使存在理论指导，实践上却存在一定的妥协。

目前语义网的 7 层体系结构如图 3.2 所示，各层功能自下而上增强。

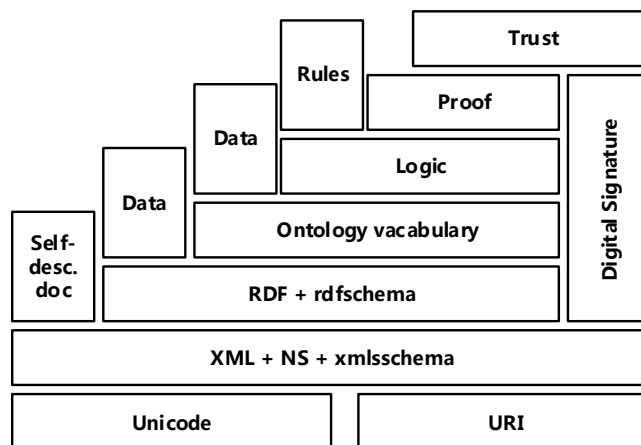


图 3.2 语义网层次结构

3.1.1 本体

本体 (Ontology) 原本是一个哲学上的概念，引入人工智能领域后被定义为“共享概念模型的明确的形式化规范说明”，其目标是捕获领域知识，提取共同理解，确定认可词汇，最终从不同层次的形式化模式上明确定义词汇之间的相互关系。领域本体是由词汇及其规则体系如对象、属性和关联构成。对象表示领域实体、属性对象的特性，关联表示对象间的关系。当本体通过特定语言表征后，便可被特定的代理如 RDF、OWL 解析和识别。

领域本体的构建尚无系统的、工程的方法，现存的多数方法都是从各自的构

建过程中通过逆向工程总结获得。如 Uschold 本体构建法^[51]从 Enterprise 本体的经验中总结, Gruniger&Fox 本体构建法^[52]从虚拟企业本体构建中提出, Meth 本体构建法用于人工智能图书馆, Bernerasetal 本体构建法用于应用开发控制。即使各种方法领域相关, 但就一般而言, 皆包括如下六个步骤:

- 1) 确定本体的领域及范围
- 2) 列举领域中的术语和概念
- 3) 建立本体框架
- 4) 设计元本体或重用已有本体, 定义领域中概念之间的关系
- 5) 对领域本体进行编码并且形式化
- 6) 对本体进行检验和评价

3.1.2 资源描述框架

资源描述框架 (Resource Description Framework, RDF) 是一种用于描述 Web 资源的标记语言。RDF 是一个处理元数据的 XML 应用, 所谓元数据, 就是“描述数据的数据”或者“描述信息的信息”。任何数据交换语言都由 3 个组成元素: 语法、数据模型和语义。其中语法规则数据的撰写, 数据模型规范数据的结构或组织形式, 语义规范数据的解释。RDF 满足以上条件, 并且提供了一个灵活并且领域无关的数据模型。其基础构件为称为“声明”的“主体—谓词—客体”的三元组。又由于 RDF 不针对任何领域及使用, 对用户而言必须定义他们在这些声明中使用的术语, 为此引入了 RDFS (RDF Schema, RDF 模式), 允许用户精确地定义它们的词汇表。

1) RDF 数据模型

RDF 中基本概念包括资源、属性、声明和图。资源可以理解为一个对象, 拥有一个统一资源定位符 (Uniform Resource Locator, URI) 来无歧义地标识资源, 但使用 URI 不一定能访问到资源。属性用来描述资源之间的关系, 其也由 URI 标识。声明断言资源的属性, 是由一个资源、一个属性和一个属性值构成的三元组, 属性值可以是另一个资源, 也可以是文字。图是 RDF 图形化的表示, 是由带标签

的节点通过带标签的边的有向边构成，强调 RDF 是一个以图为中心的数据模型的概念。如图 3.3 所示便是 RDF 图形化的结果。主体用椭圆形标记，表示一项特定的资源，谓词用箭头标记，表示被命名的属性，客体用矩形标记，表示该资源中该属性的取值。

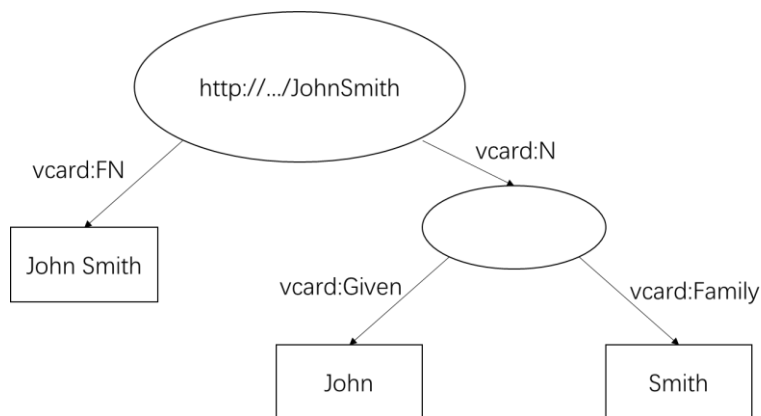


图 3.3 RDF 图形化表示

2) RDF 语法

RDF 语法有 Turtle、RDF/XML、RDFa 等多种形式，适用于不同的场景，但是其底层的数据模型和语义是完全相同的。

3) RDF 模式

RDF 模式对抽象世界中的主要关系进行描述，其包含了核心类、定义联系的核心属性以及限制属性的核心属性。详细说明如表 3.1 所示。

表 3.1 RDF 模式

类型	表示	说明
核心类	rdfs:Resource	所有资源的类
	rdfs:Class	所有类的类
	rdfs:Literal	所有字符串的类
	rdf:Property	所有属性的类
	rdf:Statement	所有具体化声明的类
定义联系的核心属性	rdf:type	将一个资源关联到它的类，资源被声明为该类的一个实例
	rdfs:subClassOf	将一个类关联到它的超类，一个类的所有实例都是其超类的实例。并且一个类可能是多个类的子类
限制属性的核心属性	rdfs:domain	指定属性的定义域
	rdfs:range	指定属性的值域

4) 应用协议

SPARQL (Simple Protocol and RDF Query Language) 是专为 RDF 开发的一种协议，用于规范信息查询和数据获取过程。虽然它的设计初衷是面向 W3C 所开发的 RDF 数据模型，但是可以应用于任何以 RDF 来表示的信息资源。

表 3.2 所示是一个典型的 SPARQL 语句。

表 3.2 SPARQL 例句

```
PREFIX jdmc: <http://xmlns.com/foaf/0.1/>
SELECT ?url
FROM      <bloggers.rdf>
WHERE {
    ?person jdmc:name "Jon Azure" .
    ? person jdmc:weblog ?url .
}
```

其中 `jdmc` 是命名空间的简称，用于简化语句缩写。`SELECT` 之后为指定查询的内容，此例中为命名为 `url` 的变量。`FROM` 是一个可选子句，表示查询数据集的 URI。最后 `WHERE` 子句由多组的三元组模式构成。在以上示例中，`WHERE` 子句的图形模式中的第一个三元组与 `jdmc:name` 属性为“Jon Azure”的节点匹配，并把它绑定到名为 `person` 的变量，此后图模型再与 `person` 的 `jdmc:weblog` 属性对应的对象匹配，绑定到 `url` 变量，形成查询结果。

除了简单的检索模式，SPARQL 还提供了各种关键词如 `OPTIONAL`、`FILTER`、`LIMIT` 等，用于实现更复杂的检索。

3.2 国内外构建状况

目前国外已存在许多语义网，拥有成熟的构建经验。其构建方法主要有两种：以人工的方式进行构建，目前比较流行的一种方式就是基于众包，且内部知识的可信度极高；第二种是计算机通过一定的算法自动从网络上抽取信息构建，相对前者而言省时省力，但存在数据错误、过时等问题，且难以发现。以下是一些著名的语义网及其构建方式。

Cyc 是一个综合集成各个领域的本体及常识知识的人工智能项目，并希望能以类似人的方式工作，实现知识的推理。Cyc 目前的大部分事实通过手工添加，并且还处于基础的知识工程建设阶段。

DBpedia 从维基百科(Wikipedia)的词条里提取出结构化的资料，并将其他资料

集连结至维基百科，以强化维基百科的搜寻功能。它是一个很特殊的语义网应用范例。DBpedia 同时也是世界上最大的多领域知识本体之一，也是 Linked Data 的一部分，DBpedia2014 版的资料集拥有超过 458 万的物件。

YAGO (Yet Another Great Ontology) 是另一个语义网应用，它从 Wikipedia、WordNet 以及 GeoNames 中抽取信息，目前有超过千万的实体以及 12 亿左右的事实，并且在样例数据上人工评测有过 95% 的准确率。

Freebase 是一个众包型的合作知识库，其内容主要来自社区成员的贡献。它以提供元数据为主，为此融合了许多网络资源，包括部分私人站点中的内容。

在国内的知识库构建则相对较迟，主要有中科院陆汝钤教授领导的研究常识为目的知识库“盘古”，并能解决相关的实际问题。曹寸根教授的构建的大规模知识系统 NKI 中包含了大量的科学本体。此外，浙江大学的高济教授以本体为基础，紧密结合知识建模、知识共享和综合集成，提出了基于表示本体的智能系统开发方法 OMSI。

3.3 相关工具

Apache Jena 是一个免费开源的 Java 语义网和链接数据应用框架，它为 RDF、RDFS、OWL 提供了一个程序开发环境。具体包括用于对 RDF 文件和模型进行处理的 RDF API，用于对基于 XML 形式的 RDF、RDFS、OWL 文件进行解析的解析器，RDF 模型的持续性存储方案，用于检索过程推理的基于规则的推理机子系统，用于对本体进行处理和操作的本体子系统，用于信息搜索的 RDQL 查询语言。相比较另一款工具 Sesame，Jena 在数据库导入、构建、查询等方面更有效率上的优势。

3.4 本章小结

本章阐述语义网的相关概念、层次结构以及 RDF 和 SPARQL，接着介绍国内外著名的语义网，并调研相关构建方法，最后简要介绍语义网构建工具。

第4章 中草药语义网的设计与构建

构建中草药语义网的首要目的是生成一个全面、关联的专业领域知识库，用作上层功能服务如检索、问答的底层数据支撑。其次，构建过程中由于需要人工对目前已有的数据做一个梳理，期间会发掘出新的概念及其关系，并且通过对概念属性和关系的定义，为后续数据的添加集成制定规范。另外，定义了专属于中草药专业知识服务系统的命名空间及其 URI，方便数据的交换传播。

语义网的概念涵盖广泛，有七层结构，本文主要实现其下 4 层作为中草药语义网的雏形。

4.1 数据基础

中草药语义网的构建依托于已有的专业结构和非结构化的异构数据。

4.1.1 关系型数据库

经过前期建设，系统数据库中已经有一定数量的专业数据，其中以单味药、方剂、疾病、证候、基源以及化合物六类的数据最为齐全，此外还包括中成药、企业、医院、医生、文献等数据。这些数据都是的单表结构，所以存在实体重复、记录稀疏、数据持续集成困难、关系对应不统一等问题，同时也造成了存储以及检索效率的下降。

1) 实体重复

若存在多条描述同一实体如“麻黄”而来源不同的数据，在数据库中则存储为多条记录，如此便造成了某些字段的重复存储，并且每一条麻黄的描述信息并不完善。

2) 记录稀疏

并非每一条新添记录的每一个字段都有值，甚至存在只有一个名字且其余字段皆为空的记录。

3) 数据持续集成困难

后续添加数据过程中若存在原结构没有的字段，则需要在结构中先添加此字段，如此也造成了先前其他记录的稀疏。同时，许多新增加的数据也缺失来源信

息。

4) 关系对应不统一

各个表之间的关系呈现形式极度不统一，编号对编号，编号对名称、名称对名称的情况皆存在。

4.1.2 非结构化数据

专业数据中也包括部分非结构化数据，例如网页自由文本、OCR 图书等。这些数据中所描述的实体，绝大多数已经存于关系型数据库中，只是对于某一属性的描述形式不同，同时非结构化数据中也存在许多关系型数据库中并不存在的字段。典型的非结构化数据源有《中药活性成分大辞典》、《中药有效成分》、《中药大辞典》、《全国中草药汇编》等。这些数据的添加将完善语义网知识的涵盖范围。

4.1.3 中国中医药学主题词表

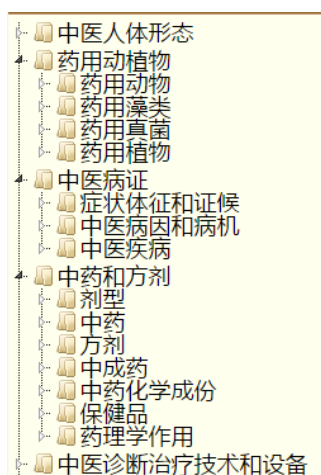


图 4.1 中医药学主题词表

《中国中医药学主题词表》由中国中医科学院中医信息研究所组织完成，是国家中医药管理局标准化研究项目。该词表是按照树状形式组织，从全面地显示了主题词之间上下位关系。词表将全部主题词按学科门类划分，排列与 15 个类目 68 个子类目。这种分类法以中医药学学科体系为基础，同时兼顾了专业特点及词汇分类的需要。该主题词表提供的分类标准，对于构造中草药领域的语义网具有

借鉴意义。其目录树如图 4.1 所示。

4.2 语义网总体结构

中草药语义网的构建主要参照 YAGO 的构建思路。YAGO 以 WordNet 作为其分类标准，从 Wikipedia 中抽取对应的信息。中草药语义网则以主题词表作为分类标准，从关系型数据库中抽取信息。

根据目前已有数据的情况，结合主题词表专业知识，本文设计中草药语义网本体概念及其关系如图 4.2 所示。

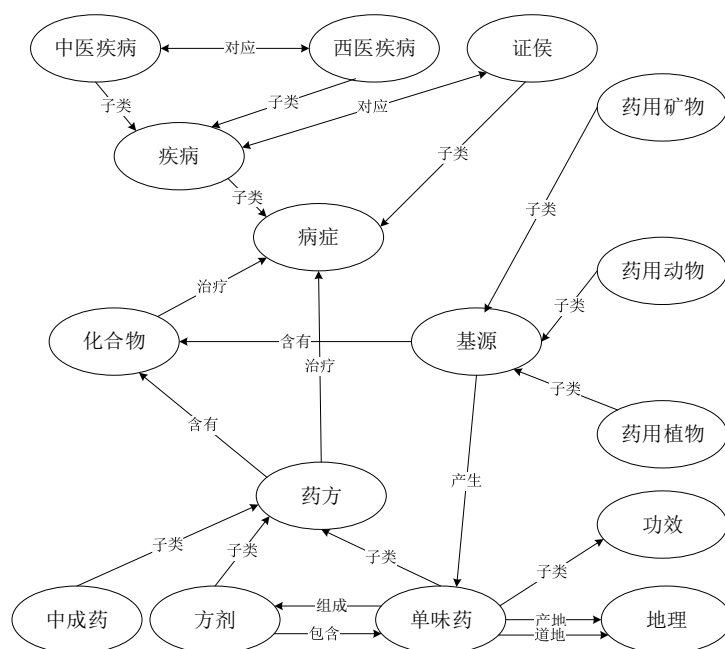


图 4.2 语义网概念关系（局部）

网络中所定义的概念中，单味药、方剂、疾病（中医疾病、西医疾病）、证候、基源（药用动物、药用植物、药用矿物）概念下的实体数量较多，属性也完整丰富，故首先从这些数据中抽取信息构建语义网。

4.3 语义网详细设计

在确定中草药语义网的概念及其关系之后，需要对其具体实现细节进行设计，

包括描述语言、URI、命名空间、概念等。考虑到专业知识背景、构建效率、实现复杂度等综合因素，最终的方案是多方面权衡下的结果。

1) 描述语言

在语义网的层次结构中，虽然 OWL 的层级高于 RDF，但最终构建选择使用 RDF，理由如下：1. 目前语义网的设计处于雏形阶段，对各种概念、实体、属性以及值的限定较少，无需用到更高级的语言，同时由于向下兼容性，今后由 RDF 转为 OWL 也较为方便；2. 许多语义网如 YAGO 是以 RDF 形式展现，具有对照参考的作用；3. 基于 RDF 的 SPARQL 语言应用广泛，开源框架相对成熟。

2) URI

在 YAGO 中，URI 以前缀加上名字的形式，指代唯一资源，而中草药语义网的 URI 是以前缀加上唯一编号的形式，如表 4.1 所示。

表 4.1 URI 对比

YAGO	<rdf:Description rdf:about="&y;Abraham_Lincoln">
中草药	<rdf:Description rdf:about="http://zcy.ckcest.cn/tcm/med#1064">

在前缀中编码分类信息，所以一个 URI 稍作规则转换便可直接访问到该资源在中草药专业知识服务系统中基础知识库中的详细信息，如图 4.3 所示。

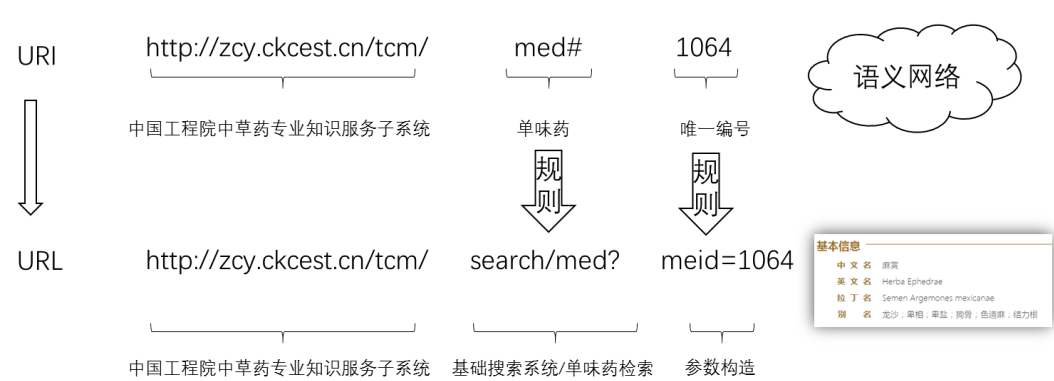


图 4.3 URI 与 URL 转换

由于 URI 中没有编码名字信息, 这种方式的缺点是其对应 SPARQL 的查询语句较前者复杂。

3) 命名空间

YAGO 虽然是通用数据库, 但是其关系以及属性比较固定, 故使用了一个唯一命名空间 “<http://www.mpii.de/yago/resource/>”, 而中草药语义网实体的属性除了中文名、英文名等, 其他差别较大, 所以采用了不同命名空间, 并且属性定义为表名加字段名的形式, 方便了语义网的自动化生成。但是对于实体间的关系, 引入了另一个独立的表示关系的命名空间。此方式也有 SPARQL 的查询语句复杂的缺点。

4) 概念表示

在中草药领域, 对单位药的功效描述形式都比较固定, 所以在设计时把功效作为一种概念, 单味药与功效是实体与实体的关系。但是由于专业知识的缺乏, 对于挖掘出的功效名称是否可以当作实体的准确性不易确定, 所以是现实时功效等暂时作为单味药的一种属性。

4.4 构建方案

语义网构建尚无系统性方法。本文在确定语义网设计之后, 作为信息抽取源, 当前数据库的结构不利于数据的自动化抽取, 所以在构建语义网之前先对数据源进行了重构和扩充, 方便信息抽取。

经过调研发现, 雪花模型^[53]的设计可以满足数据添加和抽取的需求。以化合物为例, 由原先的一张表拆成了 8 张表, 其中以一张基本表作为雪花模型的中心, 其中的每条记录都分配有作为主键的唯一编号, 代表唯一种化合物。除了编号, 其字段还包含化学式、分子式、分子量等为唯一不变的属性。其他的表还包括中文名、英文名、毒性、应用等随来源不同而值不同的表, 这些表通过外键与基本表相连, 是多对一的关系。此外, 也有一张新旧数据对照表, 此表用于表之间新关系的生成。

1) 实体消歧

实体消歧是在合并过程中对表示同一实体的信息进行合并。期间策略多以启发式规则判定，如在两个中文名相同的化合物，则比较其分子量、分子式等判断是否描述同一实体。

2) 细粒度拆分

细粒度拆分是指在信息处理过程中对存在一定模式的长文本内容进行处理，将其拆分为更小的含有语义的词组，这类文本的特征多是通过分号、逗号等标号分割或者以某关键词标注。如功效的拆分结果如表 4.2 所示。

表 4.2 功效细粒度拆分

原内容（功效主治）	拆分后的功效	拆分后的主治
健胃消食，利尿安神。治消化不良，腹胀，浮肿，膀胱炎，肺结核，失眠。	健胃消食 利尿安神	消化不良 腹胀 浮肿 膀胱炎 肺结核 失眠
清风清热；利湿解毒；散瘀消肿。主感冒发热；疟疾；关痛；高血压；泄泻；痢疾；风湿痹痛；湿疹；荨麻疹；疮疖；乳腺炎；颈淋巴结核；跌打损伤	清风清热 利湿解毒 散瘀消肿	感冒发热 疟疾 关痛 高血压 泄泻 痢疾 风湿痹痛 湿疹 荨疹 疮疖 乳腺炎 颈淋巴结核 跌打损伤

3) 来源信息

原先关系型数据库和新添的非结构化数据在新库重构的过程中，都添加了来源信息，提高了信息的可信程度。

4) 关系重建

依靠此前生成新旧关系对照表，重建实体之间以编号对应的关系。

5) 信息抽取

在完成数据库重构后，雪花状模型的结构与实体属性之间结构相似，不同雪花中心之间的结构与实体和实体之间结构相似。故可使用一套基于配置的信息抽

取算法，如表 4.3 所示，通过不同的配置文件自动完成实体与属性、实体与实体之间关系的生成。语义网的生成采用 Apache Jena 语义网应用框架。

表 4.3 基于配置的信息抽取算法

基于配置的信息抽取算法
STEP1. 配置 URI 前缀、命名空间、基本表、属性表、过滤字段（如时间戳）等信息
STEP2. 遍历基本表，读取一条表示唯一实体记录，取其主键值设为 ID，以 URI+ID 构造的方式创建一个资源 R
STEP3. 遍历基本表其他非过滤字段，以命名空间+基本表名+字段名的方式创建一个属性 P，取该字段值设为 V，以 (R, P, V) 的方式构造一个三元组
STEP4. 遍历其他属性表，获取 ID 相同的记录，遍历每一条记录，并以与 STEP3 相同的方式构造三元组
STEP5. 若基本表未遍历完毕，则重复 STEP2，否则 STEP6
STEP6. 输出 RDF/XML 形式文件

最后把生成的单独 RDF/XML 文件合并，即构造完成了中草药领域的语义网。

4.5 构建成果

鉴于语义网的构建尚未有系统性方法，本文提出的基于重构雪花模型数据库的自动信息抽取构建方案，能同时保证速度和精度，最终生成三元组共 3180259 对，其数据统计如表 4.4 所示。

表 4.4 中草药语义网数据统计

名称	数量
单味药	366369
方剂	687443
中医疾病	21840
西医疾病	41279
证候	12261
药用动物	483899
药用植物	21962
药用矿物	5841
化合物	144766
单味药化合物包含关系	36375
单味药疾病治疗关系	172392
基源单味药包含关系	9324
方剂单味药包含关系	728885
单味药证候治疗关系	308448
方剂疾病治疗关系	128216
方剂证候治疗关系	2381
疾病证候对应关系	1503
基源化合物包含关系	7736

其样例文件如表 4.5 所示。

表 4.5 中草药语义网样例文件

中草药语义网样例文件-单味药沙果（编号 4176）
<pre><rdf:Description rdf:about="http://zcy.ckcest.cn/tcm/med#4176"> <med.property:med_basic.med_name_zh>沙果 </med.property:med_basic.med_name_zh> <med.property:med_function_detail.med_function>祛风除湿 </med.property:med_function_detail.med_function> <med.property:med_function_detail.med_function>祛风湿 </med.property:med_function_detail.med_function> <med.property:med_property_detail.med_property>甘 </med.property:med_property_detail.med_property> <med.property:med_property_detail.med_property>辛 </med.property:med_property_detail.med_property> <med.property:med_property_detail.med_property>凉 </med.property:med_property_detail.med_property> <med.property:med_property_detail.med_property>涩 </med.property:med_property_detail.med_property> <med.property:med_zhuzhi_detail.med_zhuzhi>胸膜炎 </med.property:med_zhuzhi_detail.med_zhuzhi> </rdf:Description></pre>

4.6 本章小结

首先介绍系统已有的数据状况，并根据其特点，设计中草药领域的概念及其关系，并简要论述设计思想。随后介绍了为方便语义网自动生成而对抽取信息源重构的过程，包含实体消歧、细粒度拆分等处理步骤，最后统计语义网构建的成果，并展示样例数据。

第5章 中草药自动问答系统的关键技术

中草药自动问答系统属于受限领域问答系统，基于底层 RDF 语义网，问题解析是其最重要环节，即从自然问句中抽取足够多的特征信息，生成 SPARQL 查询语言。相比而言，信息检索过程主要依靠生成的问句进行子图匹配，而答案抽取则由于其底层基于专业的知识库，不存在候选文档集，即使有多个答案置信度也较高，因而本文主要研究问题解析的关键技术。

5.1 问答策略

中文自然语言的表述形式多种多样，有符合词法、语法结构的问题，也有不规整的单词片段，但如果人工识别，大致都能理解问题的实际意图。所以，需要对可能出现的问题进行分析，根据其特点采取不同的问答策略。

5.1.1 问题分析

通过对中草药领域实际的问题的调研，我们将问题定义为如下两类：

精心构造的问题 Q_1 。这类问题的特点是明确包含了实体及其属性或关系，且符合一定的语法规则。例如问句“麻黄的功效是什么”包含实体“麻黄”及其属性“功效”；又如问题“什么药可以治疗糖尿病”包含实体“治疗”及其关系“治疗”。这类问题只需能找到这些特征词并且结合一定的领域知识，很容易自动生成对应的 SPARQL 查询语句。

社区实际的问题 Q_2 。这类问题的特点是往往实体、属性和关系只包含其一甚至没有，或者语法句法杂乱无章。例如问句“吃了麻黄后怎么一直感觉不舒服”只包含实体“麻黄”却能推测出其属性“副作用”；例如问句“晚上睡觉出冷汗，是什么原因啊”，只包含属性“原因”却能推测出其概念“疾病”；又如问句“手冷，浑身没力气”单单描述某种症状，其意图既可判断为某“疾病”含有此症状，但同时亦可判断为某“方剂”治疗某症状。对于上述问题，很难仅依靠问题自身信息以及领域知识自动生成问题，需要引入额外的信息，扩展问题。有两种方式可以实现上述问题的回答：通过模式匹配把问题对应到 SPARQL 语句，特点是回

答准确但推广性差，需要随时添加模板；通过分析标注的语料，从中统计出常见的 SPARQL 语句，并以机器学习的方式训练模型用于分类问题，特点是推广性好但准确性差，而且训练出模型的好坏十分依赖于标注的语料以及训练的方式。

5.1.2 查询三元组模式

SPARQL 的查询，是一个子图匹配的过程，条件就是三元组模式。我们定义实体为s，属性为p，属性值为o，关系为r，未知变量为？，则有以下四种最基本的查询方式，可以对应到 Q_1 中某个问题。如表 5.1 所示。

表 5.1 基本查询三元组模式

基本查询三元组模式	例句
$(s, p, ?)$	〈麻黄〉的〈功效〉是什么？
$(s, r, ?)$	〈一见消〉的〈组成〉是什么？
$(?, r, s)$	哪些方剂〈包含〉〈麻黄〉？
$(?, p, v)$	哪些药的〈产地〉为〈北京〉？

其他的 Q_1 类的较复杂的问句，无非是这些基本模式的交、并、非运算，并且绝大多数是交运算。如表 5.2 所示。

表 5.2 复合查询三元组模式

复合查询三元组模式	例句
$(?, r, s1)(?, r, s2)$	什么药可以〈治疗〉〈感冒〉与〈咳嗽〉
$(?, r1, s1)(?, r2, s2)$	〈治疗〉〈咳嗽〉且〈包含〉〈麻黄〉的药
$(?, p1, v1)(?, p2, v2)$	〈产地〉〈北京〉且〈性味〉〈苦〉的药

从理论上而言， Q_1 类问题只需正确抽取出关键词构成基本查询三元组，再组合成复合三元组模式，便能生成对应的 SPARQL 查询语句。而 Q_2 类问题之所以难

以回答是因为连最基本的简单查询三元组模式都无法构成。

5.1.3 查询生成策略

通过前文的分析，本文提出了一种层次过滤式的 SPARQL 查询生成策略，所谓层次过滤，即对于一个问题依次尝试各种生成方式，直到能正确生成 SPARQL 为止，如图 5.1 所示。

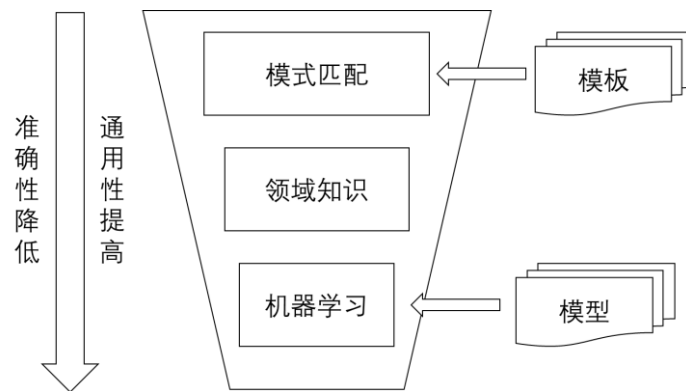


图 5.1 层次过滤式生成策略

生成策略包含三种，分别为基于模式匹配的查询生成、基于领域知识的查询生成和基于机器学习的查询生成。

基于模式匹配的查询生成，优先级最高，是最先调用的策略。其优点是准确匹配的问句能返回用户最想要的答案，并且随着模板添加可以回答无限多的问题。其面向的问题类型主要为 Q_2 ，其次为 Q_1 。

基于领域知识的查询生成，优先级其次。其优点是结合领域知识以及自然语言处理技术，能自动生成理论上无限多种的 SPARQL 问句，并且能保证相当高的准确率。其面向的问题类型为 Q_1 。

基于机器学习的查询生成，优先级最后。其优点是能回答分类正确的问题，并且扩展性较好，缺点是生成 SPARQL 语句需要预先定义，并且回答新类型的问题时需要重新训练模型，并且要有一定量的标注语料。其面向的问题类型为 Q_2 以及小部分 Q_1 。

5.2 预处理相关

由于中草药问答的领域专业性以及相关语料的缺失性，在进行相关技术研究前需要构建领域词典，并且也需要对中文语料进行标注。

5.2.1 领域词典构建

领域词典包含了中草药领域的一些专有名词，主要用于分词和命名实体识别，其内容从重构后的数据库中抽取。

1) 领域实体词典

从重构后的数据库中抽取每张基本表的名称以及拆分后的别名表，构建实体词典。由于原数据在获取以及拆分过程中，存在一定的错误，故在抽取过程中对内容进行的过滤。首先将长度为 1 或者大于 10 的词舍去，以数字或字母开头的词舍去，包含括号、逗号、空格等标号的词舍去，然后人工粗略过滤，最后获得词条 117903 项。

2) 功效词典

功效多为二字、三字、四字短语，其中四字短语可拆为一对二字短语。功效的描述首字多为动词且带某种趋势的词如“降”、“清”、“除”、“健”等，后接多为名词。功效词典从重构后数据库的拆分功效表抽取，其拆分结果中存在好多干扰项，按以下方式进行处理。首先长度为 1 的或者大于 4 的词舍去，包含数字、字母、符号的词舍去，并且在此过程中把四字词拆一对二字词，然后对每个词的首字进行统计，首字出现次数超过 5 词，则保留该词，否则舍弃，最后获得词条 1314 项。

3) 主治词典

主治多为病症名称，且在重构数据库中拆分结果较好，故只对长度为 1 以及含有标点的词做过滤，最后获得词条 6502 项

最后合并三个词典，获得分词词条 124799 项，词条样例如图 5.2 所示。

1	一丈红	100000	蠲痹丸	1000	万应保赤散	10000	先天性梅毒
2	一上散	100001	蠲痹四物汤	1001	万应内伤膏	10001	先天性气管畸形
3	一两金	100002	蠲痹散	1002	万应剪金丸	10002	先天性泪囊炎
4	一个刺二个头	100003	蠲痹汤	1003	万应剪金丹	10003	先天性淋巴水肿
5	一串红	100004	蠲痹消毒散	1004	万应化毒膏	10004	先天性溶血性黄疸
6	一串钮子	100005	蠲痹消毒饮	1005	万应午时茶	10005	先天性生精不能症
7	一串花	100006	蠲痹消痛液	1006	万应吹喉散	10006	先天性痴呆
8	一串钱子	100007	蠲痹秦艽汤	1007	万应喉中散	10007	先天性白内障
9	一串鱼	100008	蠲痹解毒汤	1008	万应喉症散	10008	先天性皮肤缺陷
10	一丸春丹	100009	蠲痹防痙汤	1009	万应回生膏	10009	先天性皮脂腺增生
11	一丸金	100010	蠲痹饮	1010	万应太乙膏	10010	先天性眼睑异常
12	一丸散	100011	蠲痹饮子	1011	万应太平丹	10011	先天性短颈
13	一丸散	100012	蠲痹煎	1012	万应头风膏	10012	先天性耳前瘻管
14	一丸金丹	100013	蠲痹散	1013	万应夺命散	10013	先天性耳胆脂瘤
15	一井散	100014	蠲痹散	1014	万应如意痢疾丸	10014	先天性聋
16	一井金	100015	蠲痹散	1015	万应宝珍膏	10015	先天性肌强直
17	一井金散	100016	蠲痹万灵汤	1016	万应山楂丸	10016	先天性肌强直病
18	一仙丹	100017	蠲痹丸	1017	万应延龄丹	10017	先天性肌肉缺如
19	一代宗	100018	蠲痹枳实丸	1018	万应愈风酒	10018	先天性肛门直肠畸形
20	一元丹	100019	蠲痹枳术丸	1019	万应抵金散	10019	先天性肢体环状狭窄
		100020	蠲痹散	1020	万应救急熊胆神丹	10020	先天性肢体缺如

图 5.2 领域词典词条样例

5.2.2 问题语料

相比于一些英文的问题集，中草药领域的问题集很少甚至于没有，所以需要构造自己的问题集，包括通过网络爬虫抓取问题并且按照一定准则进行语料标注。

5.2.2.1 语料获取

通过网络爬虫，抓取了“寻医问药”网站“有问必答”功能板块下分类为中医骨伤科、中医儿科、中医保健、中医妇科、中医综合、用药常识与不良反应、中草药问题的问题，最后获得问题 4 万多条。

5.2.2.2 语料标注

通过粗来浏览原始问题发现，语料主要涵盖了单味药、方剂以及疾病三个领域的问题。由于在证候、化合物、基源相关的问题表述等更加专业，而网上的表达比较通俗化，所以这部分问题缺失严重。同时，语料中存在许多并非符合词法、句法规则的问句如“补肾，耳鸣，强身，壮阳，提高免疫力”、“脾胃湿热引进肚子胀怎么吃就好中药都见效”、“，掉头发吃六味地黄丸可以不”等。鉴于系统作为一个开放式问答平台的辅助接口，其所需处理的问题是以交流形式出现的问题，形式绝对不会是规整的，所以在标注的过程中，不对原始问句进行改写。

对每一个问句，由两人进行三个类别的标注，当两人的标注存在出入时，交由第三人裁定。问题标注的三类分别为是否有效、答案属于哪个领域以及确定领域的那个属性或关系。

1) 问题是否有效

是否有效指目前语义网的知识的领域是否涵盖问题想要获取的信息，若没有涵盖则直接作为无效问题过滤。在标注过程中尽可能与已有的知识库产生关联，例如对于“冬虫夏草、藏红花、雪莲可以一起泡酒吗”，问题的意图可能为这些药是否相互影响，但目前的语义网未涵盖此方面知识，故转化为每一单味药的禁忌。样例标注如表 5.3 所示。

表 5.3 问题是否有效标注样例

问句	标注	分析
苍术有什么功效药量吃多呢	有效	药的功效
问下你医生，肾阴虚可以吃金匱肾气丸吗	有效	病的治疗、方的功效
肾阴虚和阳虚和肾精亏虚，怎么确认	有效	病的症状
芦荟有什么营养价值？……	无效	营养，领域无关
金银花露，下火王都吃过，但起不到一点作用	无效	意图不明确
什么牌子的芦荟软膏比较好？有图片麼？	无效	品牌，领域无关

2) 问题领域

问题领域针对标注为有效的问句，是指询问的问题可以归为单味药、方剂、疾病、证候、基源、化合物中的哪一个大类。但是语料中也存在“问下你医生，肾阴虚可以吃金匱肾气丸吗”这类表判断的问句，直接返回“是否”则缺乏可信度，并且在已知方的情况下询问“方病”关系，则人为定义此问题为“方剂”领域。又如问题“怎么回事？医生是上火了吗？应该怎么治疗”既可以归类为询问含有“上火”的疾病实体，然后返回该疾病的治疗属性，也可以归为治疗含有症

状“上火”疾病实体的方剂实体，甚至还可以归为功效中含有“清火”的方剂实体，则也需要人为定义此问题为“方剂”领域。样例标注如表 5.4 所示。

表 5.4 问题实体领域标注样例

问题	标注	分析
苍术有什么功效药量吃多呢	单味药	功效
您好，冬虫夏草如何服用效果更好	单味药	用法
枸杞和阿胶可以一起吃吗	单味药	禁忌
问下你医生，肾阴虚可以吃金匱肾气丸吗	方剂	治疗
吃什么东西可以健脾胃吃什么东西可以健脾胃。	方剂	功效
吃金匱肾气丸发热，还能吃吗	方剂	副作用
肾阴虚和阳虚和肾精亏虚，怎么确认	疾病	症状
夜间夜间盗汗是什么原因	疾病	病因

问题属性问题属性即的真实意图，即在某个领域内的具体某项属性，如求方剂的功效、主治、副作用等，不同领域的属性基本不相同，所以需要针对不同领域做不同分析。

5.2.2.3 标注结果

本文标注了 800 条问句，是否有效、问题领域、方剂问题属性如表 5.5 所示。标注结果与日常经验相符，即社区的绝大多数实际问题都是通过描述症状寻求治疗方法，其次是药的服用禁忌等问题。

表 5.5 标注结果统计

标注类型	具体分类	数量
是否有效	有效	464
	无效	336
问题领域	药	109
	方	277
	病	78
问题属性（方剂）	治疗	108
	主治	76
	副作用	71
	其他	22

5.2.3 多分词校正

在中文自然语言处理领域，分词是一个及其重要的过程，尤其是当需要进行更深入的词性标注、句法树解析的时候，第一步的分词准确率及其影响之后的处理。经过测试发现，即使引入了自定义词典，使用单一中文自然语言处理工具很难达到理想的效果。如对于问句“什么药可以治疗糖尿病”，HanLP 结果为“什么药/nz, 可以/v, 治疗/vn, 糖尿病/nhd”，其准确分出了疾病实体“糖尿病”，但是对于疑问代词“什么”却分成了“什么药”，同样使用 FNLDP 的结果为“什么/dt 药/nn 可以/av 治疗/vv 糖尿/nn 病/nn”，疑问代词“什么”分词正确，但是疾病实体“糖尿病”却分成了“糖尿”和“病”，两者皆没有达到理想的分词效果。

但经过实际测试发现，把 HanLP 分词后结果以空格分割拼接后如“什么药可以 治疗 糖尿病”作为 FNLDP 的输入，可以达到目标分词结果“什么 药 可以治疗 糖尿病”。故为了后续处理的准确性，本文采取此种分词策略，对问句的分词结果如图 5.3 所示。

```

1 海龙多少钱一条，还有和何首乌多少钱一克
2 请问当归在市场上多少钱一克
3 请问喝中药后口苦适合吃些什么
4 健脾利尿清热除湿化痰的食疗与中成药
5 三七粉都有什么吃法如何服用效果最好三七粉的吃法
6 夏天中医养生保健的方法及注意事项
7 主要症状：发病时间：化验检查结果：详细回答
8 脾虚吃什么食物能补呢
9 最近用五味子泡水可以吗
10 主要症状：2 发病时间：化验检查结果：
11 问题描述：鼻炎
12 胃癌 2009 年 9 月份作为参考，手机号码可以不填吗
13 肾阳虚和肾阴虚都有怎么治吃什么药
14 舌苔厚，舌苔红和流鼻涕一星期食疗什么好
15 中药没效果请问吃什么食物补脾胃
16 脾胃虚弱的人一定伴随脾胃疼吗？
17 化验检查结果：头孢吡肟有什么作用，还有什么副作用
18 脾胃不好应该怎样调养，应该注意点什么呢？
19 脾胃不好的人怎样调理？常常大便不成形
20 四个多月怎能引起或怎没治疗此病
21 都没效果，不知道应该怎么做才可以没那么小，长大一点
22 请问挖鼻子是上火了吗
23 六味地黄丸宜饭前还是饭后还是睡前服用
24 手脚冰凉怕冷怎么回事有什么方法可以改善
25 请问体内湿气重是什么意思，湿气是怎么来的。

```

图 5.3 问句分词样例

可见，分词过程中分出了绝大多数的领域名词，如例句 1 中的“海龙”、“何首乌”表单味药，例句 3 中的“健脾”、“清热”、“除湿”、“化痰”表功效，例句 13 中的“肾阳虚”、“肾阴虚”表疾病，例句 23 中的“六味地黄丸”表方剂。

5.2.4 领域词汇识别

领域词汇识别类似命名实体识别，其目的为发现文本中的专业词汇，并标注为类型为单味药、方剂、疾病、证候、化合物、基源、属性、关系中的某一类。但由于词汇存在一对多的关系，如“麻黄”可以识别为单味药，也可以识别为“基源”，为解决此问题，引入一种单键多值的方案，即对于一个实体名，存储其对的所有对应的 URI。若其对应的 URI 都属于同一个命名空间，则此实体判断为此命名空间的实体，否则标注为多类型，需要消歧。

5.3 特征设计

通过特征设计，抽取问题中的特征并用于分析计算，有着十分重要的意义。尤其对于问题这类短文本，经过分词、词性标注和语法树解析后，往往还需要借助一定的领域知识，通过特征组合的方式，提升分类结果。

5.3.1 Doc2Vec

词向量是一种自动提取特征的方式，把文档中的每个单词映射到一个对应的低维向量空间，方便进行余弦距离等计算。虽然其结果向量不可解释，但是依旧保留着一些语义特征。从 Word2vec 得到启发得到的 Doc2Vec，即把语句映射到一个向量空间。我们对抓取的 4 万多问句预先处理，得到了每一个问句对应的 200 维词向量，结果如图 5.4 所示。

```

1 sent_1 *** -0.057967775 -3.8123335E-4 0.022350091 -0.03814034 -0.011499845 -0.009933123 -8.
2 sent_2 *** -0.117549226 0.16592513 -0.13435535 -0.03375954 -0.0028888732 -0.066577844 0.096
3 sent_3 *** -0.06723517 0.17249823 0.16476594 -0.078306936 0.013125501 -0.09095488 -0.178378
4 sent_4 *** -0.103776604 0.041240633 0.19898513 -0.019592628 0.19530667 -0.07441089 -0.04606
5 sent_5 *** -0.013409544 9.591572E-4 0.06219206 0.007189823 -0.0138372285 -0.01370563 0.1127
6 sent_6 *** -0.22613204 0.12904188 -0.17510472 -0.010265619 -0.14644304 0.13299824 0.1811178
7 sent_7 *** -0.23530626 0.13266474 0.08319731 -0.18791358 -0.12393655 -0.08378816 -0.1076140
8 sent_8 *** -0.09667173 -0.00789469 -0.105436765 0.1683318 -0.09302202 -0.072156616 0.256260
9 sent_9 *** -0.32270044 0.011475794 -0.23533155 -0.090370975 -0.14421503 -0.21941218 0.02521
10 sent_10 *** -0.006732115 -0.056715474 2.0293311E-4 -0.052984625 -0.028931707 0.005464193 -0
11 sent_11 *** -2.6590226E-4 0.01035928 -0.04993593 0.030513596 -0.039010625 -0.009325957 0.01
12 sent_12 *** 0.1662179 -0.03670427 0.11213178 0.0791559 0.018753177 0.034056112 -0.041561615
13 sent_13 *** 0.032978773 0.026390314 -0.09552418 -0.004640639 0.0014695564 0.0445296 0.14121
14 sent_14 *** 0.041886102 0.02818261 -0.015620496 -0.030593634 -0.08793217 -0.03953337 -0.027
15 sent_15 *** 0.010353242 0.020575477 0.020876197 -0.02387086 0.029000118 -0.04089908 -0.0219
16 sent_16 *** -0.06165271 -0.12596668 -0.11096219 0.06497612 -0.13672744 0.04216621 0.1229882
17 sent_17 *** 0.007945407 0.019168133 0.017520854 0.055138946 -0.040489238 -0.002688479 -0.02
18 sent_18 *** -0.03855395 0.03867257 0.04128677 -0.017460078 -0.05917953 -0.01223285 0.028084
19 sent_19 *** -0.096319154 0.04019699 -0.03020063 -0.0041350587 -0.03762579 -0.031802714 0.02
20 sent_20 *** -0.07628074 0.106515504 -0.17340446 -0.15135233 -0.18067044 -0.05462349 -0.0562
21 sent_21 *** 0.019362919 -0.021282207 0.024252092 -0.08630113 0.0548689 -0.018703857 -0.0746
22 sent_22 *** -0.027152514 0.010345677 0.03245519 0.041508235 0.057465967 -0.040571935 0.0594
23 sent_23 *** -0.10001623 -0.010079033 0.041699987 0.028502245 0.08041919 -0.0533181 -0.23853
24 sent_24 *** 0.027895318 -0.03564397 -0.047317743 0.05947377 0.07392372 -0.12916774 0.036282
25 sent_25 *** -0.07855714 0.027391324 -0.26037258 0.025423318 0.022754908 0.022313597 0.06157

```

图 5.4 问题词向量表示

对于新输入问句向量的构造表示，可以通过词向量相加构造，而词向量则可以由百度百科或者问题语料训练的 word2vec 获取。

5.3.2 词袋模型

词袋 (Bag of Words, BoW) 模型对自然语言进行了简化，它把一个文本仅当做是一个词的集合，并且假定每个词的出现都是独立的。词袋模型忽略词序、语法和句法。其结果最终也数值化到一个与词表长度相同的向量空间，有三种形式：One-hot、词计数以及 TF-IDF。本文主要用到了后两种。对 4 万多问题分词，并对词频大于 10 的单词建立词表，得词表长度为 3052，即向量空间为 3052 维。

5.3.3 疑问词

在英文环境中，疑问代词如 what、where、when 的作用是十分重要的，直接表明了用户查询意图。但中文领域，疑问词虽然也包含着一定的信息，但是也有

问题的表述根本不包含疑问词。疑问词的提取可以通过构造疑问词词典获得，但是由于汉语词性的多变性，根据词典匹配的结果并不准确,而且一个问题基本上只有一个或没有疑问词，表 5.6 所示为疑问词的提取算法。

表 5.6 疑问词提取算法

算法：疑问词提取算法
输入：疑问句 Q
输出：疑问词集合 S
STEP1. 从百度百科中获取疑问词的集合，构造原始疑问词词典 D
STEP2. 对给定的疑问句 Q 进行分词和词性标注，标注后获得结果
$Q_p = \{w_1 / pos_1, w_2 / pos_2, w_3 / pos_3, \dots, w_n / pos_n\}$
STEP3. 从 Q_p 中提取疑问词词典 D 中的词，构成初始疑问词集合 P0
$P_0 = \{w_i w_i \in D \text{ and } w_i / pos_i \in Q_p\}$
STEP4. 对 P0 的词进行过滤，删除词性为 a、d、m 的词获得疑问词集合集合 P1
$P_1 = P_0 - \{w_i w_i \in D \text{ and } w_i / pos_i \in Q_p \text{ and } (pos_i = "a" \text{ or } pos_i = "d" \text{ or } pos_i = "m") \}$

5.3.4 核心关键词

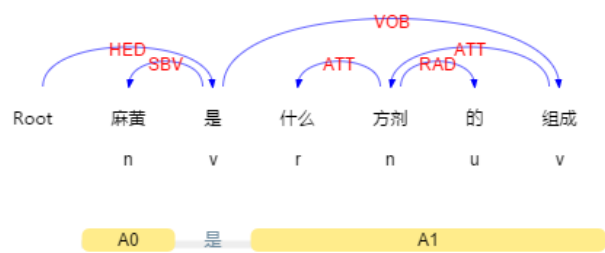


图 5.5 句法依存树

定义核心关键词为疑问词的一、二级依存词和问句中心语。其特征计算的过程主要依赖语法生成树。疑问词的一级依存词指与疑问词有依存关系的词，疑问词的二级依存词指与一级依存词有依存关系的词。当遇到一些无效词如助词“的”、“地”为依存词时，需要再向父节点推进。问句中心语则是指在依存句法分析父

节点为“HED”的词。

例如对于问句“麻黄是什么方剂的组成”，其依存句法树如图 5.5 句法依存树所示。其关键词为“什么”，一、二级依存词为“方剂”和“组成”，问句中心语为“是”。但是实际处理中由于中文复杂句语法树解析的错误率极高，复杂问题很难获得正确的核心关键词。

5.3.5 主谓宾特征

一个符合语法规则的句子，基本是由主语、谓语、宾语三者构成，其中主语是施事者，宾语是受事者，而谓语连接主谓，是句子中的动词及其辅助部分，某种程度上主谓宾承载着句子的核心信息，并且兼顾句子的语法及语义信息。谓语是父类为 HED 的词，而主宾分别为该词的左右孩子节点。

以上面的例子为例，其主谓宾特征分别为“麻黄”、“是”、“组成”。

5.3.6 领域知识特征

领域知识特征，是与中草药领域相关的特征，其抽取过程依赖于启发式规则以及词性标注。例如以疑问词加领域概念出现如“什么药”，“什么病”的问句，又如以疑问词加量词加领域概念出现如“哪个病”、“哪味药”的问句，再如以助词加领域概念出现如“的药”、“的病”的问句，其提问意图的领域分类基本明确。

同时结合领域词典，对问句中实体、属性、关系的类型以及出现次数的做统计，也可以作为领域知识特征。

5.3.7 限制特征对

限制特征对用于属性、值对的提取，用于限制问题的搜索范围。其一般由属性后面第一或第二个词构成。如问句“性味苦的治疗咳嗽的单味药”中“性味”、“苦”即为限制特征对。

5.4 基于模式匹配的方式

基于模式匹配的方式针对包含实体 Q_1 和 Q_2 类问题。通过问题和模板的映射，从问题中抽取出模板中空缺的实体，填充到 SPARQL 查询语句中，进行知识库的

查询。

5.4.1 问题类型

分析 Q_1 类问题“麻黄和防风有哪些相同的功效？”中存在两个查询三元组模式“(麻黄, 功效, ?)”以及“(防风, 功效, ?)”，这也是在基于领域知识的方式中所生成的两个查询语句，但此问题中的关键词“相同”指出所求是结果的交运算，故需设计模板在此步骤拦截此问题并生成正确的 SPARQL。又如 Q_2 类问题“为什么吃了麻黄不舒服”的意图是副作用，基于领域知识的方式由于查询三元组模式不完整，不能生成对应的 SPARQL，便会尝试使用基于机器学习的方式，但若未有相关训练模型，则将不能回答此问题，使用模式匹配的方式可解决以上问题。

5.4.2 模板设计

模式匹配的方式可分为两种，强模式和弱模式，各自拥有优缺点。对于问题“麻黄和防风有哪些相同的功效？”，对应的 SPARQL 查询语句及各自模板如表 5.7 所示。

表 5.7 模式匹配对比

类型	模板
强模式	<A>和有哪些相同的功效？
弱模式	<A> 和 相同 功效？
对应 SPARQL 语句	
SELECT ?gj WHERE {<A> 功效 ?gj. 功效 ?gj}	

强模式无需任何预处理，仅通过问题与模板逐字匹配的方式完成模板变量的填充，但是此方式仅能回答一种类型的问题。

弱模式则需要进行分类处理，同时也要依靠领域词汇识别。此方式通过问题与模板的关键词依序匹配的方式完成变量的填充，此方式扩展性较好，但极度依

赖分词结果和自定义模板，有时还会出现错误匹配。

5.4.3 自定义模板及匹配算法

下面介绍一种模板定义及其对应的匹配算法。

模板的简化形式如表 5.8 所示：

表 5.8 自定义模板

问题模板
2 // 代表只与识别有 2 个实体的问句匹配
AX BX // 模板中该替换的变量
AX BX 相同 功效 // 问题模板
x // 所求变量
SELECT ?x WHERE {AX 功效 ?x. BX 功效 ?x.} // SPARQL 查询语句

匹配成功的依据为为模板与问题之间词汇的一一映射所产生的连线不产生交叉。算法如表 5.9 所示。

表 5.9 模板匹配算法

模板匹配算法
STEP1. 定义一个长度为模板词数的数组 index,用于存储模板词问题中的位置， 定义实体变量的初始序号 start 为 1
STEP2. 从模板中取一个词，若该词为替换变量，则从实体列表选取第 start 个 的实体的词序，作为 index 的值，start 自增 1；若该词不为替换变量， 则从问句中查找该词词序，作为 index 的值；若不存在，则不能匹配
STEP3. 若词未匹配完，则返回 STEP3.
STEP4. 判断 index 时候递增，若不递增则不能匹配
STEP5. 替换 SPARQL 中对应的变量，匹配结束

5.4.4 方法评价

模式匹配方式成功的关键是实体的识别。问句中无实体不能回答，如“睡不着怎么办”；实体的个数过多不能回答，如“麻黄、三七、防风相同的功效是什么”；拦截其他正常问题，如“龙虎散和六味地黄丸相同的功效”，原因是“功效”的命名空间不同，明显产生答案。

5.5 基于领域知识的方式

基于领域知识的方式面向 Q_1 类问题。通过引入领域知识，并且组合查询三元组模式，可以无限多的 SPARQL 查询语句。

5.5.1 问题类型

表 5.10 SPARQL 生成

问句	SPARQL
麻黄和防风的归经以及功效	1. (麻黄 归经 ?) 2. (麻黄 功效 ?) 3. (防风 归经 ?) 4. (防风 功效 ?)
决明子和大黄以及萱草根中含有什么化合物	(决明子 包含 ?) (大黄 包含 ?) (萱草根 包含 ?)

Q_1 类问题，即精心构造的问题，一般只有一个意图。简单的问句形式，单领域的如“麻黄的功效是什么”，跨领域的如“什么药可以治疗咳嗽”。复杂的问句形式有多实体、多个属性问句如“麻黄和防风的归经以及功效”，多个跨领域问句如“决明子和大黄以及萱草根中含有什么化合物”。对于复杂问句，存在一定的规律。第一个问句，可以拆分为四个简单查询问句，而第二个问句，可以拆分为包含三个查询三元组模式交的查询问句。如以上两个问句的 SPARQL 形式如表 5.10 所示。

5.5.2 领域分类器

由于语义网设计时采用了不同的命名空间，直接生成问句有一定的困难，所以需要预先给问题分类，再自动生成对应的 SPARQL 问句。这里我们主要采用之前提取的领域知识特征，构建一个启发式规则的领域分类器。

分类器主要应用两类特征，我们定义如“什么病”、“哪种化合物”、“的药”之类的短语为强特征，即出现此类短语则可以立即确定回答的领域，如上分别为疾病、化合物以及方剂。我们定义与实体关联 URI 为弱特征，在强特征不能确定分类的情况下，计算 URI 类型。由于在 URI 中编码了分类信息，所以根据类型可以分为基本的六类以及多类共 7 类。

5.5.3 问句生成

在问句生成过程中，对不同的实体数目采取不同的策略。当问句中只存在一个实体时，对该实体对应的每一个 URI 都生成 SPARQL 查询语句；当问句中存在多个实体时，首先尝试将这些 URI 归一到统一一个领域，并对这个领域内的 URI 生成一个 SPARQL 查询问句。如对于问句“决明子和大黄以及萱草根中含有什么化合物”可以根据其关键短语“什么化合物”确定领域为化合物，又由于实体属于“单味药”，之间只有“包含”关系。

若问句拥有限制对特征，如对于问句“性味苦的药”的查询语句中添加“(性味, 苦)”，但是在问题中所提取的词是否准确十分影响查询结果的好坏。

5.5.4 方法评价

此方式可以回答绝大多数的 Q_1 类问题，即使如“白术、人参、生姜、甘草、大枣、益智仁在哪些方剂中含有”、“什么药可以治疗失眠、腹泻、昏迷、白喉”这类问题都能回答，但是对于与“什么病的症状有出汗”、“什么药微苦”这类问题却不能很好回答，在引入限制对特征的情况下，或许一些问题能回答得更加准确，但也会造成另一些问题由于限制对特征的值错误而得不到答案。所以对于这类问题，将尝试采取下一种方式解决。

5.6 基于机器学习的方式

基于领域知识的方式面向 Q_1 类以及小部分 Q_2 问题，这类问题很难根据其特征直接生成 SPARQL 查询语句，但是由于在网络上拥有许多类似的问题，故可以分析这些问题，挖掘出问题到 SPARQL 的对应关系。该方式主要通过训练分类模型实现。

5.6.1 有效问题分类

该分类为一个二分类问题，的目的是过滤无效问题，即那些表述不明、或者不相关的问题。此分类器的优先级高于基于模式匹配、基于领域知识的方式，目的是直接剔除无效问题。分类优劣评判的标准是尽量降低第一类错误（拒绝了实际上成立的），而第二类错误（不拒绝实际上不成立的）则没有要求，原因是 SPARQL 生成存在三套策略，所以即使在开始把负类归为正类，最后产生的语句也不一定能检索到答案。

5.6.1.1 实验设计

该实验采用 464 正例、336 负例，使用 Weka 的进行 10 折训练。

1) Doc2Vec

使用 Doc2Vec 为特征是由于其生成特征的过程简单，无需预先定义语料的向量空间，也无需对每个问句进行向量计算，同时 Doc2Vec 结果的向量空间维度较低。

以 200 维向量作为输入，采用多种分类器进行比较实验，结果如表 5.11、图 5.6 所示。

表 5.11 有效问句分类 doc2ve 特征分类结果对比

Classifier	NaiveBayes	J48	Logistic	RandomForest	SVM
Precision	0.672	0.608	0.663	0.718	0.711
F-Measure	0.673	0.609	0.664	0.715	0.699
负类 FP	0.265	0.315	0.259	0.179	0.144

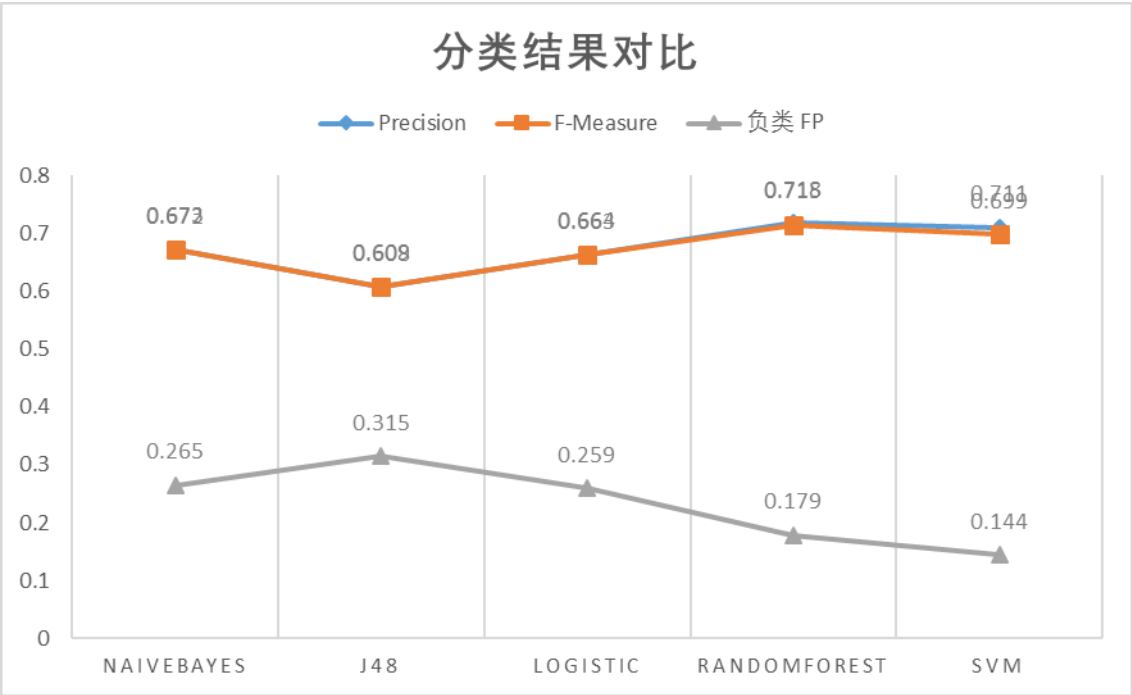


图 5.6 doc2ve 特征分类结果对比

2) 词袋模型

词袋模型是最常用的特征，一般都能获得不错的效果。

以 3052 维的词数向量作为输入,用多种分类器进行比较实验,结果如表 5.12、图 5.7 下所示。

表 5.12 有效问句分类词数特征分类结果对比

Classifier	NaiveBayes	J48	Logistic	RandomForest	SVM
Precision	0.687	0.653	0.663	0.672	0.728
F-Measure	0.685	0.654	0.658	0.672	0.715
负类 FP	0.293	0.267	0.338	0.209	0.134

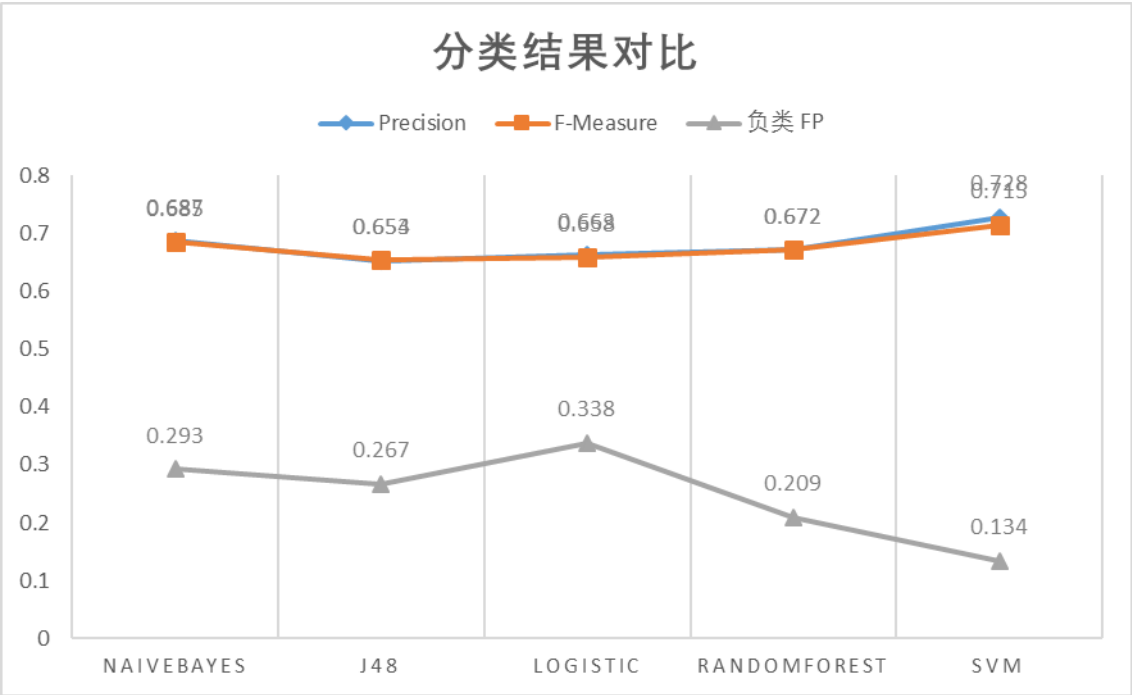


图 5.7 词数特征分类结果对比

以 3052 维的 TF-IDF 向量作为输入，用多种二分类分类器进行比较测试，结果如表 5.13、图 5.8 所示。

表 5.13 有效问句分类 TF-IDF 特征分类结果对比

Classifier	NaiveBayes	J48	Logistic	RandomForest	SVM
Precision	0.657	0.603	0.647	0.715	0.691
F-Measure	0.656	0.66	0.675	0.684	0.605
负类 FP	0.31	0.272	0.317	0.103	0.058

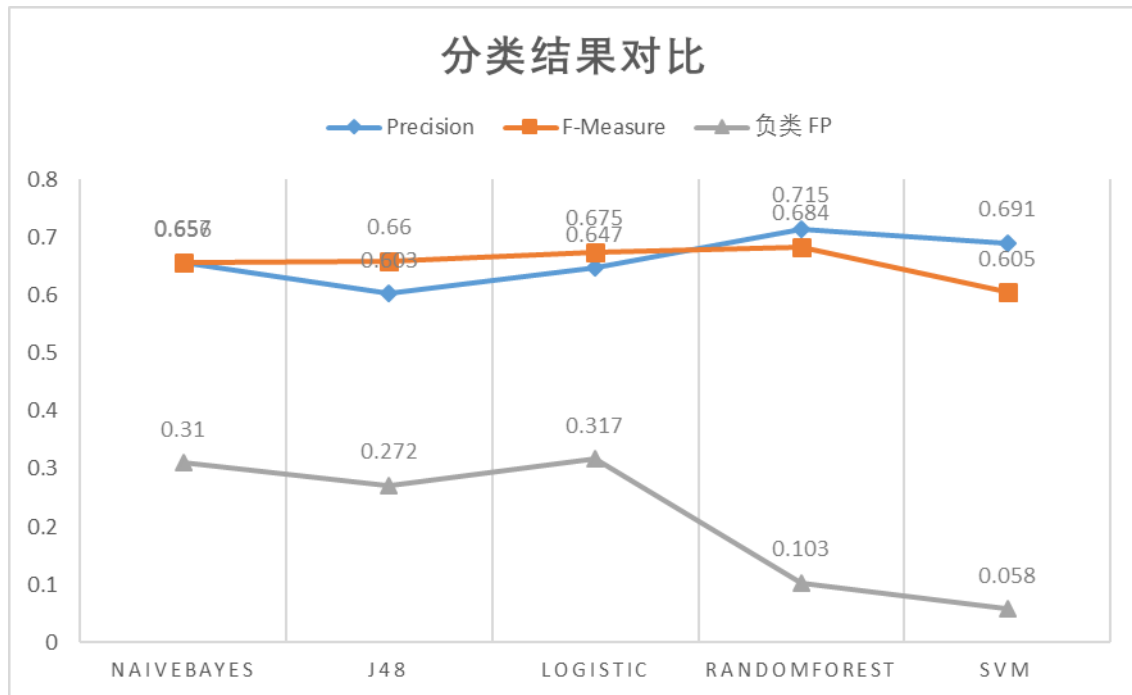


图 5.8 有效问句分类 TF-IDF 特征分类结果对比

5.6.1.2 结果分析

从以上不同特征在不同分类器下分类结果来看，RandomForest 与 SVM 的效果要明显优于 NaiveBayes、J48 和 Logistic。其中 SVM 和 RandomForest 的平均准确率相近，但是训练时间前者远远快于后者。在以 TF-IDF 为特征的情况下，SVM 的第二类错误率仅有 0.058，远小于其他值。但是也由于整个数据集分类结果向正类偏斜，造成了其 F-Measure 的下降。

我们随机抽取标注为正类却被预测为负类的问句，并进行分析。如问句“哪些中药可以口服治牛皮癣，哪些又可以泡澡治疗”标注为正类，认为其意图是获取治疗牛皮癣的方剂，但是由于在整个语料库中词“牛皮癣”出现的次数不足 10 词，故该词在向量空间中没有值，同时口服这个词常出现在中成药或西药领域，而分词过程出现的误差把“口服液”分成了“口服”和“液”，而含有其问句多数标注为负类，故此可能是为产生误分的一种原因。

我们人工构造问句“萱草根的归经是什么？”，此问题被分为负类，原因是“萱草根”和“归经”这些专业词汇不在预先定义的词表之中。此外，在领域相关的问题中，训练语料不完备，会使得训练出模型的扩展性较差。鉴于 SVM 在时间、

效果以及只确定支持向量而在小样本分类上的优势，接下来的实验都使用 SVM 分类器。

5.6.2 问题领域分类

问题领域分类是把问句分类到单味药、方剂、疾病三类。实际上还应该包括证候、基源、化合物三类，但由于语料的缺失，而我们认为后三类的问题术语精心构造类问题，可以通过前两种方式解决。只有对于形式复杂的问句，才会进入到第三种方式。

5.6.2.1 实验设计

该实验采用 464 正例、336 负例，使用 Weka 10 折训练。

1) 一般特征

采用前一个实验的三种特征进行对比实验，结果如表 5.14、图 5.9 所示。

表 5.14 单一特征对比实验

特征	Doc2Vec	词数	TF-IDF
Precision	0.705	0.77	0.752
F-Measure	0.705	0.769	0.751

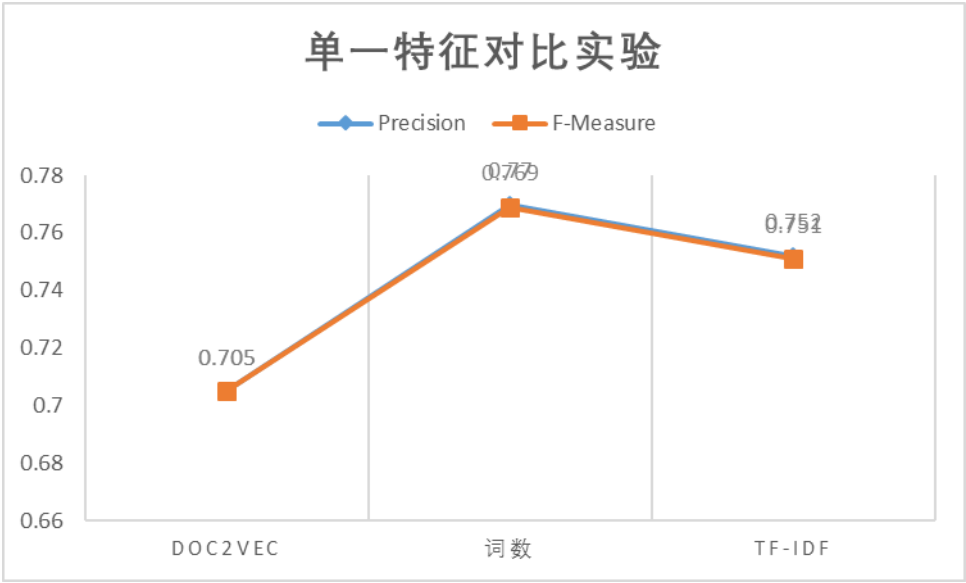


图 5.9 单一特征对比实验

通过实验发现，单单使用 BoW 特征，分类准确率就能达到 70%以上。以经验判断，是因为药、方、病类中的表述词区别明显，故能很好的表征问题。

2) 语法特征

通过前一实验，发现词数特征效果相对较好，且计算简单，则使用此特征与先前语法分析特征进行组合，并对特征进行适当加权，对比实验结果如表 5.15、图 5.10 所示。

表 5.15 组合特征对比实验

特征\权重（F1）	权值 1	权值 2	权值 5
BoW+疑问词	0.765	0.762	0.756
BoW+核心关键词	0.779	0.782	0.801
BoW+主谓宾特征	0.752	0.739	0.726

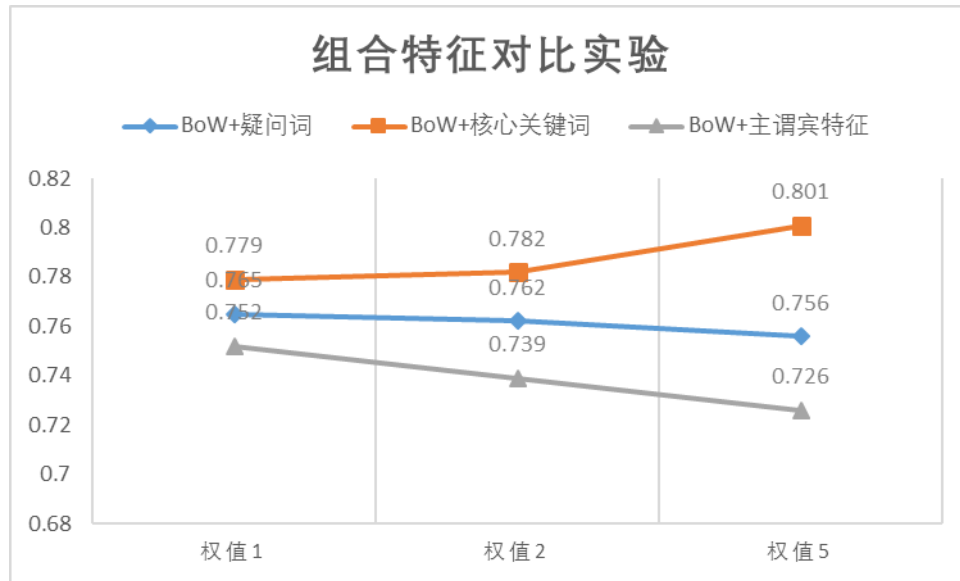


图 5.10 组合特征对比实验

从实验结果看以语法的角度去分类效果提升不大，但使用核心关键词使问题分类结果有显著提升。

3) 领域知识

从使用 HowNet 提供额外信息从而提高分类准确率受到启发，我们引入领域知识，设计如下实验：1.对问句中每一个出现在领域词典中的专业词权重增加一个数值，以强化该词的作用；2.只使用专业词；3.复用启发式规则的分类结果，直接进行分类；4.复用启发式规则的分类结果，把向量空间增加 4 个维度，分别代表药、方、病、其他，并且固定值为 10；5. 复用启发式规则的分类结果，并与专业词结合。如表 5.16、图 5.11 所示。

表 5.16 领域特征对比实验

特征\权重 (F1)	权值 1	权值 2	权值 5
BoW+专业词	0.834	0.845	0.820
专业词	0.675	/	/
启发式 (4 维)	0.692	/	/
启发式+专业词	0.801	/	/
启发式+BoW+专业词	0.754	/	/

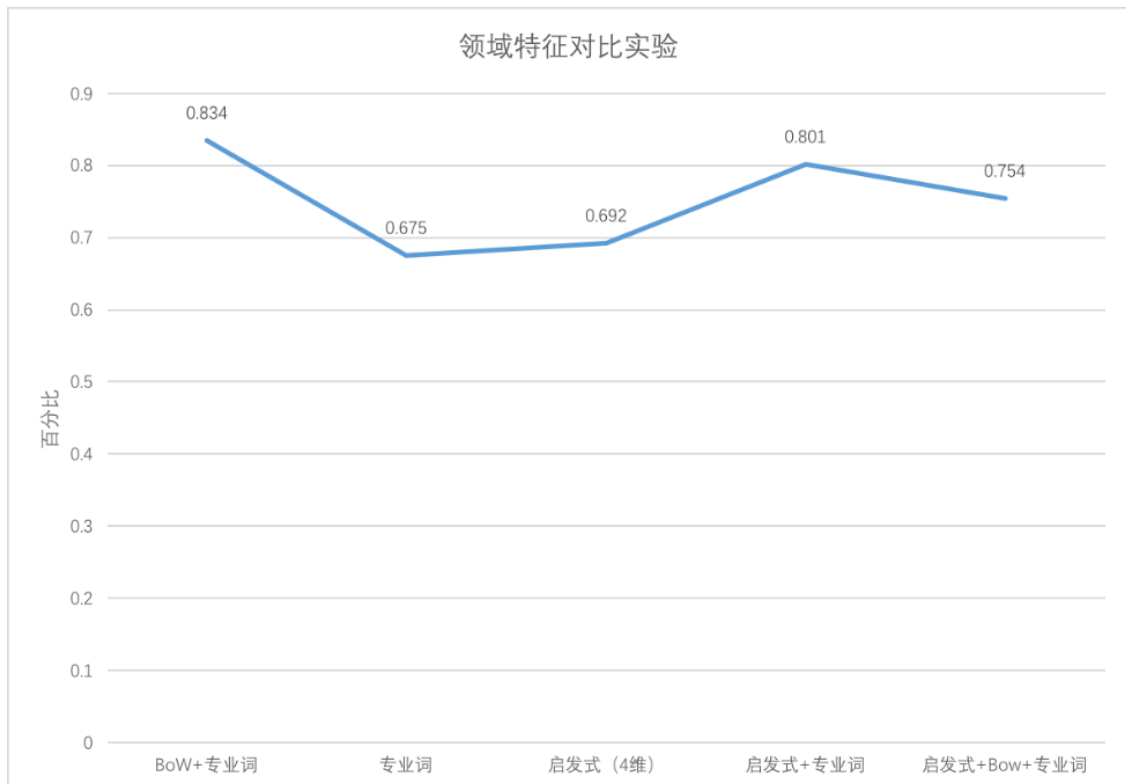


图 5.11 领域特征对比实验

引入领域特征的专业词之后，分类效果显著提升。

5.6.2.2 结果分析

由组合特征对比实验发现，即使我们加入了语法特征，依旧没有提升分类效果，比如疑问词并不能判定一个问题的类型，但是也可以看到，即使对于复杂问句的语法树解析结果完全错误，分类结果也有提升，甚至达到过 0.8。

而引入领域知识后，BoW 加专业词的方式稳定超过 0.8，而仅仅使用专业词却只有 0.67，从特征提取结果看，好多的句子都不包含专业词，使得其特征向量为零向量。当引入启发式规则后，单靠启发式规则效果不佳，原因是好多问句分类为 multi 和 none，即多实体或无实体，启发式加专业词的结果能到 0.8，但加上 BoW 后却开始下降。

从以上实验看出，引入领域特征能大大提高分类的结果，但是也不能仅使用领域特征，否则会造成一个问句的特征向量为零向量。

5.6.3 问题属性分类

问题属性分类即把问题最终归类为一个 SPARQL 查询集合，此后便可查询获取答案。不同的问题领域的问题属性也不同，其所对应的最佳特征可能也不同，下面主要对方剂训练集进行属性分类，分为治疗、功效以及副作用三大类。

5.6.3.1 实验设计

该实验采用 277 条标注为“方”的问题，其中 108 例治疗、76 例功效、71 例副作用，18 例用法（舍去）、4 例其他（舍去），使用 Weka 进行 10 折训练。

我们仅使用了专业词特征，即对属于功效和主治的词进行加权，实验结果达到了 0.85。

5.6.3.2 结果分析

由于我们的领域词典由有功效词典、主治词典、实体词典组成，而“治疗”属性中常有主治的内容，即病的症状，“功效”属性常含功效的内容，“副作用”属性常含方剂实体，所以单使用此特征就能很好的代表问句，但也依旧存在一些零向量。对于其他的领域，也可以通过添加专业词典的方式强突出关键词，如病的“症状”和“原因”的词典。虽然单个过程的分类结果较好，但是当结合领域分类器时，除了“副作用”类问题能较好分类，其他问题有存在一定的误分，归类为“疾病”类。

在分类确定后，将通过词典匹配方式提取问题中的关键词构成 SPARQL。对应 SPARQL 并非唯一，可能为一个集合。例如对于问题“一见消和一扫光可以一

起吃吗”对应为单一 SPARQL: (一见消 禁忌 ?x) (一扫光 禁忌 ?y), 又如“上火怎么办”, 将对应到如下 SPARQL: 1. (? 治疗 疾病) (疾病 症状 上火); 2. (? 主治 上火)。

5.6.4 方法评价

通过对特征的组合, 尤其是领域特征的引入, 基本上分类的 F-Measure 能达到 0.8。并且这些问题, 实际上也代表了网络上绝大多数的常问题。所以即使底层知识库的属性、关系很多, 但实际受人们关注的却只有一个很小的子库。在这种情况下, 生成 SPARQL 的关键是能发现最能代表问题的特征, 并能把问题分到正确的类别。

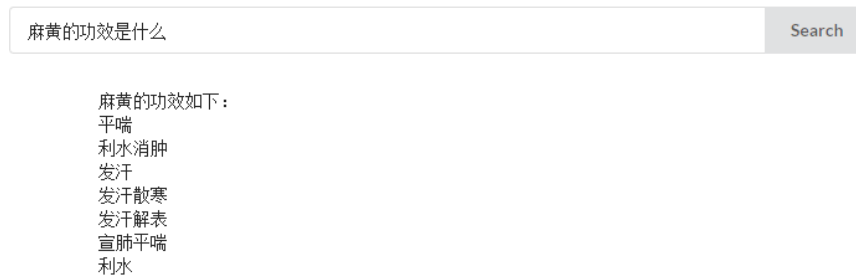
5.7 本章小结

本章首先分析潜在问句的特点, 并阐述从自然语言解析为 SPARQL 查询语言的过程。然后提出一种层次式过滤的 SPARQL 查询生成策略, 分别为基于模式匹配的查询生成、基于领域知识的查询生成和基于机器学习的查询生成。接着介绍领域词典的生成、问题语料构建以及分词优化策略。后列举问句特征及其提取方式。最后详细介绍了三种 SPARQL 生成方式, 且通过 Q_1 和 Q_2 问题的测试, 证明回答问题的能力要优于纯粹的基于规则或是基于统计的问答系统。

第6章 基于语义网的自动问答的实现

在第 4 章，我们构建完成了中草药领域的语义网，获得了拥有 300 多万三元组关系；在第 5 章，我们通过分析实际可能遇到的问题，以三种不同的策略生成 SPARQL 查询语句，并通过实验论证了方法的可行性。本章我们基于开源框架 JFinal、Apache Jena 语义网应用框架以及 Redis 内存数据库实现了 B/S 架构的问答系统。系统页面前台使用 HTML+CSS+JS，通过 AJAX 与后台进行通信。系统主要以接口的形式供中草药专业知识服务系统的其他子系统调用，并提供一个简单的测试接口的 WEB 界面，如图 6.1 所示。

中草药智能问答系统接口测试界面



麻黄的功效是什么

Search

麻黄的功效如下：

- 平喘
- 利水消肿
- 发汗
- 发汗散寒
- 发汗解表
- 宣肺平喘
- 利水

图 6.1 接口测试界面

6.1 系统架构

系统的总体架构如图 6.2 所示，可分为两大部分。

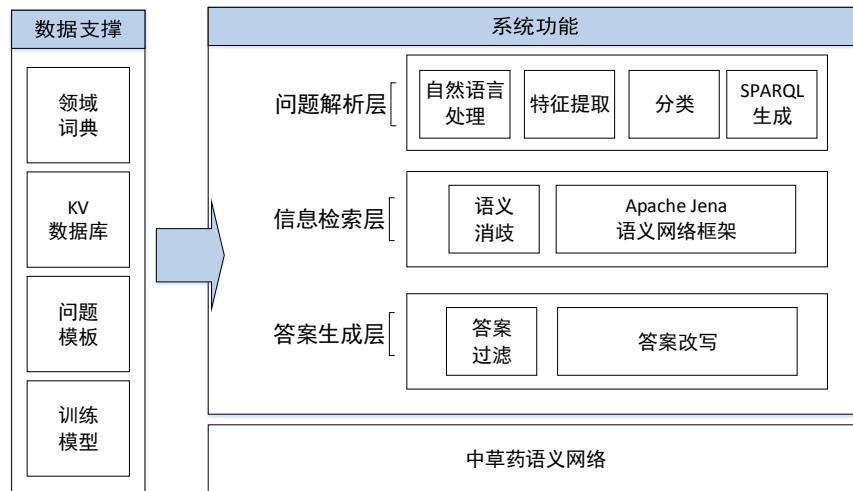


图 6.2 系统架构图

第一部分是数据支持模块，即基础数据和实验数据的准备，用于支撑系统的建设。包括中草药领域词典构建、命名实体 URI 映射库、分类器训练模型以及自定义问题模板。第二部分是系统功能模块，用于处理问答流程。系统功能模块分成四层。最上层是问题解析层，用于分析问题，确定用户意图并把其改写为 SPARQL 查询语句，此模块包含自然语言处理、特征提取、分类、SPARQL 生成四个功能模块。第二层是信息检索层，包含语义消歧和 Apache Jena 语义网框架模块，用于过滤无效问句，同时调用 Jena API 获取系统底层的语义网的信息。第三层是答案生成层，由答案过滤和答案改写构成，主要用于无结果答案的删除以及前台展示信息的添加和改写。最底层为中草药语义网。

6.2 系统功能分析

系统功能的流程图如图 6.3 所示，系统在获得一个用户输入后，把问题封装为一个 Question 对象，内含一个元素为 Answer 对象的列表，此后的所有操作的结果都存储在此对象中。系统在问题解析层首先使用开源自然语言处理工具对问题进行预处理，然后进行多维度的特征抽取，随后调用分类器集合进行类别分类，以层次过滤式的方式尝试生成 SPARQL 查询语句，期间会根据问题特征进行多次分类来缩小 SPARQL 语句的生成域。之后在信息检索层对构造的问句集合进行消

歧，删除无效问句后调用 Jena API 查询语义网获得原始答案。最后在答案生成层的答案过滤模块则会把没有结果的 Answer 对象删除，并对返回的答案进行改写以展示更多的内容。

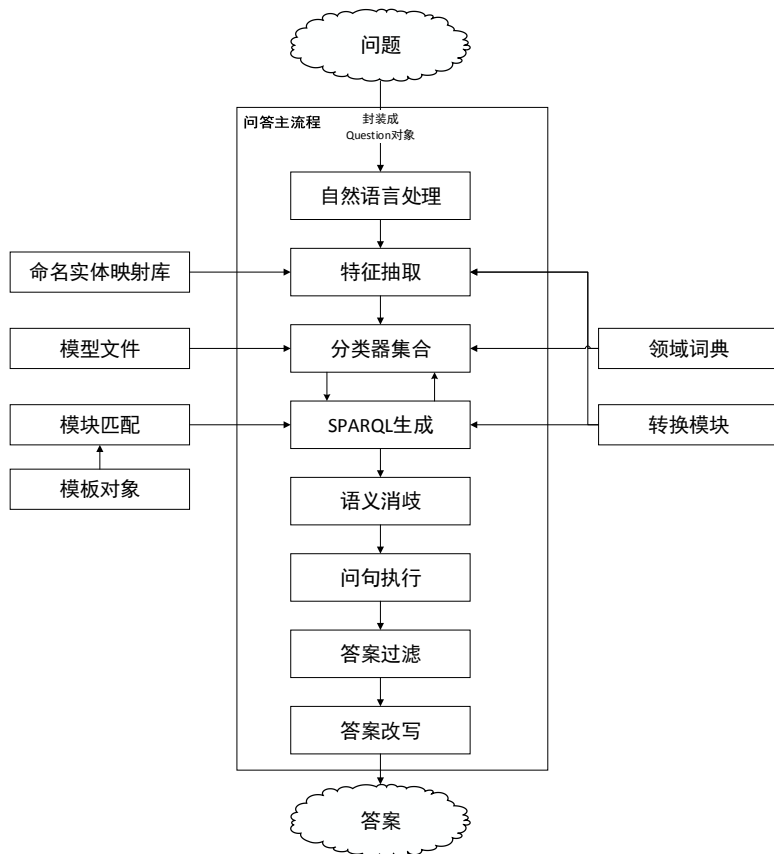


图 6.3 功能流程图

6.3 系统实现

系统实现过程中，按照功能可分成四类，分别为配置模块、应用模块、支撑模块和数据格式。配置模块是 JFinal 框架的配置，与具体功能无关；应用模块是问答过程中对问句的直接处理模块，每一个应用模块属于且只属于问题解析层、信息检索层或答案生成层中的某一层；支撑模块是供应用模块调用的辅助模块，更多起到与外部资源进行交互的作用，这些模块存在一定的复用；数据格式则以面向对象编程为设计理念，是对问答过程中一些处理对象的封装。系统的重点实现

模块为 SPARQL 生成模块，其依据之前特征抽取及分类算法的结果，生成底层的知识库查询语句。

6.3.1 数据格式

数据格式包括问题对象、答案对象、词序对、限制对、模板对象。

1) 问题对象

问题对象是对整个问题的封装，用户问题自进入系统起，便产生一个内容为该问题描述的问题对象，所有的处理都作用于此对象，并且所有的处理结果例如自然语言处理、特征提取、问句生成的结果都存储于此对象，直到最终结束将其答案字段返回给用户，问答过程结束，问题对象销毁。封装其中的字段主要有内容、答案、分词后长度、分词结果、词性标注结果、语法树标注结果、依存关系父节点编号、实体计数、实体中出现了概念类别、实体对应的 URI 列表、概念计数、概念中出现的属性/关系、概念对应的 URI 列表、答案列表、限制列表以及特征字典。其中答案列表中的元素类型为答案对象，特征字典中值对应的是元素类型为词序对的列表、限制列表中的元素类型为限制对。

2) 答案对象

答案对象封装了原始答案、产生该答案的 SPARQL 语句以及对答案的描述。原始答案即通过 Jena API 接口搜索语义网后返回的答案，是一个列表的形式，或者是空。SPARQL 语句则是为了提供 Debug 相关信息，并不呈现给用户。而答案的描述是对答案的补充性描述或者是对问答策略的描述。

3) 词序对

词序对是一个封装字符串和整形的 POJO，用于存储词及其顺序，方便处理过程中词序的获取。

4) 限制对

限制对是封装属性名、对于 URI 列表以及属性值的对象，用于存储问题的限制对特征。

5) 模板对象

模板对象是用户自定义模板的抽象，不属于问题对象的内容。该对象封装了模板实体个数、模板变量符号、模板、SPARQL 查询变量、SPARQL 查询语句的

信息，同用户自定义模板内容相符，自定义模板如表 6.1 所示。

表 6.1 问题模板样例

问题模板
2 AX BX AX BX 相同 归经 x SELECT ?x WHERE {AX <http://zcy.ckcest.cn/tcm/med/property#med_tropisw_detail.med_tropisw> ?x. BX <http://zcy.ckcest.cn/tcm/med/property#med_tropisw_detail.med_tropisw> ?x.}

6.3.2 支撑模块

支撑模块包括文件读写模块、Redis 模块、Weka 模块、词典模块、模板匹配模块、转换工具。

1) 文件读写模块

对文件 IO 的进行封装、如读取领域词典、读取模板等皆调用此模块。

2) Redis 模块

Redis 模块封装了对内存数据库的调用，其主要作用是专业领域的命名实体识别以及关键词到 URI 的映射。

3) Weka 模块

Weka 模块封装了对训练模型读取和 weka 分类器的调用。

4) 词典模块

词典模块通过调用文件读写模块获得词表，并对这些词典进行管理，供应用模块调用。

5) 模板匹配模块

模板匹配模块通过调用文件读写模块获得模板信息，封装成模板对象并对这些模板进行管理，供应用模块调用。

6.3.3 应用模块

应用模块包含自然语言处理、特征抽取、风雷器集合、SPARQL 生成、语义消歧、问句执行、答案过滤、答案改写八大模块，是直接处理问题对象的顶层功能模块。

1) 自然语言处理

由第 5 章第 2 节第 3 条的多分词校正可知，系统首先调用引入领域词典的 HanLP 进行分词并以空格作为分隔符拼接为一个完整的字符串，随后调用 FNLPLP 产生对应的分词、词性标注、语法树解析结果，并存储在问题对象的相应字段。

2) 特征抽取

特征抽取实现了第 5 章第 3 节中特征设计所描述的除 Doc2Vec 和 BoW 之外的所有特征抽取方案，并且结果以键值形式存储在问题对象特征字典字段。

3) 分类器集合

分类器集合包含四个，一个基于启发式规则的分类器以及三个基于训练模型的 SVM 分类器。前者用于基于领域知识的生成方式，后三者用于基于机器学习的生成方式，且需要通过 Weka 模块对训练获得的模型进行。

4) SPARQL 生成

SPARQL 实际的实现中没有使用判断是否为有句的分类器，问题领域分类器采用 BoW 加专业词的训练模型，问题属性分类器则使用专业词的训练模型。此外，SPARQL 会根据识别的实体数目采取不同的措施：当问句中只存在一个实体时，对该实体对应的每一个 URI 都生成 SPARQL 查询语句；当问句中存在多个实体时，首先尝试将这些 URI 归一到统一一个领域。其算法流程如图 6.4 所示：

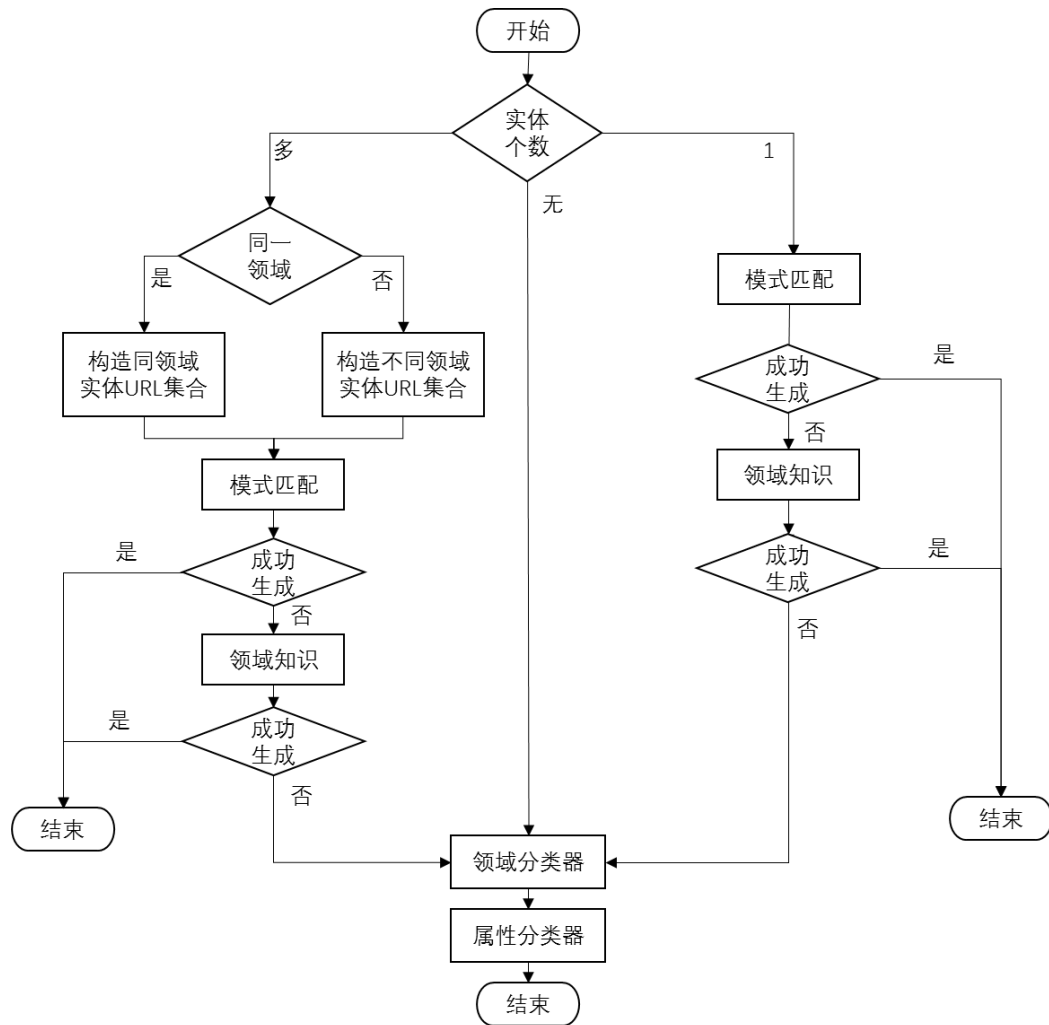


图 6.4 SPARQL 生成流程图

5) 语义消歧义

语义消歧主要在调用 Jena API 执行 SPARQL 查询语句前过滤无效的问题，典型的如实体 URI 以及属性 URI 不在同一个命名空间的语句，提升效率。

6) 问句执行

问句执行通过调用 Jena API 获得原始答案，封装为答案对象。

7) 答案过滤

在返回的答案对象中，可能存在没有结果的情况，对于此类的对象，对象则需要进行移除。

8) 答案改写

对答案对象中的内容进行处理，包括内容的拼接、HTML 标签的添加、说明

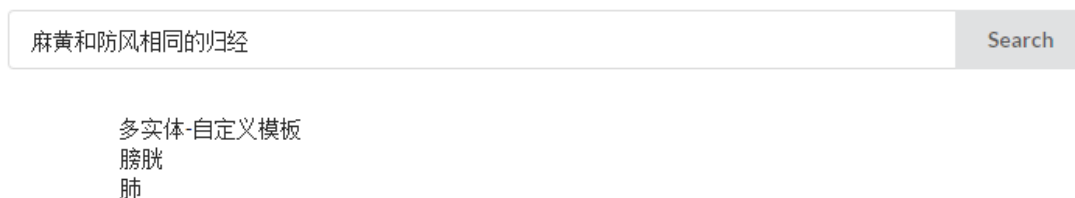
信息的添加，改写为最终呈现的给前台的答案形态。

6.4 系统展示

从第 5 章的分析，我们定义精心构造的问题为 Q_1 ，社区实际的问题为 Q_2 ，并设计了一种层次过式滤的 SPARQL 查询生成策略，即基于模式匹配的查询生成、基于领域知识的查询生成和基于机器学习的查询生成。通过组合来分别也是系统的功能。

6.4.1 基于模式匹配的 Q_1 类问题

如图 6.5 所示，基于模式匹配，体现“相同”关键词的交运算。



麻黄和防风相同的归经

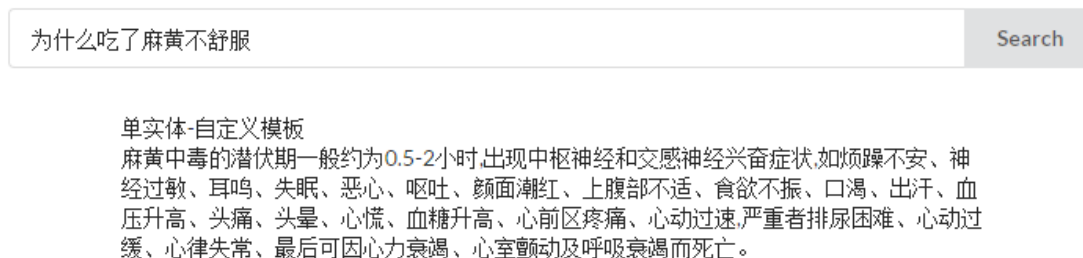
多实体-自定义模板
膀胱
肺

Search

图 6.5 基于模式匹配的 Q_1 类问题

6.4.2 基于模式匹配的 Q_2 类问题

如图 6.6 所示，基于模式匹配，查询了麻黄实体的不良反应信息。



为什么吃了麻黄不舒服

单实体-自定义模板
麻黄中毒的潜伏期一般约为0.5-2小时,出现中枢神经和交感神经兴奋症状,如烦躁不安、神经过敏、耳鸣、失眠、恶心、呕吐、颜面潮红、上腹部不适、食欲不振、口渴、出汗、血压升高、头痛、头晕、心慌、血糖升高、心前区疼痛、心动过速,严重者排尿困难、心动过缓、心律失常、最后可因心力衰竭、心室颤动及呼吸衰竭而死亡。

Search

图 6.6 基于模式匹配的 Q_2 类问题

6.4.3 基于领域知识的 Q_1 类问题

如图 6.7 所示，基于领域知识，自动生成了对应实体的属性，并与 6.4.2 相区别。

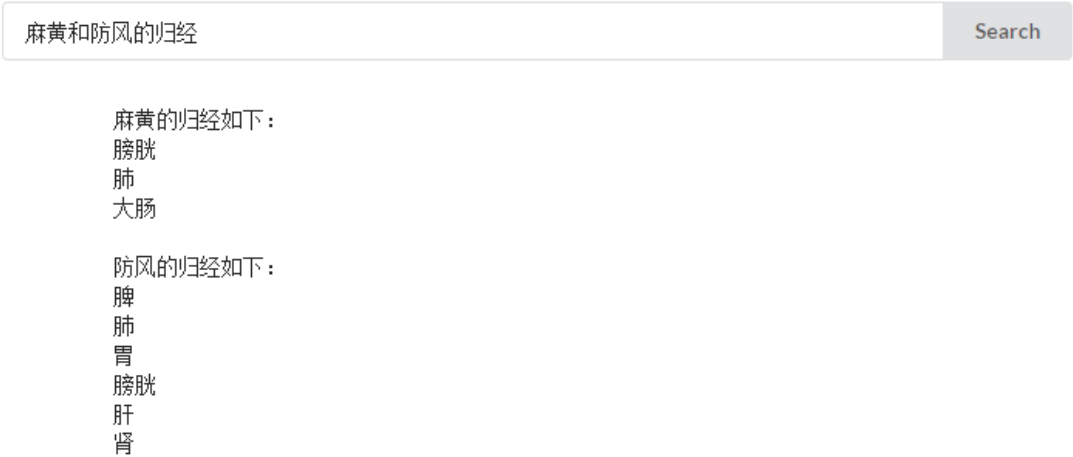


图 6.7 基于领域知识的 Q_1 类问题例一

又如图 6.8 所示，基于领域知识，系统自动生成实体间关系查询语句，并以交形式连接。



图 6.8 基于领域知识的 Q_1 类问题例二

6.4.4 基于机器学习的 Q_2 类问题

如图 6.9 所示，基于机器学习，分类到预定义的 SPARQL 问句。知识库中并没有“肥胖”实体。

方剂-治疗:

秘传飞步丸

黄末子

脾湿汤

六神辅圣丸

附子酒

大风门顶

定痛丸

牛膝汤

开关散

湿风痛风汤

消风饮

松香散

加味解醒汤

乌鱼煎

活血丹

参五秦苁汤

图 6.9 基于机器学习的 Q_2 类问题例

6.5 本章小结

本章介绍中草药专业知识服务自动智能问答的系统架构以及详细实现，并通过实际问题的测试，证明了本文相关方法的可行性以及本系统的高效性和实用性。

第7章 总结与展望

7.1 总结

为服务于中草药专业知识问答, 本文根据已有数据以及实际需求的特点, 研究并实现了基于中草药语义网的自动问答系统。本文主要可分为两个部分, 首先是下层作为知识库中草药语义网的设计与构建, 其中实现了雪花状数据库的信息自动抽取工具; 其次是上层作为应用服务的自动问答系统的研究与实现, 提出了一种层次过滤式的 SPARQL 语句生成策略。

本文首先通过调研问答系统的现状, 包括开放领域以及受限领域, 指出目前中文尤其是中文受限领域的不足以及深入研究的挑战和意义。针对依托项目目前已有专业数据的基础和面向实际大众的需求, 确定了研究知识建模和表示以及问答系统的关键技术, 并实现一个满足领域需求的问答系统的目标。

为此, 本文首先分析了问答系统的国内外现状, 并且重点研究了其中涉及的关键技术, 同时也研究了语义网及其本体、资源描述框架相关的内容, 并调研著名语义网。

随后, 本文参照《中医药学主题词表》构建了中草药的顶层本体, 并设计了 URI、命名空间等, 之后对原有的关系型数据库进行了重构, 同时集成非结构化数据, 设计并实现了自动抽取信息工具生成 RDF 的工具, 最终获得了 300 多万三元组。鉴于本体以及语义网的构建尚无系统性方法, 皆与具体工程相关, 本文提出的方法能很好地应用于可转化为雪花模型的关系型数据库的语义网构建。

接着, 针对受限领域问答系统及其问题的特点, 本文定义了两类即精心构造和社区实际的问题, 并分别设计了基于模式匹配的查询生成、基于领域知识的查询生成和基于机器学习的查询生成用于解决问题到 SPARQL 的映射。由于领域词典以及训练语料的缺失, 本文首先构建了领域词典和标注语料, 同时设计问句特征: 一般特征、语法特征、领域特征。针对模式匹配方式, 设计了弱模式匹配算法; 针对领域知识方式, 定义了多种概念间的关系; 针对机器学习方式, 通过多组对比实验, 论证的方法的正确性。本文采用规则和统计相结合的方式, 使得回答问题的类型大大提升。

最后，根据以上的设计及研究，采用 JFinal 框架实现了 B/S 架构的基于中草药语义网的自动问答系统，并测试能有效回答问题。

7.2 展望

基于受限领域知识库的中文自动问答，本文只是进行了初步探究，其中的每个部分都存在一定的问題：

- 1) 在设计语义网的概念及其关系时也缺少专业知识的指导，很难判断设计是否准确合理，所以需要和领域专家合作交流。
- 2) 目前中草药语义网的构建仅仅是从数据库中抽取信息并进行简单的拆分，但目前存在许多为属性值的内容可以提升为实体。
- 3) 中文自然语言处理中词性标注、语法树解析存在很大的问题，即使处理精心构造的问句。由于是受限领域问答，可以尝试标注领域相关的语料来训练自然语言处理的模型。
- 4) 基于机器学习的问句生成在实验中效果不错，但在实际应用中却很依赖问题形式，需要提取新的特征或改进相关方法。

参考文献

- [1] Dang H T, Kelly D, Lin J J. Overview of the TREC 2007 Question Answering Track[C]. TREC, 2007, 7: 63.
- [2] Cui H, Kan M Y, Chua T S. Soft pattern matching models for definitional question answering. ACM Trans Inf Syst (TOIS)[J]. Acm Transactions on Information Systems, 2007, 25(2):107-108.
- [3] Xu J, Weischedel R, Licuanan A. Evaluation of an extraction-based approach to answering definitional questions[C]. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2004: 418-424.
- [4] Blair-Goldensohn S, Mckeown K, Schlaikjer A H. A Hybrid Approach for QA Track Definitional Questions.[C]. In Proc. of the 12 th Annual Text Retrieval Conference, 2003:185-192.
- [5] Voorhees E M, Tice D M. The TREC-8 Question Answering Track Evaluation[C]. TREC, 1999, 1999: 82.
- [6] 秦兵, 刘挺, 王洋,等. 基于常问问题集的中文问答系统研究[J]. 哈尔滨工业大学学报, 2003, 35(10):1179-1182.
- [7] 王树西, 刘群, 白硕. 红楼梦人物关系问答系统[C]. 第一届学生计算语言学研讨会论文集, 2002:168-174.
- [8] 中国工程科技知识中心 [OL]. <http://www.ckcest.cn/>.
- [9] 中草药专业知识服务子系统 [OL]. <http://zcy.ckcest.cn/tcm/>.
- [10] 申晨. 中草药问答系统的设计与实现[D]. 浙江大学, 2014.
- [11] Apache Jena [OL]. <http://jena.apache.org/>.
- [12] Katz B, Borchardt G C, Felshin S. Natural Language Annotations for Question Answering[C] FLAIRS Conference, 2006: 303-306.
- [13] YAGO [OL]. <http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/>.
- [14] DBpedia [OL]. <http://wiki.dbpedia.org/>.
- [15] FreeBase [OL]. <http://www.freebase.com/>.
- [16] NELL [OL]. <http://rtw.ml.cmu.edu/rtw/>.
- [17] Frank A, Krieger H U, Xu F, et al. Question answering from structured knowledge sources[J]. Journal of Applied Logic, 2007, 5(1):20-48.
- [18] Liu K, Zhao J, He S, et al. Question Answering over Knowledge Bases[J]. Intelligent Systems IEEE, 2015, 30(5):26-35.
- [19] Fader A, Zettlemoyer L, Etzioni O. Open question answering over curated and extracted knowledge bases[C]. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014: 1156-1165.
- [20] 陆汝钤, 石纯一. 面向 Agent 的常识知识库[J]. 中国科学: 技术科学, 2000, 30(5):453-463.
- [21] 王树西, 刘群, 白硕,等. 基于动态知识库的问答系统研究[C]. 全国第七届计

- 算语言学联合学术会议, 2003: 597-602.
- [22] WordNet [OL]. <http://wordnet.princeton.edu/>.
- [23] HowNet [OL]. <http://www.keenage.com/>.
- [24] k-nearest neighbours algorithm [OL]. http://www.wikiwand.com/en/K-nearest_neighbors_algorithm.
- [25] support vector machine [OL].
http://www.wikiwand.com/en/Support_vector_machine.
- [26] naive bayes classifier [OL].
http://www.wikiwand.com/en/Naive_Bayes_classifier.
- [27] decision tree learning [OL].
http://www.wikiwand.com/en/Decision_tree_learning.
- [28] Zhang D, Lee W S. Question classification using support vector machines[C]. International Acm Sigir Conference on Research & Development in Informaion Retrieval, 2003:26-32.
- [29] Bae K, Ko Y. An effective category classification method based on a language model for question category recommendation on a cQA service[C]. Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012:2255-2258.
- [30] Huang Z H, Thint M, Qin Z. Question classification using head words and their hypernyms[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008: 927-936.
- [31] 文勘, 张宇, 刘挺,等. 基于句法结构分析的中文问题分类[J]. 中文信息学报, 2006, 20(2):33-39.
- [32] 孙景广, 蔡东风, 吕德新,等. 基于知网的中文问题自动分类[J]. 中文信息学报, 2007, 21(1):90-95.
- [33] 牛彦清, 陈俊杰, 段利国,等. 中文问句分类特征的研究[J]. 计算机应用与软件, 2012, 29(3):108-111.
- [34] Phan X H, Nguyen L M, Horiguchi S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections[C]. Proceedings of the 17th international conference on World Wide Web. ACM, 2008: 91-100.
- [35] 翟延冬, 王康平, 张东娜,等. 一种基于 WordNet 的短文本语义相似性算法[J]. 电子学报, 2012, 40(3):617-620.
- [36] Liang P, Jordan M I, Dan K. Learning Dependency-Based Compositional Semantics[J]. Computational Linguistics, 2011, 39(2):389-446.
- [37] Zettlemoyer L S, Collins M. Learning to Map Sentences to Logical Form: Structured Classification with Probabilistic Categorical Grammars[J]. Eprint Arxiv, 2012:658--666.
- [38] Wong Y W, Mooney R J. Learning Synchronous Grammars for Semantic Parsing with Lambda Calculus.[J]. Annual Meeting, 2007, 960-967.
- [39] Yih W, Chang M W, He X, et al. Semantic parsing via staged query graph generation: Question answering with knowledge base[C]. Association for Computational Linguistics (ACL), 2015.
- [40] Bordes A, Weston J, Usunier N. Open Question Answering with Weakly Supervised Embedding Models[M]. Machine Learning and Knowledge Discovery in

- Databases. Springer Berlin Heidelberg, 2014:165-180.
- [41] TF-IDF [OL]. <http://www.wikiwand.com/zh/TF-IDF>.
- [42] Zhang D, Lee W S. Web Based Pattern Mining and Matching Approach to Question Answering[C]. TREC, 2002, 2: 497.
- [43] Cui Hang, Kan M-Y, Chua T-S. Unsupervised learning of soft patterns for generating definitions from online news [C]. Feldman S I, Uretsky M, Najork M, et al. Proceedings of the 13th International Conference on World Wide Web (WWW 2004), May 17-20, 2004. New York, NY, USA: ACM, 2004: 90-99.
- [44] Moldovan D, Harabagiu S, Girju R, et al. LCC tools for question answering[C]. Proceedings of the 11th Text Retrieval Conference (TREC 2002), Department of Commerce, National Institute of Standards and Technology, 2002:144-155.
- [45] FNL P [OL]. <https://github.com/xpqiu/fnlp>.
- [46] HanLP [OL]. <http://hanlp.linrunsoft.com/>.
- [47] Weka [OL]. <http://www.cs.waikato.ac.nz/ml/weka/>.
- [48] Chang C C, Lin C J. LIBSVM: A library for support vector machines[J]. Acm Transactions on Intelligent Systems & Technology, 2011, 2(3):389-396.
- [49] Berners-Lee T, Hendler J, Lassila O. The semantic web[J]. Scientific american, 2001, 284(5): 28-37.
- [50] Shadbolt N, Hall W, Berners-Lee T. The semantic web revisited[J]. Intelligent Systems IEEE, 2006, 21(3): 96-101.
- [51] Uschold M, Gruninger M. Ontologies: Principles, methods and applications[J]. The knowledge engineering review, 1996, 11(02): 93-136.
- [52] Uschold M, King M, Moralee S, et al. The enterprise ontology[J]. The knowledge engineering review, 1998, 13(01): 31-89.
- [53] 雪花模型 [OL]. baike.baidu.com/view/5732452.htm.

攻读硕士学位期间主要的研究成果

- [1] 项目：中草药专业知识服务系统，国家财政部专项，2012.1 月 – 至今
- [2] 软件著作权：《中草药基础知识搜索系统》登记号 2014SR033597
- [3] 软件著作权：《企业之友制药助手》登记号 2015SR012246
- [4] 软件著作权：《中草药词典系统》登记号 2015R11L497655
- [5] 论文：Yao, L., Zhang, Y., Wei, B., Qian, H., & Wang, Y. (2015). Incorporating Probabilistic Knowledge into Topic Models. *Advances in Knowledge Discovery and Data Mining*. Springer International Publishing. PAKDD 2015.

致谢

三年的研究生生涯转瞬即逝，在这里我对每位老师，同学和朋友表示深深的感谢！

首先，我要感谢我的导师张引老师，她不仅从工程科研上给予我很大帮助，在工作生活上也对我十分关心。在她的培养下，我开阔了眼界，丰富了知识，从一个什么都不会的大学生，成长为可以独立完成项目的研究生。

其次，我也要感谢实验室的魏宝刚老师、吴江琴老师、张寅老师、鲁伟明老师对我平日的指导。

同时，我还要感谢中草药组的各位师兄师弟师姐师妹，特别感谢的王一兵对我技术提升之路的莫大帮助，同时也感谢姚亮、黄祥洲对我科研方向上的指引，感谢张月娇对我情绪的鼓励，感谢胡直峰、李哲蓉帮我分担任务，感谢黎磊、凌超对我工作的关心和帮助，感谢任晓琳、申晨对于刚来实验室我的手把手指导，感谢组里的新生力量张扬扬、张锐和金哲。

此外我要感谢 220 的伊灯、杨善松、高鹏程、李戈、余姚、刘军以及 215 的边亚丽、李彬、谭亮、张占江、龚军、金登科、姜利成、曹欣怡、陈辉、李一鸣，有你们的实验室生活丰富多彩。

最后衷心的感谢我的父母，有你们陪伴的人生道路并不孤单。

感谢在浙大度过的丰富而充实的三年，往事值得回味！

钱宏泽

二零一六年一月