密级：



硕 士 学 位 论 文



论文题目 基于超链数据的中医药语义查询系统

作者姓名 盛浩

指导教师 陈华钧 副教授

姜晓红 副教授

学科(专业) 计算机应用技术

所在学院 计算机学院

提交日期 2011年1月

A Dissertation Submitted to Zhejiang University for the Degree of

Master of Engineering



TITLE: Linked Data based Semantic Query System of Traditional Chinese Medicine

Author: Hao Sheng

Supervisor: Vice-professor Huajun Chen

Vice-professor Xiaohong Jiang

Subject: Computer Application Technology

College: College of Computer Science

Submitted Date: January, 2011

# 摘要

现代计算机信息技术的飞速发展，极大地推动了各科学研究领域的学术与科技进步，同时也积累了大量的科学数据。例如，医学与生命科学数据库，国家地理信息科学数据库，大气基础科学数据库等。如何有效的组织、共享和利用这些数据，将成为科学技术进一步发展的关键手段之一。另外，科学技术的创新与多学科研究领域的交叉融合，也越来越倚重科学数据的发掘、集成以及将其转化为知识与信息的能力。

在本文中，我们从中医药学科的海量信息出发，利用语义本体，语义查询等技术手段，设计并实现了一个面向中医药领域的语义查询系统。该系统将关系型数据库，超链数据等多种数据来源集成到一个统一的平台，并提供语义查询接口，根据用户的查询请求，去查找语义相关的数据，最终把结果以一种集成的方式展示给用户。

本文的主要内容如下：

1．介绍查询系统的应用背景和语义方面的相关技术；

2．分析系统的整体结构设计，分别阐述数据模块、语义本体模块、语义查询模块等子模块的具体功能；

3． 介绍基于中医药关系型数据库和中医药超链数据的语义搜索的具体实现。首先提出一个数据到语义视图的映射，然后在语义视图的基础上，用查询重写算法将来自不同数据源的查询结果语义包装成XML格式后返回，并展示给用户。

4． 提出了一种基于语义相似度的模糊查询技术。该技术以超链数据为基础，结合了词法分类和词频统计等方法，实现了一个混合型的语义相似度模型。利用这个模型，计算出一组与输入关键词在与语义上较为相近的词语集，然后将这一组词语作为查询系统的关键词，实现模糊查询。

关键词：语义查询，超链数据，数据挖掘，语义相似度

# Abstract

目 录

[摘要 i](#_Toc278631895)

[Abstract ii](#_Toc278631896)

[第1章 绪论 1](#_Toc278631897)

[1.1 课题背景 1](#_Toc278631898)

[1.2 技术背景 2](#_Toc278631899)

[1.2.1 语义web和知识表示框架 3](#_Toc278631900)

[1.2.2 URI与XML 4](#_Toc278631901)

[1.2.3 资源描述框架RDF 4](#_Toc278631902)

[1.2.4 本体与本体语言 5](#_Toc278631903)

[1.2.5 Sparql查询语言 6](#_Toc278631904)

[1.3 相关工作 6](#_Toc278631905)

[1.3.1 中医药信息化与中医药本体建设 7](#_Toc278631906)

[1.3.2 中医药语义网格平台 7](#_Toc278631907)

[1.4 本文工作与论文结构 8](#_Toc278631908)

[1.5 本章小结 9](#_Toc278631909)

[第2章 语义查询系统架构 9](#_Toc278631910)

[2.1 搜索系统总体设计与框架 9](#_Toc278631911)

[2.2 数据模块 9](#_Toc278631912)

[2.2.1 关系型数据库模型 9](#_Toc278631913)

[2.2.2 超链数据模型 9](#_Toc278631914)

[2.3 语义本体模块 9](#_Toc278631915)

[2.4 语义查询模块 9](#_Toc278631916)

[2.5 展示模块 9](#_Toc278631917)

[2.6 本章小结 9](#_Toc278631918)

[第3章 基于异质异构数据库的语义搜索 10](#_Toc278631919)

[3.1 中医药数据库介绍 10](#_Toc278631920)

[3.2 基于sparql语义查询设计与实现 10](#_Toc278631921)

[3.2.1 sparql语义查询实现 10](#_Toc278631922)

[3.2.2 sparql查询重写 10](#_Toc278631923)

[3.3 具体应用案例展示 10](#_Toc278631924)

[3.4 本章小结 10](#_Toc278631925)

[第4章 基于超链数据的语义搜索 11](#_Toc278631926)

[4.1 背景概述 11](#_Toc278631927)

[4.1.1 Linked Open Data介绍 11](#_Toc278631928)

[4.1.2 相关研究工作 11](#_Toc278631929)

[4.2语义搜索算法设计 11](#_Toc278631930)

[4.2.1 数据抽取与预处理 11](#_Toc278631931)

[4.2.2 LDSS语义搜索算法设计 11](#_Toc278631932)

[4.3 算法实现与实验分析 11](#_Toc278631933)

[4.4 本章小结 11](#_Toc278631934)

[第5章 基于超链数据的语义相似度 11](#_Toc278631935)

[5.1 背景介绍 11](#_Toc278631936)

[5.1.1 Linked Open Drug Data 12](#_Toc278631937)

[5.2 语义相似度模型 14](#_Toc278631938)

[5.2.1语义相似度与研究意义 14](#_Toc278631939)

[5.2.2基于词性分类的语义相似度 15](#_Toc278631940)

[5.2.3 基于词频统计的语义相似度 15](#_Toc278631941)

[5.2.4 混合型相似度模型 15](#_Toc278631942)

[5.3 算法实现与实验分析 15](#_Toc278631943)

[5.4 本章小结 15](#_Toc278631944)

[第6章 总结与展望 15](#_Toc278631945)

[参考文献 15](#_Toc278631946)

[攻读硕士学位期间主要的研究成果 15](#_Toc278631947)

[致谢 15](#_Toc278631948)

# 第1章 绪论

## 1.1 课题背景

伴随计算机科学与信息技术的不断发展，自然科学、人文科学、社会经济等领域积累了大量的科学数据。作为当代科学技术创新与研究的重要资源与基础，这些科学数据蕴含着丰富的科学价值、社会价值与经济价值，同时也代表了人类文明的发展和社会的进步。随着21世纪全球信息化时代的到来，各国政府、学术界研究机构和科学研究人员越来越重视对海量科学数据信息的获取、挖掘与利用。同时，大规模的数据共享与分析也成为当前科学技术研究的一个重要特色。因此，科学高效地利用、分析、集成与共享科学数据的水平，已经成为一种衡量国家科技水平与综合实力的重要指标。

中医学是中华民族5000多年历史文化沉淀下来的瑰宝，积累了大量的珍贵文献与临床实验数据。中医学作为中华民族的传统医学，其理论基础成型于春秋战国时期，而后经过数代医学专家的研究和发展，逐渐形成了以阴阳五行、气血经络为基础，分中医临床疾病、中药药理、中医针灸、中药化学实验、中医个案诊疗等多个中医学科的综合性医学理论。随着计算机技术的不断深入，传统中医学与计算机信息科学的融合而成的中医药信息学极大地推动了中医学的普及推广和发展。中国中医科学院与浙江大学计算机学院CCNT实验室合作的中医数字信息化建设项目，经过大量领域专家和研究人员十余年的不懈努力，积累了大量的中医药实验数据，开发了以DartGrid为代表的中医药语义网格平台。

目前针对中医学的信息检索系统存在以下几个问题：

第一、中医药学科拥有海量数据，数据的表示形式多样化，无法实现异质异构数据库之间的透明化访问与共享。

第二、中医药领域内的疾病、药物等概念之间存在着大量的关联关系，仅仅通过基于字符串的查询，无法满足检索需求。因此我们需要利用语义本体技术等手段来实现这一目标。

第三、之前开发的系统针对于单个中医库，系统之间的整合与维护相对困难，缺乏一个统一的用户访问界面。

在这个背景下，我们针对中医药领域数据的特点，通过构建语义本体等手段来定义领域概念之间的关联，使用语义查询来解决异质异构数据资源的统一访问，设计并开发出这个基于语义本体的中医药语义查询系统。

## 1.2 技术背景

互联网的出现，改变了人们交流信息和获取知识的方式。人们可以通过互联网方便地搜索和查看网页、阅读新闻、购买商品、与亲朋好友交流等等。但是现在的互联网应用程序并不是十分完美的， 它不能完全满足用户的需求。就如我们最常用到的像Google、百度和Yahoo这样优秀的搜索引擎也存在下面这些问题：

1. 信息准确性不高。尽管搜索引擎会根据网页的相关性进行排序，把相关性较高的网页排在前面，但是由于海量的网页数据，很多有价值的信息往往容易被淹没在浩瀚的数据海洋中。

2．查询结果是基于字符串的匹配，对关键词依赖过高。由于在语言上存在一词多义、一义多词的情况，可能很多网页包含的关键词在语义上和用户输入的关键词是相似或者相近的，但是不会出现在搜索结果中，很多情况下这些内容正是用户所需要的。

3．缺乏语义关联。用户所需要的信息往往是多方面的，比如搜索一个手机型号，用户需要了解这个手机的品牌、参数、功能、价格和用户评价等等多方便的信息。目前的搜索引擎在语义上还缺乏一定的支持。

为了解决这些问题，W3C提出了语义web的概念，它的目标是建立一套标准，用于表示网络上的各种资源，以及通过URI将相关的信息资源联系在一起。下面就其中一些关键的标准和技术做一个介绍。

### 1.2.1 语义web和知识表示框架

语义web(Semantic Web) 是由万维网之父蒂姆·伯纳斯·李(Tim Berners-Lee)在1998年提出的一个概念，它的想法是通过给互联网上的网页文档数据添加语义标签，从而将整个互联网的信息联系在一起，使得计算机更容易地处理这些信息。随后，万维网标准化组织W3C语义web工作组制定了一系列相关的标准和框架。

语义web的知识表示框架如图1.1所示。其中XML作为语法层；RDF作为

数据层；Ontology作为语义层；Logic提供了基于本体层的逻辑推理机制；而证据层支持代理之间的证据交换。下面会对这些概念作个具体的介绍。

Rules

Trust

Proof

Data

Logic Vocabulary

Documents

Ontology Vocabulary

RDF + RDF Schema

XML + NS + XML Schema

Unicode

URI

图1.1 语义web技术的层次结构

### 1.2.2 URI与XML

URI统一资源标识符（Uniform Resource Identifier)是用于标识互联网上某一个资源的字符串。它允许用户对网络中的资源进行定义与操作，计算机则可以根据URI来定位到该资源。

XML(Extensible Markup Language)是一种可扩展的标记语言，用于定义数据的交互格式。它的核心是在原有的文本数据上添加自定义的标签来表示不同的含义，方便计算机处理，并且有第三方的类库支持XML文件的解析。需要注意的是XML这一层并不涉及到任何语义方面的知识。

### 1.2.3 资源描述框架RDF

RDF(Resource Description Framework)是由W3C制定的一种用于描述web信息的数据模型标准，现在已经作为语义web资源的数据交换通用标准。它通常包含了一组陈述(Statement)，每一个陈述是由一个三元组(Triple)组成。每个三元组包含三部分：主体(Subject)、谓语(Predicate)、和客体(Object)。其中主体可以是网络中的任何资源(Resource)，例如书籍、人物、建筑、地点等等，每一个资源通常都用一个URI表示。谓语描述的是主体与客体之间的关系，比如书籍与作者之间存在“written by”关系，人物与城市之间存在“born in”关系。客体既可以是资源(Resource)也可以是文字(Literal)。图1.2是一个用XML格式表示RDF数据的例子。在这个例子中描述的主体是一张名字叫Empire Burlesque的唱片，其中有4个谓语也就是4条关系，分别是唱片的演唱歌手、出版公司、价格和出版年份。这样四条RDF的陈述可以用图1.3这个关系图来表示。

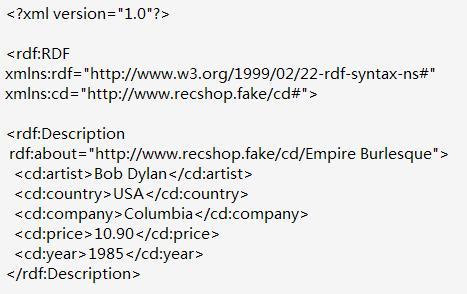


图1.2 一个RDF的例子

Bob Dylan

USA

cd:country

cd:artist

**http://www.recshop.fake/cd/Empire Burlesque**

cd:company

Columbia

cd:year

cd:price

10.90

1985

图1.3 RDF数据转化过来的RDF关系图

### 1.2.4 本体与本体语言

在RDF层上面的就是本体层。什么是本体(Ontology)？本体这个词最早来源于希腊语中的哲学术语“onto”，即存在的意思。从哲学的角度看，本体关注的是世界上存在的物质或者实体，任何只要是存在的东西都可以称之为本体。

在计算机领域，我们使用本体来描述和定义某一个领域里的概念、现象和实体，以及这些实体之间的关系。本体论的建立，可以使特定领域的知识、词汇、数据具有一个统一的标准去访问和表示。因此，通过运用本体我们可以使用语义视图来管理数据；为用户提供语义查询；支持分布式的异质异构数据资源的集成与透明访问。

### Sparql查询语言

Sparql是由W3C开发的标准语义web查询语言，就像SQL查询语言用于查询数据库数据，Sparql用于在RDF数据中查询相匹配的实例。下图是针对图1.2的一个sparql查询例子：

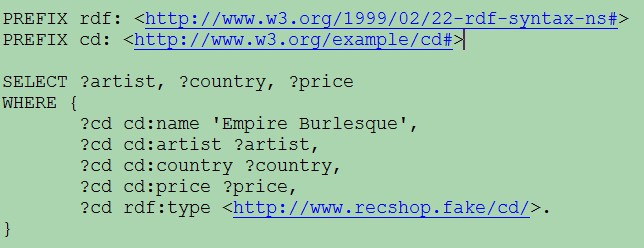


图1.4 一个简单的sparql查询例子

在这个例子中，?cd 是一个类型为<http://www.recshop.fake/cd/>的变量，它有四个属性name, artist, country和price。其中name是已知条件，它的值是“Empire Burlesque”，其他三个是未知变量。这个sparql语句查询的是名字为Empire Burlesque的唱片的演唱者、国家和价格。

## 相关工作

中医药领域是一个数据密集型的领域，积累了大量的文献数据。但是由于中医理论的复杂性和汉语在语义表达的固有特点，导致中医药领域知识的描述和表达会出现多种不同的方式。另外，互联网的出现为人们获取信息知识提供了一条新的途径，不管中医药领域的研究人员或是普通用户也希望通过互联网这个平台能够方便快捷地共享和利用中医药的信息资源。浙江大学计算机学院CCNT实验室和中国中医科学院信息所合作的中医药信息化联合实验室正致力于解决中医药领域的知识表达和信息共享等方面的问题。下面分别对中医药的本体建设和数据网格平台做一个详细的介绍。

### 中医药信息化与中医药本体建设

中医药领域本体的建设目的是为了解决中医药领域内异质异构数据资源的统一知识表达和透明化访问。它包含两个部分，一方面是结合中医药领域专家的知识，将中医药领域内的一般性概念现象进行提取和归纳，最后总结出一套共性的本体。并通过语义本体的推理、分类等人工智能方法，将不同中医领域内的本体知识统一起来形成一个整体的知识表达框架。另一方面，实验室推出的中医药本体开发平台为用户提供了一个更为方便快捷的本体编辑和维护工具。它具有以下几个特点：便捷性。用户不需要安装额外的软件，直接通过互联网就能接入到本体开发平台进行加工；协同性。支持全国各地的用户同时在本体开发平台上加工；安全性。系统采用了多级用户账号权限管理，保证了平台的安全性；高效性。由于系统采用了ajax异步传输技术，使得用户操作的时候不需要一次性的载入所有的数据，提高了程序的性能。

### 中医药语义网格平台

DartGrid是浙江大学计算机学院CCNT实验室自主研发的语义网格平台软件。它采用语义技术实现了异质异构数据库的统一化访问，并支持语义本体层的语义查询功能。目前，基于DartGrid构建实现的中医药数据网格平台集成了大约 40多个专题的中医药科学数据库资源。该平台包含了100多个语义类和800多个语义属性关系，向用户提供了一体化的语义集成查询服务，方便用户从海量的中医药数据中获得更加精确、全面的信息。

## 本文工作与论文结构

本文从中医药的海量信息出发，利用语义网络，语义本体等相关技术，实现了一个集成中医药临床疾病库、药理数据库和中医药Linked Data的多数据源的中医药语义查询系统。本文的主要工作分为以下几个部分:

* 设计并实现语义查询系统的整体架构，分为前台界面模块、语义查询逻辑模块、语义本体模块和数据模块。
* 针对不同中医药科学数据库的特点，实现了sparql语义查询的自动构造，支持多属性、多级关联的语义查询以及单属性的统计查询。
* 实现中医药Linked Data与关系型数据库的融合，支持对Linked Data的语义查询访问。
* 提出一个基于词性分类和词频统计的语义相似度模型，用于计算两个事物之间的语义相似度。该模型用于实现查询系统的模糊查询。

论文的章节安排如下：

第一章介绍论文相关的应用背景和技术背景，重点介绍中医药的发展现状和信息利用、共享方面存在的问题。

第二章详细介绍了中医药语义查询系统的整体设计框架，然后分模块的说明系统各个部分的功能和实现。

第三章是关于异质异构关系型数据库的语义查询，着重介绍了sparql语义查询的设计与实现。

第四章对基于Linked Data的语义查询做了详细的介绍，包括Linked Data的数据预处理，语义本体查询在Linked Data的应用实现。

第五章提出了一个基于词性分类和词频统计的语义相似度模型，重点介绍语义相似度算法的实现以及模型在模糊查询中的应用。

第六章对本文工作的贡献和存在的问题做了总结，并提出了新的工作方向。

## 1.5 本章小结

本章首先对论文相关的课题背景做了介绍，分析了当前中医药领域在信息化发展道路上和知识获取利用中存在的一些问题，并点出了本文工作的必要性和意义，就是为了解决中医药领域的知识共享与利用。

其次，介绍了本文工作所涉及到的语义web等相关的技术背景，重点说明这些技术在解决上述问题中起到的作用，同时帮助读者理解相关的技术术语。

第三，介绍了实验室在中医药本体建设和语义网格平台方面的相关工作情况，这些也是本文工作的基础。最后描述了本文的主要工作和文章的组织结构，突出每章的重点，帮助读者阅读。

# 第2章 语义查询系统架构

## 2.1 系统总体设计与框架

随着Internet和浏览器技术的不断发展，B/S结构(即浏览器/服务器)的互联网应用程序越来越受到企业和用户的欢迎。相比较传统的C/S(客户端/服务端)结构，B/S结构具有如下几个优点：

1. 跨平台，无需安装客户端，只需要浏览器即可访问。
2. 开发速度快，可以快速地添加新的功能。
3. 程序升级维护方便，只需要更新服务端程序，所有用户就可以在浏览器上看到更新后的应用。
4. 大部分的业务逻辑都放在服务端，减轻了客户端的压力。

但是也同样存在一些缺点，比如，响应速度不够快，界面表现形式不够丰富等。因此需要根据实际的应用需求和用户使用习惯和特点来选择适合的系统架构。在前面的章节我们提到中医药领域的信息化建设正在快速的发展中，同时中医学的理论知识体系也随着计算机信息化的融入而不断变化发展，因此针对中医药知识利用和分享的语义查询系统在功能上也会不断的更新与改进。另一方面，查询系统的使用者多为中医药领域的相关研究人员，对计算机软件的安装和维护不是很了解。基于上述这些原因我们的中医药语义查询系统采用B/S结构，前台用户界面使用Flex+XML开发，后台逻辑使用java开发，这样既能保证用户界面的表现力，给用户提供良好的操作体验，同时又能保证系统的快速更新和维护。

系统主要包含一下几个功能模块：底层数据模块（包含关系型数据库和Linked Data两部分）、语义本体模块、语义查询模块和界面展示模块。系统的整体架构设计如图2.1。

数据库

数据库

展示层

逻辑层

本体层

图2.1 语义查询平台整体架构

界面展示层和逻辑层属于前端，后台部分包括本体层、数据层和DartGrid。语义查询服务的整个流程是界面展示层获得用户输入的关键词，然后通过web service将查询条件传递给逻辑层；逻辑层根据不同的查询请求类型自动构造针对语义本体的Sparql查询语句，接着将Sparql查询语句传递给本体层，在本体层中的Sparql解析器负责解析传递过来的查询语句，查找对应的本体实例，然后根据本体和数据库之间的映射关系生成一组SQL查询语句，最后执行SQL语句，将查询到的结果进行语义封装最终展示给用户。

## 2.2 数据模块

底层数据模块分为两个部分，一部分是关系型数据库，其中包含了临床疾病库、药理库、化学库、针灸库和个案库。这些数据的来源是全国各地中医药领域的研究人员将大量的中医文献信息中抽取出具体的数据然后通过数据加工平台录入到对应的数据库。由于中医学本身的复杂性，因此各个中医数据库的表结构和关联关系都大不相同，在这一层的工作就需要我们结合中医学的知识，理清数据库的表结构以及表之间的关联关系，这也是本体层的技术，它能帮助我们更好地构建中医药本体。另一部分是超链数据（Linked Data）。在本文中，我们使用Linked Open Drug Data作为数据集，在这个数据集中，包含了上百种药物以及药物与疾病、基因、功效之间的关系。我们在第四章中会详细介绍基于Linked Data的语义查询实现。

## 2.3 语义本体模块

本体层的主要工作是建立中医药语义本体。这部分工作是由中医药领域专家和信息技术人员共同合作完成的。首先领域专家根据中医学的理论，系统地归纳和整理中医药领域的基本概念和现象，例如疾病、药物、症状等。然后把这些概念抽象成本体，并定义本体的属性和本体之间的关联关系。信息技术人员在这个基础上将中医药本体按照RDF、OWL等本体语言标准进行本体建模，在这个过程中需要考虑本体类之间的层次关系，构建出符合规范和领域知识的中医药本体。

我们使用实验室自主开发OntoTool本体编辑工具来构建本体。相比目前比较流行的Protégé本体工具，OntoTool加入了一些新的功能，它可以使用拖拽的方式自动生成本体之间的关联关系，并支持以XML格式导出已经构建好的本体，其中包含本体URI，本体属性以及本体与数据库表字段之间的映射关系。

## 2.4 语义逻辑模块

语义逻辑模块的主要任务是完成语义查询的自动构造和查询结果的语义封装。在这里语义查询指的就是针对本体的Sparql查询。整个查询过程分为以下几个步骤：

1. 接收展示层传递过来的查询请求，包含查询的数据库、查询关键词、查询入口和查询关联项。
2. 读取本体文件。根据查询条件，查找对应的本体信息，包括查询入口对应的本体和查询关联项对应的本体，然后给本体的各个属性设置唯一的查询变量名称，生成符合Sparql语法的查询语句。
3. 将生成的Sparql查询语句交给Sparql解析器，等待底层查询的结果返回。然后将结果包装成XML格式，返回给前台用户界面。

## 2.5 展示模块

展示层的目的是为用户提供良好的系统访问界面。目前，FLASH技术和RIA富互联网应用程序的普及，为我们提供了一种全新的选择，它比传统网页具有更为丰富的用户体验和优秀的视觉效果。我们系统采用的Flex就是其中一种RIA技术。Flex包含了丰富的控件，可实现拖拽、数据表格、列排序、分页等功能。另外它的开发效率很高，程序员可以方便的开发出界面原型。在中医药语义查询系统的开发中，我们也采用了这样的方式，根据需求快速的开发出界面原型，然后交给用户。这样用户就可以用一种直观的方式来对界面原型提出修改意见，开发人员根据修改意见进行修改，经过多次迭代，直到满足用户的需求，最终开发出符合用户操作习惯和审美要求的用户界面来。

## 2.6 本章小结

本章主要介绍了中医药语义查询系统的整体架构设计。系统的设计充分考虑到中医药领域的发展状况以及系统使用人员的具体需求和知识背景。系统采用B/S的架构，一共分为四个模块，从上到下一次是界面展示模块、语义逻辑模块、语义本体模块和底层数据模块，前两块属于系统前台，后两块属于后台服务端。各个模块之间相互独立，通过接口的方式进行通信。使用这样的架构，降低了系统各模块的耦合度，并且具有很高的扩展性和灵活性，方便系统的维护和功能更新。

# 第3章 基于异质异构数据库的语义查询

## 3.1 中医药数据库介绍

中医药是中华民族5000多年文明史中流传下来的一颗璀璨的瑰宝，积累了大量珍贵的文献。中医药信息化工作中的一个重要部分就是把这些文献中的知识提炼并保存到不同的数据库中。浙江大学计算机学院CCNT实验室与中国中医科学院信息所共同合作的中医药数据共建加工平台帮助全国十余个中医药数据加工中心完成了大量的中医数据录入工作。目前共有各类数据库40余个，数据总量大约110万条，包括中医药期刊文献数据库、临床疾病诊疗数据库、中药方剂数据库、中药药理实验数据库、中药化学实验数据库，中医针灸个案等相关数据库。

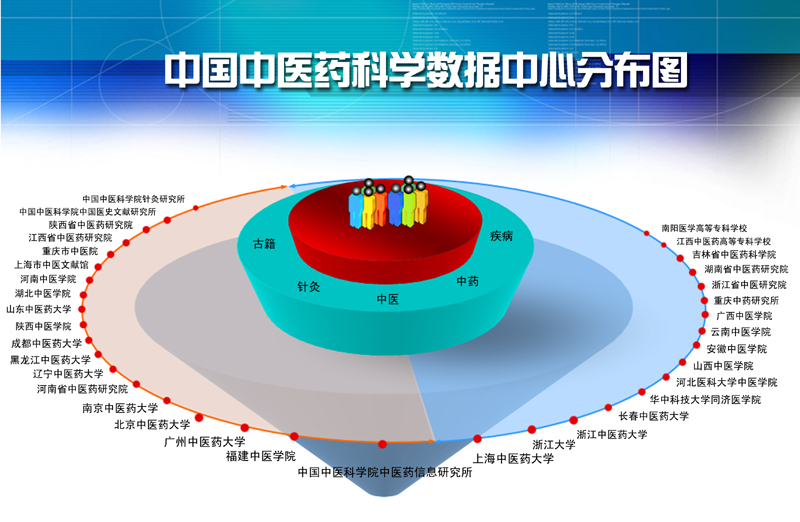


图 3.1 中医药科学数据中心分布图

## 3.2 中医药数据库语义查询实现

目前已经完成对中医药疾病临床研究数据库和中药药理数据库的语义查询。

本章以临床疾病库为例子，具体介绍基于数据库的语义查询系统实现。

### 3.2.1 中医药临床数据库与临床库本体

中医疾病临床研究数据库收录1988年以来公开出版在中文生物医学期刊上的，临床病例数在10例及10例以上的文献，或者是文献中所研究的病例数不足10例，但是对同一种疾病进行的治疗研究的文献。数据库主要采用文献数据结构化处理的方法，根据文献标引的特征词采集文献所研究的疾病和证候的诊断及其中医药、针灸、按摩等治疗信息。通过对文献数据结构化的处理，找出一些中医理法方药之间的关系，实现数据间的关联检索和主题概念的转换检索，为临床科研人员提供一种新的文献检索利用的方法及途径。中医临床疾病库共有21张表，表之间的关系图如图3.2所示。根据表之间的关系我们把这些表分为三类：

* 临床研究文献（包括临床文献、临床诊疗、实验检查等信息）；
* 临床研究对象（包括疾病、症状、证候）；
* 临床治疗方法（包括药方疗法、西药疗法、针灸疗法、按摩疗法和其他疗法）。

按照上述分类，我们可以建立临床库的本体，如图3.3所示。其中研究对象和治疗方法是两个父本体，研究对象有三个子本体分别是疾病、症状和证候。治疗方法有五个子本体，分别是药方疗法、西药疗法、针灸疗法、按摩疗法和其他疗法。临床文献、研究对象和治疗方法两两之间关联。

临床研究主表

抽象临床研究对象

疾病

实验检查

临床诊疗

对照组

证候

临床症状

抽象具体治疗过程

西药疗法

其他疗法

药方疗法

按摩疗法

针灸疗法

穴位

药物

治疗过程对象关联

疾病证候关联

证候症状关联

疾病症状关联

图3.2 临床库表结构关系图

证候

疾病

症状

研究对象

治疗方法

药方方法

西药方法

针灸方法

按摩方法

其他方法

临床文献

对象-文献关联

治疗方法-文献关联

对象-治疗方法关联

图 3.3 临床库本体示意图

### 3.2.2 基于本体的语义查询实现

## 3.3 具体应用案例展示

## 3.4 本章小结

# 第4章 基于超链数据的语义搜索

## 4.1 背景概述

### 4.1.1 Linked Open Data介绍

### 4.1.2 相关研究工作

## 4.2语义搜索算法设计

### 4.2.1 数据抽取与预处理

### 4.2.2 LDSS语义搜索算法设计

## 4.3 算法实现与实验分析

## 4.4 本章小结

# 第5章 基于超链数据的语义相似度

## 5.1 背景介绍

Linked Data作为语义网的一个重要组成部分，它将互联网上的关联数据联系在一起，方便用户查看与共享。与传统网页相比，Linked Data具有以下几个优势：

1. Linked Data采取RDF作为它的表示模型，是一种结构化的数据，方便计算机处理。
2. 具有更强的关联性，它通过URI命名的方式，将关联信息整合在一起。比如，查看“Tim Berners-Lee”的个人信息，就能通过关联找到他发表的论文、出版的书籍等等信息。
3. Linked Data有大量的数据集，包含信息科学、生命科学、医学科学等众多自然科学学科。在不同学科的数据之间，存在着大量的关联关系。

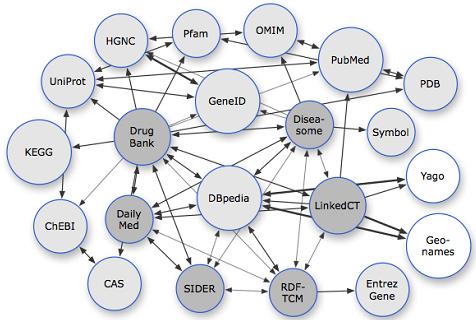


图5.1 Linked Open Data Cloud

### 5.1.1 Linked Open Drug Data

伴随着生命科学与医疗卫生科学的发展，各种各样的医学信息例如，药物、疾病、基因、临床实验，药理治疗等信息呈现出极速增长的趋势。Linked Open Drug Data的目标是将不同来源的医学数据用Linked data的方式集成起来，方便研究人员获取，处理与使用，并最终服务于医学研究。

目前，越来越多的医学科学数据正加入到Linked Open Drug Data[1]这个数据云当中。图5.1表示的是Linked Open Drug Data这个项目现在已经发布的数据集。据统计，其中包含约超过800万条的RDF三元组数据。下面简单介绍几个主要的数据集。

1. Linked Clinical Trials[2] 临床实验数据集，它来源于ClinicalTrials.gov数据库，其中保存了来自158个国家的大约60,000条临床实验。每一条临床实验有一个简单的说明，以及它关联的病例、实验检查、实验条件等等。这些数据是将原有的XML格式数据转换成关系型数据，并保存在relational与XML混合类型的数据库中，然后利用D2R Server[3]把这些数据用RDF的形式发布出来。
2. DrugBank[4]是一个关于西药药物的数据集。它包含5000多种药物的详细信息，例如化学数据、药学数据、制药配药数据。同样地，它也是通过D2R Server来发布到Linked Data数据云中。
3. DailyMed[5]来自美国国家医药数据库，它提供了大量的高质量的药物数据信息。数据的覆盖范围包括药物化学结构组成、药物作用机理、药物用途、药物禁忌、副作用等。
4. TCMGeneDIT[6]是一个关于中医药的数据库，它包含大约848种不同的中草药，以及它们相关的药物组成、药物功效、关联基因和治疗的疾病。这些关联关系都是通过文本挖掘等技术从大量现存的中医文献出提取出来的。
5. <http://esw.w3.org/HCLSIG/LODD>
6. <http://www.linkedct.org>
7. Bizer, C., Cyganiak, R.: D2R Server - Publishing Relational Databases on the Semantic Web. Poster at the 5th International Semantic Web Conference, 2006.
8. <http://www4.wiwiss.fu-berlin.de/drugbank/>
9. <http://www4.wiwiss.fu-berlin.de/dailymed/>
10. Fang, Y.-C., Huang, H.-C., Chen, H.-H., Juan, H.-F.: TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining. BMC Complementary and Alternative Medicine. 8(1):58, 2008.

## 5.2 语义相似度模型

### 5.2.1语义相似度与研究意义

语义相似度表示的是两个单词或者两个事物之间的语义相似程度。例如，“汽车”和“摩托车”在语义相似程度上就要比“牙刷”和“电脑”来的高。可以看出，语义相似度涉及到词性、词法以及语义概念。不同与字符串相似度，语义相似度具有不确定性和一定的主观性。

语义相似度被广泛用于信息检索、文本分类、自然语言处理等研究领域。比如，目前的搜索引擎通过用户输入关键字，系统自动去查找包含目标关键字的网页，然后将结果排序输出。举个例子，用户输入关键字“电脑”，搜索引擎自动去后台的数据库查找包含字符串“电脑”的相关网页，然后按相关程度输出查找结果。但是它不会去查找包含字符串“计算机”的相关网页，因为计算机并不知道“电脑”和“计算机”在语义上表示的是同一个意思。因此这就需要我们利用自然语言处理和语义相似度分析等手段来帮助计算机更好地理解词语在语义层面的上的意思，进而为用户提供更为人性化的服务。

一般地，我们用一个数学公式来表示两个词语间的相似度，公式如下：

### 5.2.2基于词性分类的语义相似度

### 5.2.3 基于词频统计的语义相似度

### 5.2.4 混合型相似度模型

## 5.3 算法实现与实验分析

## 5.4 本章小结

# 第6章 总结与展望

# 参考文献

# 攻读硕士学位期间主要的研究成果

# 致谢