|  |
| --- |
| 中国中医科学院中医药信息研究所 |
| **基于文本的中医药语义关系发现研究** |
| 技术报告 |
|  |
| **于彤** |
| **2013/9/27** |

|  |
| --- |
| 中医药学语言系统是中国中医科学院中医药信息研究所于2002年起开始研制的一项大型中医药术语系统，语义网络是语言系统的重要组成部分。为了促进中医药学语言系统的发展，提高系统中语义关系的准确性，本课题拟进行基于文本的中医药语义关系发现研究，以求通过对中医药文献的文本进行分析，并与中医药学语言系统现有语义网络结合，得到准确的语义关系。  研究主要内容包括：（1）对目前中医药学语言系统语义网络进行分析，发现问题，并归纳得到基于语义类型层的语义关系网；（2）对中医药文献文本进行分析，发现获得语义关系，并归纳为基于语义类型的语义网络；（3）对文本获得的语义网络与目前的语义关系网进行比较，优化，形成新的语义网络。 |

**目录**

[1 项目概述 1](#_Toc376030710)

[1.1 项目目标 1](#_Toc376030711)

[1.2 研究内容 1](#_Toc376030712)

[1.2.1 分析整理中医药学语言系统语义网络 1](#_Toc376030713)

[1.2.2 基于文本的语义关系发现 2](#_Toc376030714)

[1.2.3 语义网络归纳总结和比较分析 3](#_Toc376030715)

[1.3 研究方法和技术路线 3](#_Toc376030716)

[1.4 创新点 5](#_Toc376030717)

[1.5 预期研究成果 5](#_Toc376030718)

[2 术语与定义 1](#_Toc376030719)

[3 文献综述 2](#_Toc376030720)

[3.1 文本挖掘的概念及代表性工作 2](#_Toc376030721)

[3.2 文本挖掘在中医药领域的应用 6](#_Toc376030722)

[4 需求分析 9](#_Toc376030723)

[4.1 基于文献的语义关系发现 9](#_Toc376030724)

[4.2 语义网络归纳方法 9](#_Toc376030725)

[4.3 语义网络比较方法 10](#_Toc376030726)

[5 技术方案 11](#_Toc376030727)

[5.1 中医药语义网络归纳与浏览系统 11](#_Toc376030728)

[5.2 从万方知识服务平台获取文献资源 18](#_Toc376030729)

[5.2.1 万方数据知识服务平台的检索接口概况 19](#_Toc376030730)

[5.2.2 万方数据知识服务平台的检索接口样例 22](#_Toc376030731)

[5.2.3 基于PHP的万方数据知识服务平台访问程序 24](#_Toc376030732)

[5.3 基于TCMLS的中医药文献分词处理 27](#_Toc376030733)

[5.4 语义关系发现方法 31](#_Toc376030734)

[5.4.1 常见的语义关系发现方法 31](#_Toc376030735)

[5.4.2 基于TCMLS的语义关系发现方法 32](#_Toc376030736)

[5.4.3 Hamster：基于Java的文本语义关系发现程序 35](#_Toc376030737)

[5.4.4 文本挖掘实验过程及结果 38](#_Toc376030738)

[5.4.5 基于Web的文本语义关系检阅系统 39](#_Toc376030739)

[5.5 语义网络比较 43](#_Toc376030740)

[5.6 实现技术 46](#_Toc376030741)

[5.7 小结 47](#_Toc376030742)

[6 项目计划 48](#_Toc376030743)

[网络资源 49](#_Toc376030744)

[参考文献 50](#_Toc376030745)

**图目录**

图 1‑1 技术路线图 5

图 3‑1 BioPatentMiner 对“p53 is activated by amylase.”的标注结果 5

图 5‑1 语义网络展示界面 12

图 5‑2 语义网络展示界面：查看关系实例 13

图 5‑3 语义关系浏览界面（以“方剂 治疗 证候”为例） 13

图 5‑4 概念“风热证”的展示界面 14

图 5‑5 语义关系搜索界面 15

图 5‑6 基于分类体系的检索方式与基于顶层语义网络的检索方式的对比图 16

图 5‑7 语义关系搜索界面可帮助用户理解TCMLS语义网络的用法 16

图 5‑8 语义网络浏览系统可帮助术语学家发现术语系统中极不常用的类型 17

图 5‑9 语义网络浏览系统可帮助术语学家发现错误的语义关系 17

图 5‑10 语义关系搜索系统可帮助术语学家发现错误的语义关系（以“方剂 有…部分 中药”为例） 18

图 5‑11 语义关系搜索系统可帮助术语学家发现错误的语义关系（以“方剂 影响 中医疾病”为例） 18

图 5‑12 从万方数据知识服务平台获取文献的流程图 19

图 5‑13 万方数据知识服务平台检索结果样例：与“人参”相关的文献 23

图 5‑14 万方数据知识服务平台检索结果样例：以“袁隆平”为作者的文献 24

图 5‑15 将从“万方数据知识服务平台”获取的文献题录信息存入MySQL中 26

图 5‑16 中医药文献检索系统界面 27

图 5‑17 基于TCMLS的中医药文献分词处理 29

图 5‑18 文本语义关系发现过程示意图 35

图 5‑19 文本语义关系检索界面 39

图 5‑20 文本语义关系展示界面 40

图 5‑21 文本语义关系的相关文献界面 40

图 5‑22 文本语义关系的相关Baidu搜索结果界面 41

图 5‑23 文本语义关系在TCMLS中的相关陈述界面 41

图 5‑24 文本语义关系录入界面 42

图 5‑25 语义网络比较界面 42

**表目录**

表 5.1 源自文献的语义网络与TCMLS语义网络的比较报表 43

表 5.2 系统实现技术列表 46

表 6.1 项目开发的总体计划 48

# 项目概述

中医药学语言系统是中国中医科学院中医药信息研究所于2002年起开始研制的一项大型中医药术语系统，语义网络是语言系统的重要组成部分。为了促进中医药学语言系统的发展，提高系统中语义关系的准确性，本课题拟进行基于文本的中医药语义关系发现研究，以求通过对中医药文献的文本进行分析，并与中医药学语言系统现有语义网络结合，得到准确的语义关系。

研究主要内容包括：（1）对目前中医药学语言系统语义网络进行分析，发现问题，并归纳得到基于语义类型层的语义关系网；（2）对中医药文献文本进行分析，发现获得语义关系，并归纳为基于语义类型的语义网络；（3）对文本获得的语义网络与目前的语义关系网进行比较，优化，形成新的语义网络。

## 项目目标

（1）整理研究现有语义关系，发现冗余和问题等;

（2）获得基于文本的中医药语义关系，形成新的语义网络。

## 研究内容

### 分析整理中医药学语言系统语义网络

中医药学语言系统以中医药学科为主导且遵循中医药学科思路，旨在建立中国第一个计算机化的可持续发展的中医药学及与其学科相关的中国医药学检索语言集成系统，形成中医药学及其与该学科相关的语言系统平台。该系统以中医药学科为核心体系，以分类层次保持完整复杂性网状结构的中医药学科特点，按照中医药学科体系保留了完整的分类体系。同时，根据中医药语言学特点，设计了126种语义类型和58种语义关联关系。利用语言学与中医药学科结合关系，用结构标准化特点，充分完整表达了概念词间的关联与体系，为中医药学科数字化与规范化打下基础。

目前中医药学语言系统语义网络是由30多万个术语和58种语义关系组成的超过60万的语义关联而构成，其中30多万个术语归属于由126个语义类型。由于中医药学语言系统以前的加工规则和加工方式，现有的概念间的语义关联是由加工人员根据自己的专业知识来加工的，这就不可避免地造成了目前语义关联的严重不足和错误率高的问题。鉴于目前语义网络的状况，本研究先对现有的语义关联进行分析整理，修改错误关联，并**最终总结出以语义类型为基础的语义关联，形成准确完整的中医药学语言系统语义网络。**

### 基于文本的语义关系发现

**对中医药文献文本进行分析，获得实际的语义关系。**中医药学是经过几千年的发展而形成的，文献记载是其重要的知识流传的方式之一。从中医药学语言系统建设来源上来说，我们也是主要采用了各种文献中的知识。在系统建设的前期，由于缺乏相应中医药学文献文本挖掘方法，在对概念间的语义关联信息的抽取上，我们主要依赖于加工人员的个人知识和手工操作，但这种方法与加工人员的个人知识素养和责任心有很大的关系，并且会随着中医药学语言系统的不断发展而变得愈发困难。随着系统规模的不断扩大，加工者对内容的理解和掌握难度也在不断加大，这就进一步增加了手工加工的难度，从而制约着中医药学语言系统的进一步发展。

为了中医药学语言系统的进一步发展，本研究**拟选用近十年（2001——2011）的文献，对文献文本进行分析挖掘，对文本进行分句，分词，对每句中的语义关系进行向量标示，然后得到具体的概念间的关系；对每个概念进行语义归类，得到语义类型间的语义关系，形成语义类型间的语义网络。**

### 语义网络归纳总结和比较分析

**对文本获得的语义关系与目前的语义关系网进行比较，补充完善现有的语义关系网络。**中医药学语言系统语义关系是在UMLS(统一的医学语言系统)的54种语义关系的基础上，根据中医药学的学科内容，增加了4种语义关系而形成的。**根据上述研究内容总结出的以语义类型为基础的语义关系网和从文献文本中实际抽取的语义关系网，两者进行分析比较，补充完善现有的语义关系网络，形成能够指导中医药学语言系统实际发展和应用的语义网络。**

## 研究方法和技术路线

**语义本体方法**

本体(Ontology)一词来源于哲学领域，用来研究事物的特征以及它们之间的关系，在信息学领域中，它被用来描述一个由类型，属性以及关系组成的领域知识框架。语义本体(Semantic Ontology)是本体技术在语义Web中的应用，是其核心部分之一，其主要理念是用一套规范的表示方式来表达某一领域中的概念以及它们之间的相互关系。这些知识可以被用来分析领域的特征或者定义这一领域。

**文本挖掘方法**

文本挖掘旨在通过识别和检索令人感兴趣的模式，进而从数据源中抽取有用的信息。文本挖掘的数据源是文本集合，令人感兴趣的模式不是从形式化的数据库记录里发现，而是从非结构化的数据中发现。它的核心功能是分析文本集合中各个文本之间共同出现的模式。

**分类法**

分类法是指将类或组按照相互间的关系，组成系统化的结构，并体现为许多类目按照一定的原则和关系组织起来的体系表，作为分类工作的依据和工具。

**比较法**

比较法是通过观察，分析，找出研究对象的相同点和不同点，它是评价事物的一种基本方法。

**归纳法**

归纳论证是一种由个别到一般的论证方法。它通过许多个别的事例或分论点，然后归纳出它们所共有的特性，从而得出一个一般性的结论。归纳法是从个别性知识，引出一般性知识的推理，是由已知真的前提，引出可能真的结论。它把特性或关系归结到基于对特殊的代表的有限观察的类型；或公式表达基于对反复再现的现象的模式(pattern)的有限观察的[规律](http://baike.baidu.com/view/183124.htm)。

**德尔菲法**

专家会议预测法，是一种主观预测方法。它以书面形式背对背地分轮征求和汇总专家意见，通过中间人或协调员把第一轮预测过程中专家们各自提出的意见集中起来加以归纳后反馈给他们。

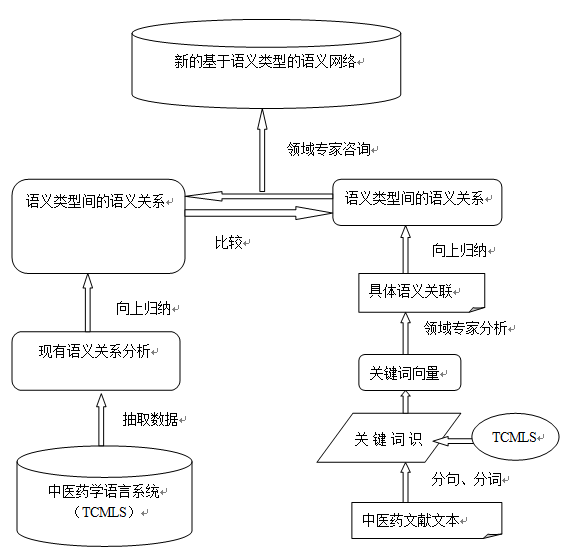


图 1‑1 技术路线图

## 创新点

（1）基于文本挖掘发现中医药语义关系的方法

（2）基于中医药语义类型间的语义网络的建立

## 预期研究成果

（1）建立“基于中医药语义类型的语义网络”

（2）在核心期刊发表论文2篇。

（2）撰写关于基于文本挖掘发现中医药语义关系的研究报告一份；

# 术语与定义

**D**

**顶层语义网络**

语义类型之间相互关联构成的语义网络。

**Y**

**语义关系（Semantic Relationship）**

概念之间具有明确含义的关联关系。

# 文献综述

在本章中，介绍了文本挖掘技术的基本概念和最新进展，讨论文本挖掘技术在中医药领域的应用现状和未来的发展方向。

## 文本挖掘的概念及代表性工作

随着各领域文本文献资源的迅速增长，对自动化的文本处理方法提出了越来越迫切的需求。文本挖掘（Text Mining）或文本中的知识发现（knowledge discovery in text），旨在利用数据挖掘、机器学习、统计、自然语言处理（Natural Language Processing, NLP）等技术，从海量文本（包括文学、电子邮件和文件等）中提取结构化信息，或发现新颖的知识（如科学假说）（Berry&Castellanos, 2007; Cohen & Hersh, 2005; Feldman&Sanger, 2006; Hearst, 1999; Simpson & Demner-Fushman, 2012）。

在过去的几十年里，生物医学领域取得了突飞猛进的发展，该领域已发表的文献量也得到了前所未有的增长。研究人员仅靠传统方法，已很难跟踪领域内最新的信息和知识。在海量文献和数据中蕴含的大量知识尚未得到有效的挖掘和利用。文本挖掘也因此成为生物医学的热点话题之一。

Cohen *et al*.(2005)，de Bruijn&Martin(2002)，Hirschman *et al*.（2002），Jensen *et al*. (2006)，Krallinger&Valencia(2005)，Shatkay&Feldman(2003)，Simpson *et al*. (2012)，Weeber *et al*. (2005)，Yandell&Majoros(2002)，Yetisgen-Yildiza&Pratt(2006)，Zweigenbaum *et al*.（2007）等已对生物医学文本挖掘相关的方法、应用程序和工具进行了深入评论。生物医学文本挖掘的核心任务包括：识别生物医学领域的命名实体（如基因、蛋白质、疾病和药物等）（Cakmak&Ozsoyoglu, 2008；Jimeno et al., 2008；Tsuruoka Y *et al*., 2008），发现生物医学实体之间的相互关系（Hristovski *et al*., 2005; Jenssen et al., 2001; Zhou&He, 2008）以及从生物医学文献中发现科学知识和假设（Ganiz et al., 2006）等。下面介绍两项代表性工作：

（1）Swanson Linking

<http://en.wikipedia.org/wiki/Swanson_linking#Swanson_linking>

美国芝加哥大学医学教授Don.R Swanon于上世纪80年代创立了一种新颖的情报学研究方法——Swanson Linking，这是一种基于文献的知识发现方法，用于揭示蕴含于公开文献中、尚未被人们认识的关联关系，提出科学假设，以供科学研究人员进一步证实（Swanson *et al*., 1986）。Don.R Swanon 提出了“未发现的公开知识（Undiscovered public knowledge）”这一概念（Swanson, 1986）。他认为，公开发表的文献中存在着许多有价值但未被发现的关联；从文献中挖掘这些隐含关联会对科学发展产生重要的推动作用。

Swanson的知识发现过程可分为两个步骤:形成假设和检验假设。我们将形成假设的过程称为开放式的非相关文献知识发现方法（open），它以感兴趣的主题C为初始点，通过中间集合B与文献集合A之间的关系，确定C与A之间的关联，其具体过程即C→B→A的过程，如图1所示。而检验假设的过程称为闭合式的非相关文献知识发现方法，即以A和C为出发点，研究人员需要努力寻找共同的中介词B，其知识发现过程的路径为C→B←A，如图2所示。A和C的联系越多，所做的假设越有价值。Swanson的第一个发现是食用鱼油可以治疗雷诺氏病，第二个发现是偏头痛和镁缺乏之间关系的研究。这两个发现都已被实验和临床证实（马明等, 2003；许建阳等，2003）。随后，Swanson还与他的合作者Smalheiser利用Arrowsmith系统进行另外5项研究：镁缺乏与神经系统疾病(1994年)、消炎痛与阿尔茨海默病(1996年)、雌激素与阿尔茨海默病(1996年)、游离钙磷脂酶A2与精神分裂症(1998年)、可作为生物武器的潜在病毒(2001年) （马明等, 2003）。Swanson还开发了的网络版知识发现工具Arrowsmith，该工具与PubMed实现了紧密结合，使知识发现过程变得更加容易、快捷和灵活。Arrowsmith的网址为http://kiwi.uchicago.edu和http://Arrowsmith.psych.uic.edu，可供科研工作者免费使用。Swanson *et al*. (1986,1991,1997）所做的开拓性研究，向我们揭示了从生物医学文献中发现新颖知识的巨大潜力。2000年，美国情报科学与技术学会(ASIST)授予D. R. Swanson该学会的最高成就奖。

（2）BioPatentMiner

Mukherjea et al. (2005) 研发了一个从生物医学专利中挖掘语义关系的系统——BioPatentMiner。该系统通过一个爬虫（Crawler）从USPTO网站下载用户关心的专利（它还通过其他方法收集了一些专利），对这些专利进行解析以提取发明人（Inventors）、授权人（Assignees）、标题（Title）、摘要（Abstract）。当前，该解析器（Parser）所能处理的专利为USPTO网站上的HTML文件，但也可以调整为其他格式。

该系统进一步通过BioAnnotator识别文献中出现的术语，并对它们进行归类。BioAnnotator使用生物医学词典（当前为UMLS）和一个规则引擎来实现其标注功能。UMLS 是一个跨语言、跨学科（化学、生物学等）的术语系统，用于表达医学术语及其之间的关系。UMLS的核心组件是一个语义网络（Semantic Network），它具有135个语义类型（如Gene or Genome and Amino Acid, Peptide, or Protein等），这些语义类型被通过54种语义关系（如prevents、causes等）连接起来。在UMLS中，每个生物医学概念都被赋予了若干个语义类型。一个概念可对应若干个词汇（即同义词），其中一个词汇为指定为正名（Primary name or baseform）。BioAnnotator在文献中识别出词汇之后，会找出这些词汇所对应的正名和语义类型，作为文献的语义标注。



图 3‑1 BioPatentMiner 对“p53 is activated by amylase.”的标注结果

从自由文本中挖掘实体之间的关系是一个非常困难的问题，因为这些关系有多种表达方式（Mukherjea et al., 2005）。为简化这个问题，Mukherjea et al.（2005）将挖掘的目标设定限定为仅挖掘某些特定类别的实体的几种特定关系（用户将这些限定条件写入一个XML配置文件）。

该系统采用RelationAnnotator对已标注的文本进行解析。它首先根据配置文件识别用户关心的实体，再尝试发现这些实体之间的关系。该方法通过模板（Template）来表达语义关系所对应的模式。一些通常的模式为：

1. 主体 动词组 客体 (例如：p53 activates amylase).
2. 客体 被动词组 主体 (例如：amylase is activated by p53).
3. 动名词 客体 主体 (例如：activation of amylase by p53).

当RelationAnnotator发现文本与某一模式相匹配时，就判定文本所含实体之间存在着对应的关系。在模式匹配的过程中，仅考虑句子中的领域实体和动词。标注后的专利被表示为如 所示的XML文件。

## 文本挖掘在中医药领域的应用

与生物医学领域的大量研究工作相比，文本挖掘在中医药领域的发展仍处于早期阶段。中医药文献（特别是中医药古文）的计算机处理技术尚存在很多困难。Zhou et al.（2010）对文本挖掘及其在中医药领域的应用进行了综述，讨论了相关的研究问题和未来的方向。下面我们介绍一些有代表性的研究。

Wu *et al*. (2004)开展了一项中医药文本挖掘的开创性工作。文本挖掘被用于从MEDLINE等文献库中发现中医证候、疾病和基因之间的关系，从而构建功能性基因网络。该研究所基于的假设是：与同一证候相关联的一组基因应具有某种程度的相互作用。因此，同一证候所关联的基因可以构成一个功能性基因网络，该网络对应某种生物学功能。据此，他们从中医药文献中发现了证候与疾病之间的关系，从MEDLINE中发现了疾病与基因之间的关系，再已疾病为中介建立证候与基因之间的关系。“证候-基因”关系被用于生成基于证候的基因网络，它帮助我们从中医药的角度分析基因的功能。

他们开发了文本挖掘系统MeDisco/3S，以完成这项研究。MeDisco/3S是一个综合性的数据挖掘系统，旨在从MEDLINE和中医书目文献中揭示基因之间的功能性关系（Peng&Zhang, 2007; Wu et al., 2004; Zhou et al., 2007）。在挖掘过程中，他们采用一种简单、有效的Bootstrapping法，从TCMLARS所含的文献中抽取证候和基因之间的关系；并采用共词分析（co-occurrence）从MEDLINE的文摘和题目中识别疾病和基因之间的关系；最后，以疾病为中介，建立证候与基因之间的关系。挖掘得到了约200,000条证候-基因关系，用于生成基于证候的基因网络。MeDisco/3S还将数据挖掘与OLAP相结合，获得了良好的分析效果。例如，该项研究发现了一组与肾阳虚（Kidney-Yang Deficiency syndrome，KYD syndrome）相关的基因，如CRH（促肾上腺皮质激素）、PTH（甲状旁腺激素）、PRL（催乳素）、BRCA1基因（乳腺癌1，发病初期）和BRCA2（乳腺癌2，发病初期）等，从而构建了一个与肾阳虚对应的基因网络。通过分析这些基因的功能，发现与这个证候相关的基因，它们彼此之间具有功能性的关联关系，并聚集成一个功能性的网络模块。 jenssen等。

Cao et al.（2004）提出了一种基于本体从半结构化文本（Semi-structured Text）中抽取中药和方剂知识的方法。该方法被用于构建“国家知识基础设施（Cao et al., 2002）”中的中医药部分。他们根据课本、抄本、百科全书、字典等7个知识来源，开发了中药和方剂等2个领域本体。他们基于这些本体，使用一种可执行的知识抽取语言（EKEL），从中药和方剂文本中获取知识。该系统已在一系列中药和方剂的文献库上得到了测试，产生了一个包含2710种草药和5900中方剂的知识库。

另外，Zhou *et al*. (2005)研制了MeDisco/3T系统，旨在从中文书目数据库中自动提取方剂知识。该系统以少量方剂名称作为种子，反复提取新的中医药名词和模式；其提取方剂名称的准确率超过95％。他们进一步利用所提取的方剂名称，通过启发式规则从半结构化的文献摘要中提取的方剂药物成分的信息。最后，他们通过关联规则挖掘算法，即Apriori算法（Agrawal&Srikant, 1994），从方剂组成信息中挖掘方剂配伍规律。

Li *et al*. (2007)提出构建一个被称为“表型网络-生物网络-草药网络”的多层次网络，从而整合中医药和系统生物学的知识，为中医理论和事件提供生物学解释[Li et al. 2007]。他们在神经-内分泌-免疫（neuro-endocrine-immune，NEI）网络的背景下，研究热证和寒证的生物学实质。该项研究综合使用了文本挖掘、网络分析和拓扑比较等方法，对与热证和寒证相关的NEI网络进行了系统分析。研究发现，激素在寒证网络中占主导地位，免疫因素在热证网络中占主导地位，且这两个网络通过神经递质相互连接。另外，治疗寒证的中药一般会影响寒证网络的“枢纽节点”；治疗热证的中药一般会影响热证网络的“枢纽节点”。这项研究表明中医证候与NEI系统具有密切的关系。

Fang et al.（2008）通过生物医学文本挖掘的方法，构建了数据库TCMGeneDIT，其中记录了848种草药的知识（如化学成分、疗效等），以及疾病、基因、草药之间的关联关系。从公共数据库（如HPRD，KEGG）中获得的“蛋白质-蛋白质相互作用”和“生物路径”也被用来探索与中药疗效相关的基因作用。从中药名称，基因，Mesh疾病，中药成分和效果等方面对文献进行了标注。经过标注的文献语料又被用于发现各种关联关系，包括（草药，基因），（草药，疾病），（草药，基因，疾病），（草药，成分），（草药，疗效）和（基因，成分）。基于Swanson的闭合式发现模型，推断传递性关联关系。一个基于Web的搜索系统，帮助用户搜索相关的关系和网络（http://tcm.lifescience.ntu.edu.tw/）。 TCMGeneDIT能帮助我们了解中药化学成分的作用以及中药的作用机理。

陶金火等（2011）提出了一种利用中医药学语言系统，从大量中医药文献中发现语义关系的方法。该工作基于如下的基本假设：经常在同一句话中出现的两个词汇或许存在某种语义关系。该方法首先基于词典对目标文献以句子为单位进行关键词的抽取（得到的关键词采用向量模式进行存储，关键词保持在句子中的出现的先后顺序），再计算高频关键词组，最后根据本体知识库来识别高频关键词组之间的语义关系。陶金火对多种不同类型的文献（其中还包括中医古文）进行了分析，从中发现中药及其治法的知识。实验结果证实该方法可以对中医药文献进行语义关系发现。

此外，周雪忠等（2003，2004）提出了一种基于文本挖掘的中医学文献主题自动标引方法。付志宏(2010)、杨克特（2010）等开发了面向中医药领域的语义搜索引擎DartSearch，该系统提供丰富的内容和强大的搜索功能，有效满足Web用户的知识检索需求，促进虚拟组织内的知识共享。

# 需求分析

**拟基于TCMLS从文本中抽取语义关系，归纳总结为一个基于语义类型的语义网络，并将其与TCMLS的语义网络进行对比,从而构建一个完整的语义网络。**

## 基于文献的语义关系发现

该软件应能处理如下形式的数字文献：（1）存储于常用的关系型数据库（如Access）中的文本字段；（2）文字内容可以正常解析的PDF文件（不支持扫描版的PDF文件）；（3）纯文本文件。

该软件依赖于“中医药学语言系统”中关于中医药概念/术语的信息；通过“中医药学语言系统”的结构性数据导出文件来访问这些信息。

该软件应能对中医药文献的全文或摘要进行解析，以合理的效率和准确率生成中医药领域具体概念（或术语）之间的语义关系，其中：（1）语义关系的形式为（主体、谓词、客体），其中主体和客体来自“中医药学语言系统”中的概念及术语，谓词是由原文的含义决定的；（2）所生成的语义关系与原文内容具有相关性；（3）所生成的语义关系实质上是假设性的，有待领域专家加以验证。

该软件应将所发现的语义关系以可读文本文件的形式输出，其中包括可由Excel软件打开和处理的文件形式 (Excel形式为主——谓——宾形式) 。

## 语义网络归纳方法

该软件将概念间的语义关系，归纳总结为语义类型之间的关系。所得到的语义网络通过语义关系将语义类型连接起来。每个概念都属于一个语义类型，将每个概念间的语义关系，都总结为语义类型之间的关系（例如：大黄的语义类型是中药，腹痛的语义类型是疾病，从类似于“大黄 治疗 腹痛”的一系列语义关系中，可以总结出“中药 治疗 疾病”）。

语义关系的形式为（语义类型、语义关系名称、语义类型、实例数量、...），其中：（1）语义类型（如“中药”、“疾病”）来自“中医药学语言系统”；（2）语义关系名称（如“治疗”）可以来自“中医药学语言系统”，也可以是新颖的；（3）“中药 治疗 疾病”的实例数量是指“大黄 治疗 腹痛”这类具体的语义关系的数量。

该软件应将所发现的语义关系以可读文本文件的形式输出，其中包括可由Excel软件打开和处理的文件形式 (Excel形式为主——谓——宾形式) 。这些语义关系可被融合成一个基于语义类型的语义网络。

## 语义网络比较方法

将TCMLS的语义关系归纳为一个语义网络（简称A）；将从文本中挖掘的语义关系归纳为一个语义网络（简称B）。对A和B进行比较分析，并将它们综合起来构建一个准确、完整的TCMLS语义网络框架，以指导中医药学语言系统实际发展和应用。

该软件应找出在B而不在A中的语义关系，将其展示出来供专家分析。专家可将一些语义关系加入B中，从而对TCMLS现有的语义网络进行补充。

该软件应通过Web界面，将语义网络以表格和图形的方式展示出来，支持用户比较它们之间的异同。

# 技术方案

在本项目中，开发了一套基于TCMLS的语义关系发现技术，包括分词、文本关系发现、语义网络归纳和语义网络比较等功能。

1. 开发语义网络归纳系统，对TCMLS中的语义网络进行归纳，支持用户对其进行分析和修改。
2. 以TCMLS中的词汇作为关键词，从“万方数据知识服务平台”中获取中医药文献。
3. 从TCMLS中导出中医药领域词库，基于词库对“文献库”中的文本进行分词处理。
4. 开发语义关系发现系统，支持文本语义关系的发现以及用户对关系进行分析和标注。
5. 使用语义网络归纳系统，对文本语义关系进行归纳，构成基于文本的语义网络。
6. 开发“语义网络比较界面”，对TCMLS的语义网络与基于文本的语义网络进行比较。

下面分别进行介绍。

## 中医药语义网络归纳与浏览系统

顶层语义网络对于TCMLS的规范化加工具有重要意义，但我所目前尚缺乏从TCMLS的100多万条语义关系中归纳出“顶层语义网络”的系统。在本项目中，我们实现了一套中医药语义网络归纳与浏览系统，它能够从大量的中医药语义关系中归纳出一个“顶层语义网络”，并以Web的方式实现语义网络的浏览和展示功能。

如 所示，系统以Web的方式实现语义网络的浏览和展示功能。在界面左侧，统计每个语义类型的实例个数，并按实例数量大小顺序对语义类型进行排列。当用户选择某一类型时，系统会在右侧显示该类型涉及的语义关系。

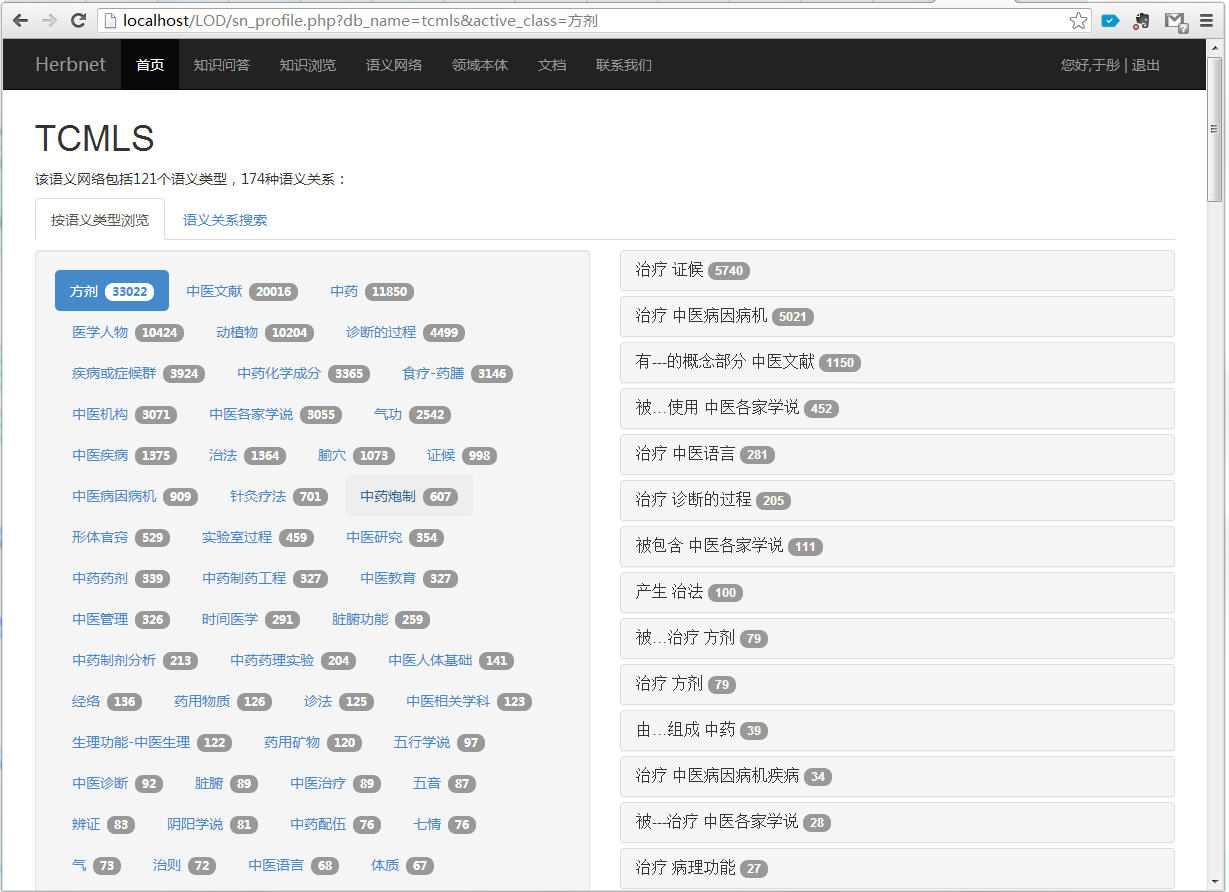


图 5‑1 语义网络展示界面

如 所示，当用户选择某一语义关系时，系统会显示该语义关系的一些实例。例如，当用户选择“方剂 治疗 证候”时，系统会给出“银花解毒汤 治疗 风热证”、“红轮散 治疗 热证”、“三光散 治疗 风淫证候”、“三黄丸 治疗 便血”、“红豆散 治疗 湿淫证候”等。

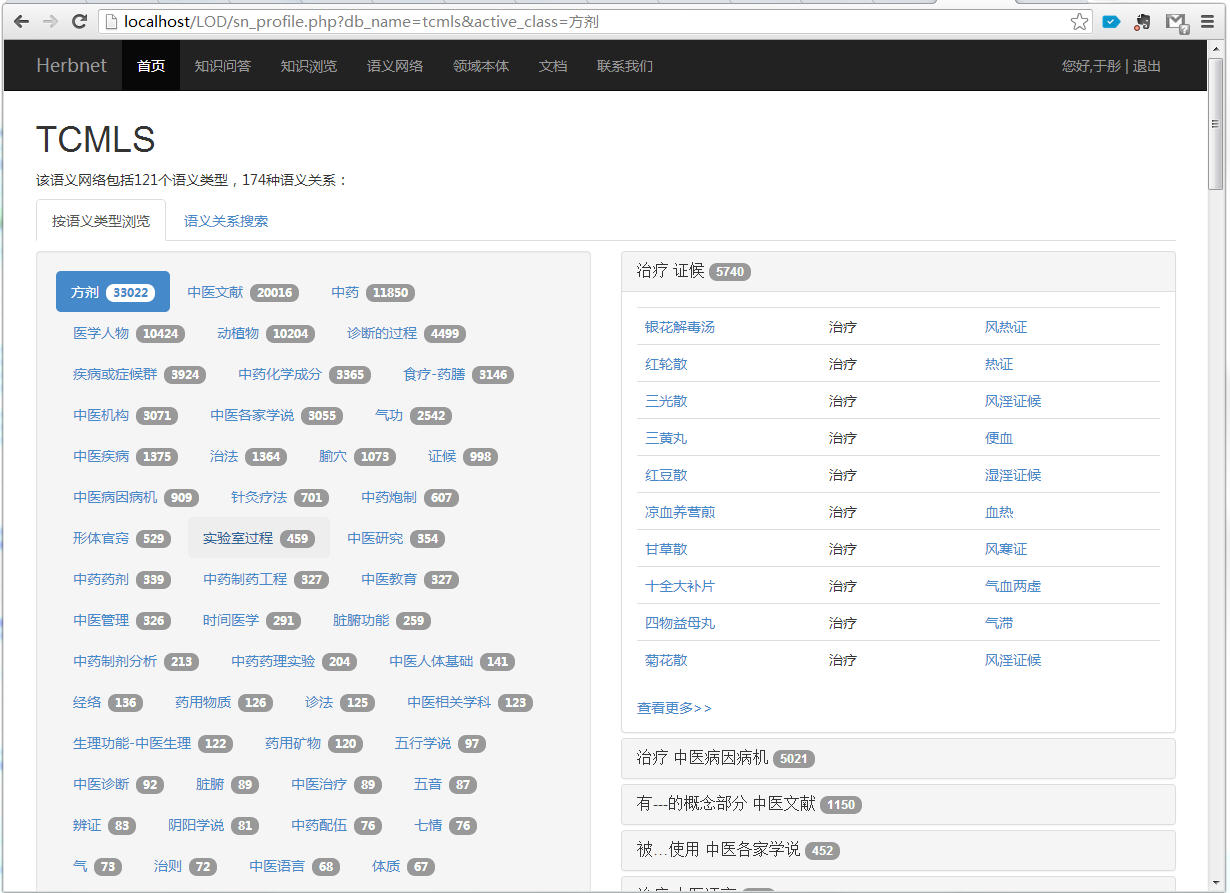


图 5‑2 语义网络展示界面：查看关系实例

若用户希望查看更多的实例，可点击“查看更多>>”，则系统将转到如 所示的语义关系搜索界面（以“方剂 治疗 证候”为例展示）。

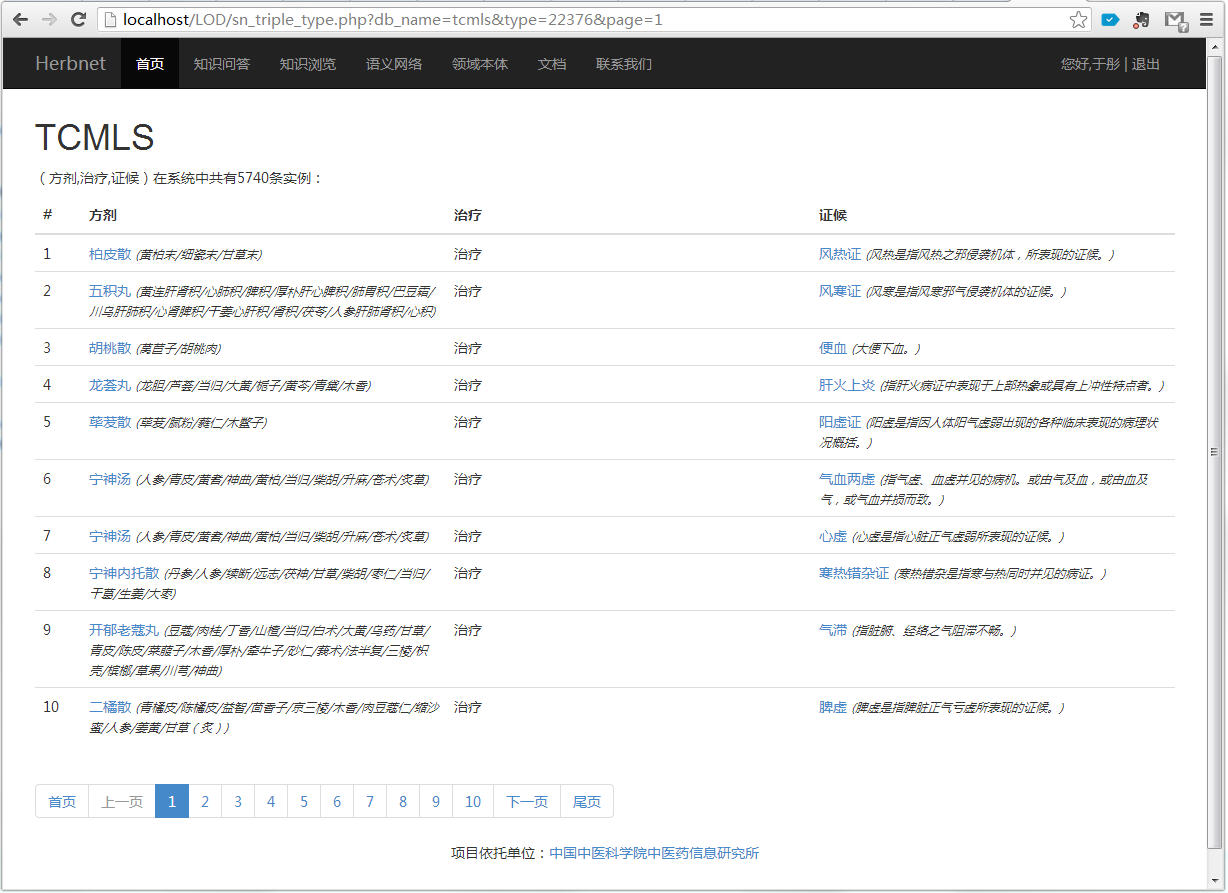


图 5‑3 语义关系浏览界面（以“方剂 治疗 证候”为例）

在语义关系浏览界面中，每个实体都带有超链接。用户可以若用户希望查看更多的实例，可点击“查看更多>>”，则系统将转到如 所示的概念信息展示界面（以“风热证”为例展示）。



图 5‑4 概念“风热证”的展示界面

如 所示，系统还提供了另一种语义网络展示工具，被称为“语义关系搜索”。用户可以任选主体、谓词和客体的类型，系统则将符合条件的语义关系分门别类地展示出来。例如，用户若想了解“哪些类型的事物可以治疗证候”，则可选择主体为“任意事物”，谓词为“治疗”，客体为“证候”，系统会显示一个列表:“方剂 治疗 证候”、“治法 治疗 证候”、“食疗-药膳 治疗 证候”等等。对于每条顶层语义关系，系统都给出了一些实例供用户参考。例如，“方剂 治疗 证候”的例子包括“银花解毒汤 治疗 风热证”,“红轮散 治疗 热证”,“三光散 治疗 风淫证候”等等。用户若想查看更多的实例，可点击“查看更多>>”，则系统将转到如 所示的语义关系搜索界面（以“方剂 治疗 证候”为例展示）。

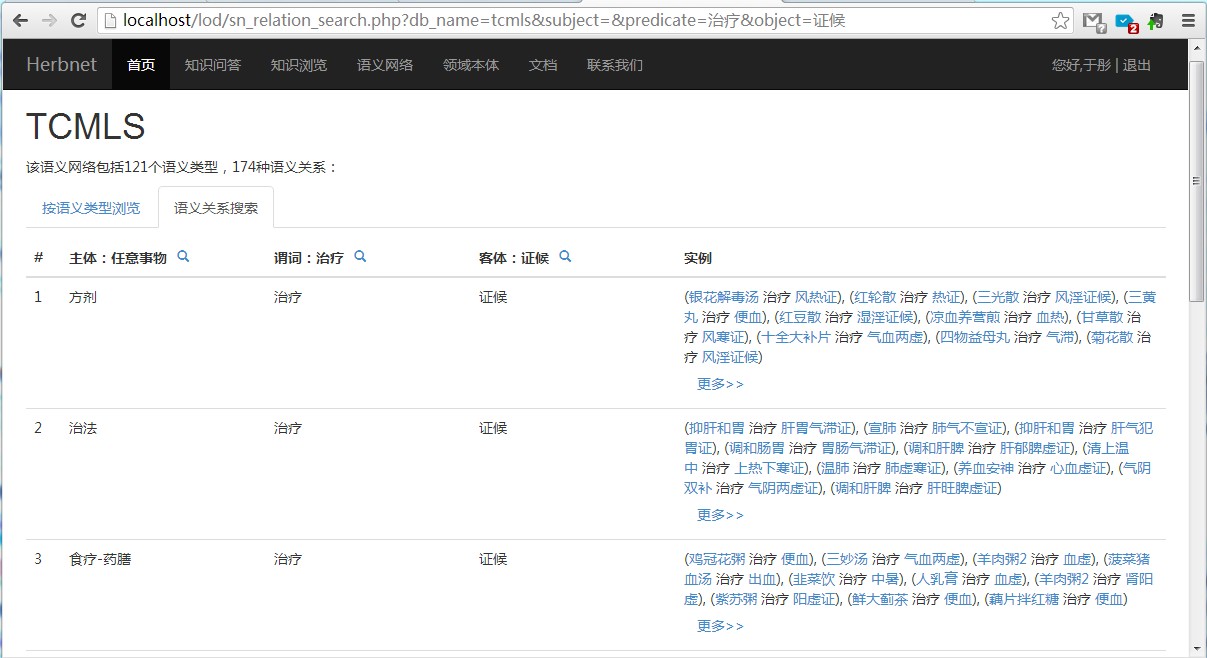


图 5‑5 语义关系搜索界面

该系统能够从大量语义关系中归纳出“顶层语义网络”，并支持Web用户对归纳出的“顶层语义网络”进行检索和浏览。它还支持用户以“类型间语义网络”为框架，浏览实例间语义关系，从而丰富了TCMLS等复杂术语系统的检索方式。

如 所示，我们传统上主要是基于分类体系对TCMLS等术语系统进行检索。用户通过概念的层次结构来定位某个概念，再以这个概念为入口来浏览语义网络。在本项目中，我们实现了基于顶层语义网络的检索方式。例如，用户可以检索“哪些事物 治疗 证候”，系统会列出“针灸 治疗 证候”、“治法 治疗 证候”、“中药 治疗 证候”、“药膳 治疗 证候”等顶层语义关系。用户可选择某条顶层语义关系（如“方剂 治疗 证候”），系统会列出它的实例（如“四君子汤 治疗 气虚证”）。

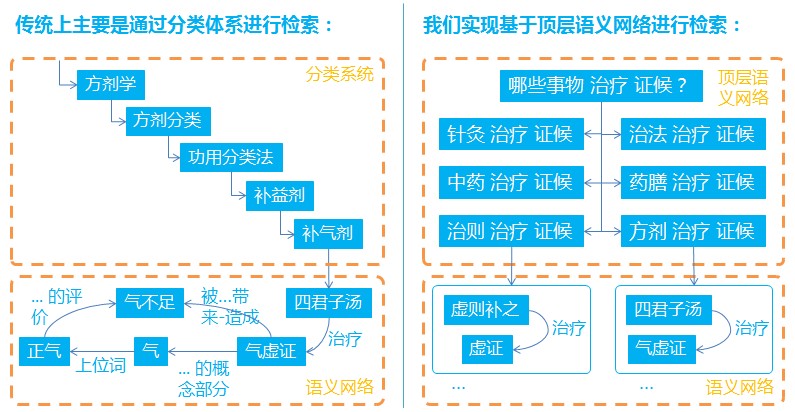


图 5‑6 基于分类体系的检索方式与基于顶层语义网络的检索方式的对比图

系统可帮助用户理解TCMLS语义网络的用法。如 所示，用户可查到“方剂”和“中药”之间主要有“由...组成”和“有...的概念部分”这两种语义关系，并通过实例来理解它们的用法。这有助于语义网络加工的规范化。



图 5‑7 语义关系搜索界面可帮助用户理解TCMLS语义网络的用法

该系统还可帮助用户检出TCMLS中的不规范之处。如 所示，系统可以检出一些极少使用的语义类型，或可在TCMLS“语义网络”的新版本中去除。



图 5‑8 语义网络浏览系统可帮助术语学家发现术语系统中极不常用的类型

系统还可帮助术语学家发现术语系统中错误的语义关系。如所示，“方剂 执行 中医病因病机”仅有一个实例：“人森养荣汤 执行 瘟疫”，这或许是个错误用法。



图 5‑9 语义网络浏览系统可帮助术语学家发现错误的语义关系

另外，术语学家也可用语义关系搜索系统来发现错误的语义关系。如 所示，用户通过该系统可以看出，TCMLS中一般通过“由…组成”来表达“方剂”与“中药”之间的组成关系，而极少用“有...部分”关系将“方剂”与“中药”相连。用户可以据此将“有...部分”改为“由...组成”。



图 5‑10 语义关系搜索系统可帮助术语学家发现错误的语义关系（以“方剂 有…部分 中药”为例）

又如 所示，“方剂 影响 中医疾病”极少出现，应被替换为“方剂 治疗 中医疾病”。



图 5‑11 语义关系搜索系统可帮助术语学家发现错误的语义关系（以“方剂 影响 中医疾病”为例）

## 从万方知识服务平台获取文献资源

采用文本挖掘技术所构建的语义网络，是从海量文献中涌现出的一种语义学模式。简单地说，文献量越大、越全面，所得到的语义网络就越具有统计价值。文献的数量和相关程度是影响挖掘结果价值的重要参数。鉴于此，我们希望能够获取尽量全面的文献资源作为文本挖掘的基础。

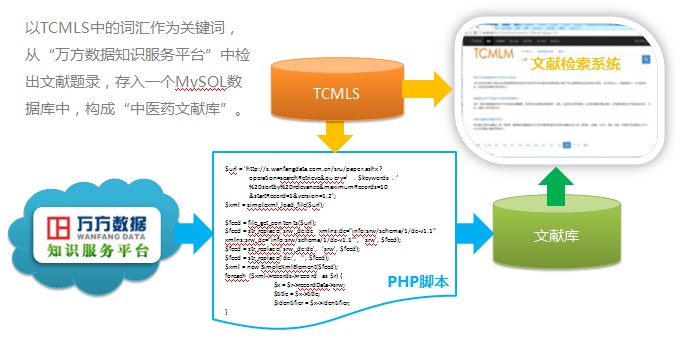


图 5‑12 从万方数据知识服务平台获取文献的流程图

经调研发现，我们有2个主要的技术选择。其一，是使用我所已经建成的文献库；其二，是用TCMLS术语作为检索词，从万方等我所有权访问的文献库中获取文献。贾李蓉和胡雪琴等都使用过我所文献库作为文献来源。如 所示，在本次研究中，我们尝试采用第二种方案来获取中医药文献资源，探索构建中医药领域相关文献库的新手段，为今后从CNKI 、超星等网络资源中获取文献积累经验。另外，从网络资源中抓取的文献库，也可用于丰富我所的文献库。

### 万方数据知识服务平台的检索接口概况

万方数据知识服务平台对外提供了检索接口，基本符合SRU/SRU标准（<http://www.loc.gov/standards/sru/>）。下面简要介绍该服务接口的概况[[1]](#footnote-1)：

服务接口地址：

* <http://s.wanfangdata.com.cn/srw/paper.asmx>
* <http://s.wanfangdata.com.cn/sru/paper.ashx>

服务接口描述文（WSDL）：

* <http://s.wanfangdata.com.cn/srw/paper.asmx?wsdl>

接口参数说明：

|  |  |
| --- | --- |
| 参数名称 | 参数说明 |
| Operation | 操作类型。取值：searchRetrieve（固定值） |
| Version | 接口版本。取值：1.2（固定值） |
| Query | 检索表达式。取值：用户检索请求对应的表达式，CQL格式。 |
| startRecord | 检索结果起始记录。取值：正整数 |
| maximumRecords | 检索结果最大记录数。取值：正整数。 |
| sortKeys | 排序字段，此参数不使用。具体的排序请求直接追加到query参数的CQL表达式中。 |

检索字段说明：

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名称 | 字段含义 | 检索方式 | 示例或说明 |
| ArticleID | 论文ID | exact | ArticleID exact Periodical\_lssfzkxxxb201001007 |
| PeriodicalID | 期刊ID | exact | PeriodicalID exact ahdxxb |
| Title | 标题 | All,= | Title = "抗弯强度" |
| Creator | 责任者 | exact | Creator exact "袁隆平" |
| Keywords | 关键词 | all,= | Kewords = "文献检索" |
| Abstract | 摘要 | all,= | Abstract = "文献检索" |
| Date | 日期 | >=、<=、within |  |
| CitedCount | 被引次数 | >、>= |  |
| DBID | 数据库标识 | = | 取值：   * WF\_XW ： 学位论文 * WF\_QK ： 期刊论文 * WF\_HY ： 会议论文 |

排序字段说明：

知识服务平台支持对检索结果按进行多项指标的加权排序。

各排序字段如下：

|  |  |
| --- | --- |
| 字段名称 | 字段含义 |
| CitedCount | 被引次数 |
| Date | 论文出版时间 |
| Relevance | 相关度 |
| CoreRank | 文献核心刊 |

排序示例：

* sort by relevance/weight=5 CitedCount Date/weight=2 CoreRank

样例：

* <http://s.wanfangdata.com.cn/sru/paper.ashx?operation=searchRetrieve&query=hnc%20sortby%20relevance&maximumRecords=10&startRecord=1&version=1.2>

签约主站通用登陆接口：

* http://login.wanfangdata.com.cn/login.aspx?userid=xxx&password=xxx

根据该接口声明，可采用PHP，Java等各种主流的编程语言来实现万方文献库的文献抓取程序，下面介绍通过PHP实现抓取万方文献库中文献的技术。

### 万方数据知识服务平台的检索接口样例

下面举几个万方数据知识服务平台的查询实例：

（1）以“人参”为关键词对万方数据知识服务平台进行检索的字符串：

|  |
| --- |
| [http://s.wanfangdata.com.cn/sru/paper.ashx?operation=searchRetrieve&query= "人参" sortby%20relevance&maximumRecords=10&startRecord=1&version=1.2](http://s.wanfangdata.com.cn/sru/paper.ashx?operation=searchRetrieve&query=Creator%20exact%20%22%E8%A2%81%E9%9A%86%E5%B9%B3%22sortby%20relevance&maximumRecords=10&startRecord=1&version=1.2) |

如图 5‑13 所示，万方返回的检索结果为一个XML文件，其中包括“人参”相关文献题录信息，该文件符合DC规范。

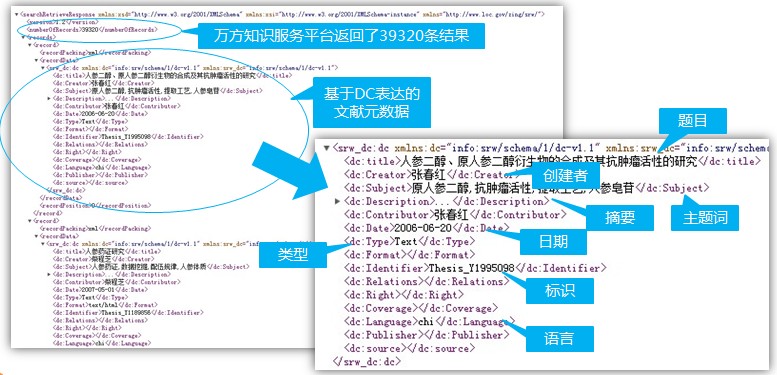


图 5‑13 万方数据知识服务平台检索结果样例：与“人参”相关的文献

（2）检索以“袁隆平”为作者的文献：

|  |
| --- |
| [http://s.wanfangdata.com.cn/sru/paper.ashx?operation=searchRetrieve&query= Creator exact "袁隆平" sortby%20relevance&maximumRecords=10&startRecord=1&version=1.2](http://s.wanfangdata.com.cn/sru/paper.ashx?operation=searchRetrieve&query=Creator%20exact%20%22%E8%A2%81%E9%9A%86%E5%B9%B3%22sortby%20relevance&maximumRecords=10&startRecord=1&version=1.2) |

如图 5‑14 所示，万方返回的检索结果为一个XML文件，其中包括“袁隆平”所作的文献题录信息。



图 5‑14 万方数据知识服务平台检索结果样例：以“袁隆平”为作者的文献

（3）检索同时包括“人参”、“大黄”的文献：

|  |
| --- |
| [http://s.wanfangdata.com.cn/sru/paper.ashx?operation=searchRetrieve&query= "人参" and "大黄"sortby%20relevance&maximumRecords=10&startRecord=1&version=1.2](http://s.wanfangdata.com.cn/sru/paper.ashx?operation=searchRetrieve&query=%20%22人参%22%20and%20%22大黄%22sortby%20relevance&maximumRecords=10&startRecord=1&version=1.2) |

### 基于PHP的万方数据知识服务平台访问程序

如上文所述，我们希望以TCMLS中的词汇作为关键词，从“万方数据知识服务平台”中获取中医药文献。下面给出对万方数据知识服务平台进行访问的核心PHP脚本，它对“万方数据知识服务平台”进行关键词检索，并将检索内容显示出来。

|  |
| --- |
| //连接“万方数据知识服务平台”的URL字符串，其中变量$keywords存储用于检索的关键词。  $url = 'http://s.wanfangdata.com.cn/sru/paper.ashx?operation=searchRetrieve&query=' . $keywords . '%20sortby%20relevance&maximumRecords=10&startRecord=1&version=1.2';  $feed = file\_get\_contents($url); //从“万方数据知识服务平台”获取查询结果文件  $feed = str\_replace('srw\_dc:dc xmlns:dc="info:srw/schema/1/dc-v1.1" xmlns:srw\_dc="info:srw/schema/1/dc-v1.1"', 'srw', $feed);  $feed = str\_replace('srw\_dc:dc', 'srw', $feed);  $feed = str\_replace('dc:', '', $feed);  $xml = new SimpleXmlElement($feed); //从查询结果文件中读出一个XML模型。  //对XML结果集中的每条记录进行遍历访问：  foreach ($xml->records->record as $r) {  $x = $r->recordData->srw;//$x中记录了当前正在处理的一条文献题录。  $title = $x->title;//读取一条文献的题目。  $Identifier = $x->Identifier;//读取一条文献的标识。  //将文献题录内容显示出来。  echo '<h4><a href="http://d.wanfangdata.com.cn/' . $Identifier . '.aspx">' .  $title . '</a></h4>';  echo '<p>' . $x->Creator . '等发表于' . $x->Date . '</p>';  echo "<p>" . $x->Description . "</p>";  } |

将从“万方数据知识服务平台”中检出的20余万篇文献题录存入一个MySQL数据库中，构成如所示的“中医药文献库”。

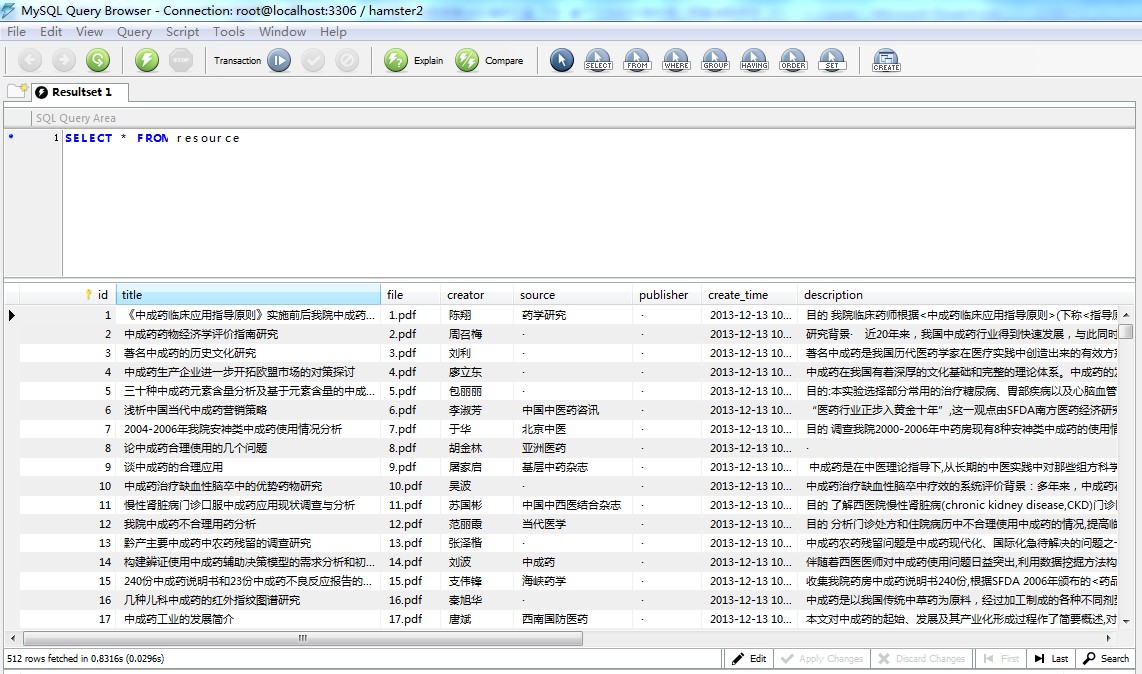


图 5‑15 将从“万方数据知识服务平台”获取的文献题录信息存入MySQL中

如 所示，在“中医药文献库”的基础上，实现了基于Web的文献检索系统，支持基于关键词的文献检索功能。



图 5‑16 中医药文献检索系统界面

这样一个从万方数据知识服务平台中获取的中医药文献库，可以支持文本关系发现，文献标引等文本挖掘实验和文献检索应用，为后续工作奠定了基础。

## 基于TCMLS的中医药文献分词处理

分词是中文处理的一个特殊问题，且往往是其中的第一步。对文献进行准确分词是后续算法良好运行的基础。当前，自动化中文分词算法主要分为3类：（1）基于理解的分词方法，（2）基于字符串匹配的分词方法，以及（3）基于统计的分词方法。

基于理解的分词方法利用句子的语义和句法进行分词。这种方法需要对语言学知识具有充分的理解。汉语的语法比较复杂，表达方式灵活多变，因此这种方法的实现难度很大，尚有待完善。

基于字符串匹配的分词方法，即基于词典的分词方法。该方法需要有一个全面的词典库。此类方法的优点是准确率较高。其主要缺点是分词效果受制于词典的规模，且发现新词的能力较弱。MMSeg4J是用Chih-Hao Tsai的MMSeg算法实现的中文分词程序。

基于统计的方法是一类较为新颖的方法。它不需要字典库和语法知识，仅通过海量文献中呈现出来的统计规律来进行分词。其主要思想是文本中多次同时出现的字是词语的概率比较大。这种方法分词效率较高，无需庞大的知识库，还可以识别出新词。但目前受限于文献规模，这种方法的识别率较低，还有很长的路要走。

在这三种方法中，基于词典的分词方法较为适合中医药领域的需求。中医药领域的概念、术语和表达方式都具有显著的领域特色，因此基于语言学的方法处理中医药文献的效果并不理想。中医药文献的表达方式灵活多变，且总量相对较少，因此也不适合采用基于统计的方法。相比之下，由于中医药领域已经出现了TCMLS等较为完备的术语系统，因此基于词典的分词方法可在该领域取得相对较好的效果。

基于词典的分词方法也与本项目的特定需求相吻合。在本项目的后续处理中，需要获得关键词的语义信息（如定义、别名、上位词、下位词、相关关系等），因此分词算法所识别出的关键词应预先存在于语言系统之中。鉴于此，本项目采用TCMLS作为词库，对中医药文献进行了分词处理。所提取的关键词属于中医药领域，为后续处理提供了噪音较少的数据源，提高了语义信息的质量，减少了无关信息。

下面对分词的方法进行具体解释。如 所示，分词处理就是将“... 人参有‘补五脏、安精神、定魂魄、止惊悸、除邪气、明目开心益智’的功效...”的文本，拆分为类似于“人参 | 有 | 补 | 五脏 | 安 | 精神 | 定 | 魂 | 魄 | 止 | 惊悸 | 除 | 邪气 | 明 | 目 | 开 | 心 | 益 | 智 | 的 | 功 | 效 |”的序列。在分词处理中，用到了两个词库：其一是中医药领域专业词库，来源于TCMLS；其二是从中文动词中筛选出的谓词词表，用于推测语义关系的类型。



图 5‑17 基于TCMLS的中医药文献分词处理

分词算法所用的“词库”就是一个如下表所示的词汇列表，存于一个文本文件中。

|  |
| --- |
| ...  国肠汤  遗尿灵汤  地血散  走马牙疳  大瘕泄  葛根石膏汤  干葛石膏汤  柴胡加芒消汤  朱黄丹  脂草饮  凉血散  破合汤  引血归经  ... |

分词程序是基于Java语言编写的，实现了“正向最大匹配”的分词方式。可通过Complex类的segWords方法来调用这个分词程序，下面是segWords方法的代码：

|  |
| --- |
| /\*\*  \* 对一段文字进行分词的代码。  \* **@param** input 输入的一段文字，作为分词的对象  \* **@param** wordSpilt 词汇之间的分隔符  \* **@return 返回分出的词汇序列**  \* **@throws** IOException  \*/  **public** String segWords(Reader input, String wordSpilt) **throws** IOException {  //sb用于缓存分词的结果  StringBuilder sb = **new** StringBuilder();  //Seg为分词器  Seg seg = getSeg();  //MMSeg利用Seg对input中的一段文字进行分词，提供分词结果的迭代访问  MMSeg mmSeg = **new** MMSeg(input, seg);  Chunk chunk = **null**;  **boolean** first = **true**;  //从MMSeg对象中读出词汇序列，存入sb中  **while**((chunk=mmSeg.next())!=**null**) {  **for**(**int** i=0; i<chunk.getCount(); i++) {  Word word = chunk.getWords()[i];  **if**(!first) {  sb.append(wordSpilt);  }  String w = word.getString();  sb.append(w);  first = **false**;  }  }  //返回词汇序列  **return** sb.toString();  } |

如下是调用Complex类的segWords方法的一个例子：

|  |
| --- |
| String txt = "人参有“补五脏、安精神、定魂魄、止惊悸、除邪气、明目开心益智”的功效，“久服轻身延年”。";  Complex complex = **new** Complex();  String wordSequence = complex.segWords(txt, " | ");  System.out.println(wordSequence); |

分词的结果为一个词汇向量，其中的词汇保持其在原句中出现的顺序：

|  |
| --- |
| 人参 | 有 | 补 | 五脏 | 安 | 精神 | 定 | 魂 | 魄 | 止 | 惊悸 | 除 | 邪气 | 明 | 目 | 开 | 心 | 益 | 智 | 的 | 功 | 效 | 久 | 服 | 轻 | 身 | 延 | 年 |

其中，“人参”、“五脏”、“精神”、“惊悸”、“邪气”为来源于TCMLS中的词汇，因此被分词算法识别出来；而其它词汇则未被识别出来，而被拆分为单字处理。

## 语义关系发现方法

目前，实用的知识库系统显现出大型化的趋势，往往包含百万条语义关系。若由人工编辑这些关系，则会是一个耗时费力的大工程。若能实现从文献中自动抽取语义关系的方法，则可大幅提升知识获取的效率。因此，语义关系发现方法对领域知识库的构建具有很大意义。

### 常见的语义关系发现方法

从自由文本中挖掘实体之间的关系是一个非常困难的问题，因为同一种关系在文本中会有多种表达方式（Mukherjea et al., 2005）。常见的语义关系发现方法分为如下2类：

第一、基于语法分析的语义关系发现

学者们在自然语言处理（NLP）技术的基础上，提出了基于语法分析的语义关系发现。其主要思路是：通过对文本进行语法分析，构建出语法树，再通过语法和词性的分析得到其中的语义关系。例如，Jiang et al.（2007）提出了一种基于语法分析和关联规则挖掘的模式挖掘方法；他们使用语法分析来构建语法树，然后通过对语法树的结构分析来得到语义关系。这类方法的优点是对语义关系定位比较准确，并可以通过语法特征得到文本中的隐含信息。

但此类方法的缺点也是显而易见的：它们的效果严重依赖于语法分析的结果，其应用范围也受语法分析算法的约束。而语法分析研究的现状是：除了对几种使用较广的语言如英语等有较为成熟的语法分析算法外，很多语言都没有成熟的语法分析方法。而即使对于一些语法分析研究较为成熟的语言来说，对于不同的研究领域，其行文以及语法规范也都有着各自的特点，对于一些特殊的领域，现有的语法分析方法往往无法取得令人满意的结果。因此，基于语法分析的语义关系抽取算法其应用范围受到了很大的限制，特别是在一些有着独特语法规则的领域中更是如此，本文中提到的中医药领域就是一个例子。

第二、基于模式匹配的语义关系发现

这种方法是用某种模式对文本进行匹配，根据匹配的情况得到相应的语义关系。根据匹配模式的不同来源，可以分为两类：基于领域知识的模式匹配和基于学习的模式匹配（Hotho et al., 2005）。基于领域知识的模式是由领域专家将其领域知识总结、升华得到的通用知识模式，然后再使用这些模式作为模板，从文献中找到相应的关系。基于学习的模式是使用机器学习方法，通过对文献特征的分析，得到有用的模式。这类方法的目标一般限定为仅挖掘某些特定类别的实体的几种特定关系。

综上所述，这两类方法都不适合本项目的需求。在本项目中，我们采用基于已有语言系统的文本语义关系发现方法。我们选取TCMLS作为已有的语言系统，为语义关系发现方法提供重要的背景知识。该方法以TCMLS中的词汇为基础，发现语义关系中的主体和客体；它还以TCMLS中已有的语义关系为根据，推测从文本中挖出的语义关系的类型。总之，我们的基本策略，就是基于TCMLS从文本中发现更多的关系，经用户验证后插入TCMLS之中，从而丰富TCMLS的语义关系；再用丰富后的TCMLS进行新一轮的实验，进一步丰富TCMLS。以此类推，从而形成一套基于文本语义关系发现来驱动TCMLS加工的技术方案。下面具体介绍基于TCMLS的语义关系发现方法。

### 基于TCMLS的语义关系发现方法

如 所示，所谓文本语义关系发现，就是从“... 人参有‘补五脏、安精神、定魂魄、止惊悸、除邪气、明目开心益智’的功效...”这样的文本中，发现“人参 补 五脏”、“人参 安 精神”、“人参 止 惊悸”、“人参 除 邪气”这样的关系。另外，我们会统计每条关系出现的频数。如果在文档D1, D2, ..., Dn中都出现了某条关系，如“人参 除 邪气”，则“人参 除 邪气”出现的频数即为n。在本项目中，我们的基本策略是：以TCMLS作为领域词库，从文献库中找出在同一文档中出现的两个领域术语（如“人参、邪气”、“人参、五脏”等），构成候选的文本语义关系，并统计每条关系的频数，以供语言学家进行检阅和处理。由机器判断语义关系的谓词（如“人参”与“邪气”之间的谓词为“除”）仍是一个技术难题。在我们的策略中，会在这两个词附近找出一些候选性谓词（如“补”、“除”等）推荐给用户；并提供TCMLS中的相关用法，供用户参考（例如，针对“人参”和“肾阳虚证”，我们会根据TCMLS中的用法向用户推荐“治疗”这一谓词）。在本项目中，还开发语义关系检索与标注的系统，支持术语学家对文本语义关系进行检查、分析和标注等工作，并使用户了解语义关系与相关文本的关联。

为实现我们的语义关系发现策略，首先需要将中医药文本分解成一系列独立的语义单元。中医药文献资源包括书籍、期刊、会议论文集、病历、报告等，语义单元划分要针对不同类型的文献进行具体分析。本项目主要考虑书籍、期刊和会议论文集。一般而言，书籍会细为章节，期刊和会议论文集会分为一系列论文。因而，这些文献都可以被进一步细分为“文章”（在书籍中对应一章或一节，在期刊和会议论文集中对应一篇论文）。我们可以对文章进行进一步的细分：将文章分为小节（摘要也可被看成一小节）；将小节分为段落；将段落分为句子。

理论上，在一篇文章中出现任意两个词汇之间都可能具有某种或强或弱、或直接或间接的语义关系，因此也或可以整篇文章作为一个语义单元进行语义关系发现。然而，在语义关系识别阶段，关键词组过长将极大地增加算法的时间复杂性和空间复杂性。为保证算法效率，关键词组内词汇数量不宜过多。因此，本项目没有采用整篇文章作为语义单元。

与全文相比，句子是一个相对较小且具有完整语义的单元。句子通常描述一个完整的意思，且其内部的词汇有比较密切的关系。最为基本的“主语-谓语-宾语”的句式就代表了一条语义关系。因此，句子为中文分词以及后续处理提供了天然的单元。但在很多情况下，语义关系的主体和客体会分散在不同的句子中，因此仅以句子作为语义单元会遗失掉很多的关系。对于这类情况，可以通过以小节（摘要）或段落为单位的语义关系发现来处理。鉴于此，本项目分别以句子和小节作为基本的语义单元进行了语义关系发现的实验。

我们在方案设计的过程中，遵循了如下的一些设计原则：

1. 在文中距离越近的“名词、动词、名词”，越有可能表达一条语义关系。为此，我们会记录语义关系中的词汇在文中的最短距离，作为反映语义关系真实性的一个参数。
2. 在各种文献中多次出现的“主语-谓语-宾语”，更有可能代表一条语义关系。若一条关系频繁出现于各种文献中，则其很可能是领域专家认可的。鉴于此，我们对从各种文献中发现的关系进行了汇总，统计了每条关系中出现的文献频数。在文本语义关系展示系统中，将关系出现的频数作为一个参数提供给术语学家。
3. 文献量越大、越全面，所得到的语义网络就越具有统计价值。鉴于此，我们采用TCMLS对万方文献库进行检索（详见第5.2节），以期获得尽可能全面的文献。
4. 以常用动词表为基础，推测语义关系中的谓词。在常用动词表中，罗列了一些在中医药领域中常用的动词。我们将主语或宾语附近出现的动词记录下来，作为语义关系的候选谓词，推荐给术语学家。
5. 由术语学家对文本语义关系进行检验。鉴于文本语义关系发现方法尚不能保证结果准确性，我们需要一个文本语义关系的检阅系统，对文本语义关系进行检查、分析和标注等工作。该系统也可使术语学家了解语义关系与相关文本的关联。

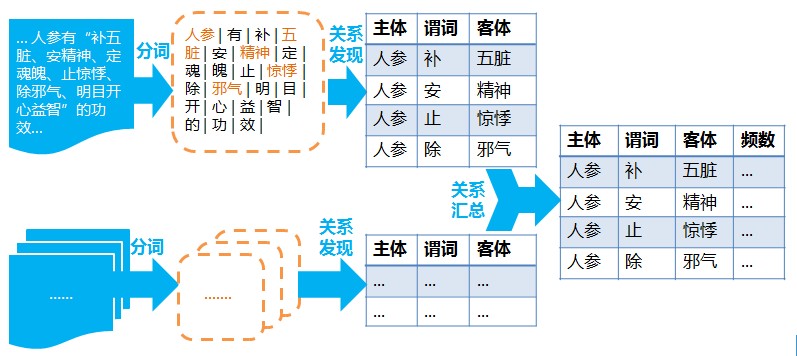


图 5‑18 文本语义关系发现过程示意图

下面介绍文本语义关系发现程序和实验，以及基于Web的文本语义关系检阅系统。

### Hamster：基于Java的文本语义关系发现程序

在本项目中，采用Java语言开发了一套文本语义关系发现程序，被称为Hamster。Hamster的命名空间为“com.cintcm.hamster”，它按照不同的包（package）可划分为4个不同的模块：

* **com.cintcm.hamster.relation**:它实现了几种语义关系发现的方法和策略，并包括一些辅助性的类，为Hamster的核心模块。
* **com.cintcm.hamster.relation.experiment**:它包括实现各种文本语义关系发现实验的脚本。
* **com.cintcm.hamster.relation.io**:它包括实现文件(或数据库)的输入/输出功能的程序。
* **com.cintcm.hamster.metadata：**它用于从PDF等格式的文件中提取题名、作者等文献元数据。
* **com.cintcm.hamster.db：**它包括访问数据库的程序。

下面介绍**com.cintcm.hamster.relation**中与语义关系发现相关的核心代码。首先介绍对某一段文本进行语义关系发现的调用代码：

|  |
| --- |
| /\*\*  \* 对某一段文本进行语义关系发现  \* **@param** txt 为文本的内容  \* **@param** docId 为文本的ID  \*/  **public** RelationExtractor(String txt, String docId) {  **this**.text = txt;  //对传入的文本进行分词处理  List<List<String>> word\_lists = Utils.*breakWords*(txt);  **this**.words = **new** ArrayList<String>();  **for** (List<String> l : word\_lists) {  **this**.words.addAll(l);  }  **this**.docId = docId;  //对传入的文本进行语义关系发现  **this**.extractRelations();  } |

下面是对语义关系进行发现的代码：

|  |
| --- |
| **private** **void** extractRelations() {  String s = **null**;  //对词汇序列进行遍历  **for** (**int** c1 = 0; c1 < words.size(); c1++) {  // s即为一个潜在的主体  s = words.get(c1);  // 如果s为中医药领域的一个核心术语，则进行进一步的处理  **if** (TCMCoreDecider.*isCore*(s)) {  //对词汇序列进行遍历，搜寻潜在的客体  **for** (**int** c2 = 0; c2 < words.size(); c2++) {  // obj即为一个潜在的客体  String obj = words.get(c2);  // 确保obj即为一个潜在的客体  **if** (obj.equalsIgnoreCase(s))  **continue**;  // 如果obj为中医药领域的一个词汇，则进行进一步的处理  **if** (TCMTermDecider.*isCore*(obj)) {  Set<Integer> verb\_nums = getVerbNums(c1, c2);  **boolean** verbFound = **false**;  **for** (Integer verb\_num : verb\_nums) {  **if** ((verb\_num >= 0) && (verb\_num < words.size())) {  //在主体和客体附近，找出一个词汇作为候选谓词  String pred = words.get(verb\_num);  //若谓词出现于动词词表中，则进行进一步的处理  **if** (VerbDecider.*isVerb*(pred)  && !pred.equalsIgnoreCase(s)  && !pred.equalsIgnoreCase(obj)) {  //生成新的关系：（主体、谓词、客体），记录文本内容和ID  Relation rel = **new** Relation(s, pred, obj)  .setDocId(**this**.docId).setText(  **this**.text);  //记录关系的取值，若客体为核心词汇则赋3分；否则赋2分  **if** (TCMCoreDecider.*isCore*(obj)) {  rel.setValue(3);  } **else** {  rel.setValue(2);  }  //记录关系中主体和客体之间的距离  rel.setDistance(Math.*abs*(c2 - c1) + 1);  verbFound = **true**; relations.add(rel);  }  }  }  **if** (!verbFound) {  //生成新的关系：（主体、""、客体），其中谓词为空，记录文本内容和ID  Relation rel = **new** Relation(s, "", obj).setDocId(  **this**.docId).setText(**this**.text);  //若未找到谓词，则赋1分  rel.setValue(1);  //记录关系中主体和客体之间的距离  rel.setDistance(Math.*abs*(c2 - c1) + 1);    relations.add(rel);  }  }  }  }  }  } |

下面介绍采用Hamster进行文本挖掘实验的过程以及结果。

### 文本挖掘实验过程及结果

我们以TCMLS作为关键词，从万方数据知识服务平台获取了217667条文献题录信息（含摘要）；采用Hamster工具，从中挖出了87826条关系。实验脚本如下图所示：

|  |
| --- |
| //建立文献数据库（其中存储着被挖掘的文献，并将存储挖出的语义关系）的连接  Connection conn = MySQLUtils.*getConnection*();  //创建SQL语句对象  Statement stmt = conn.createStatement();  //获取文献数据库中的文献数量  **int** num\_articles = getNumArticles(stmt);  //relationManager用于处理和存储挖出的语义关系  RelationManager relationManager = **new** RelationManager();  **for** (**int** i = 1; i <= num\_articles; i++) {  //用于获取一条文摘的SQL语句  String sql = "select id, description from resource where id=" + (i++);  //执行的SQL语句，获取一条文摘  ResultSet rs = stmt.executeQuery(sql);  **if** (rs.next()) {  String doc\_id = resultSet.getString(1);  String text = resultSet.getString(2);  //构建RelationExtractor对象  extractor = **new** RelationExtractor( text, doc\_id);  //得到RelationExtractor挖出的关系  List<Relation> relations = extractor.getRelations();  //将挖出的关系交给relationManager处理  relationManager.addRelations(relations);  }  //关闭结果集  rs.close();  }  //通过relationManager将挖出的关系存入数据库中。  relationManager.insertPairs("relation");  //关闭SQL语句对象  stmt.close();  //关闭数据库的连接对象  conn.close(); |

实验得到的关系被存入一个MySQL数据库中，并通过一个文本语义关系检阅系统展示出来，供语言学家进行检阅。在下一节中，将对这个文本语义关系检阅系统进行介绍。

### 基于Web的文本语义关系检阅系统

如 所示，本项目还开发了一套文本语义关系的检阅系统，支持用户完成文本语义关系的检阅、分析和标注工作，查看文本语义关系的文献依据和相关网页（借助Baidu搜索结果），并将文本语义关系正式插入某个术语系统（如TCMLS）。

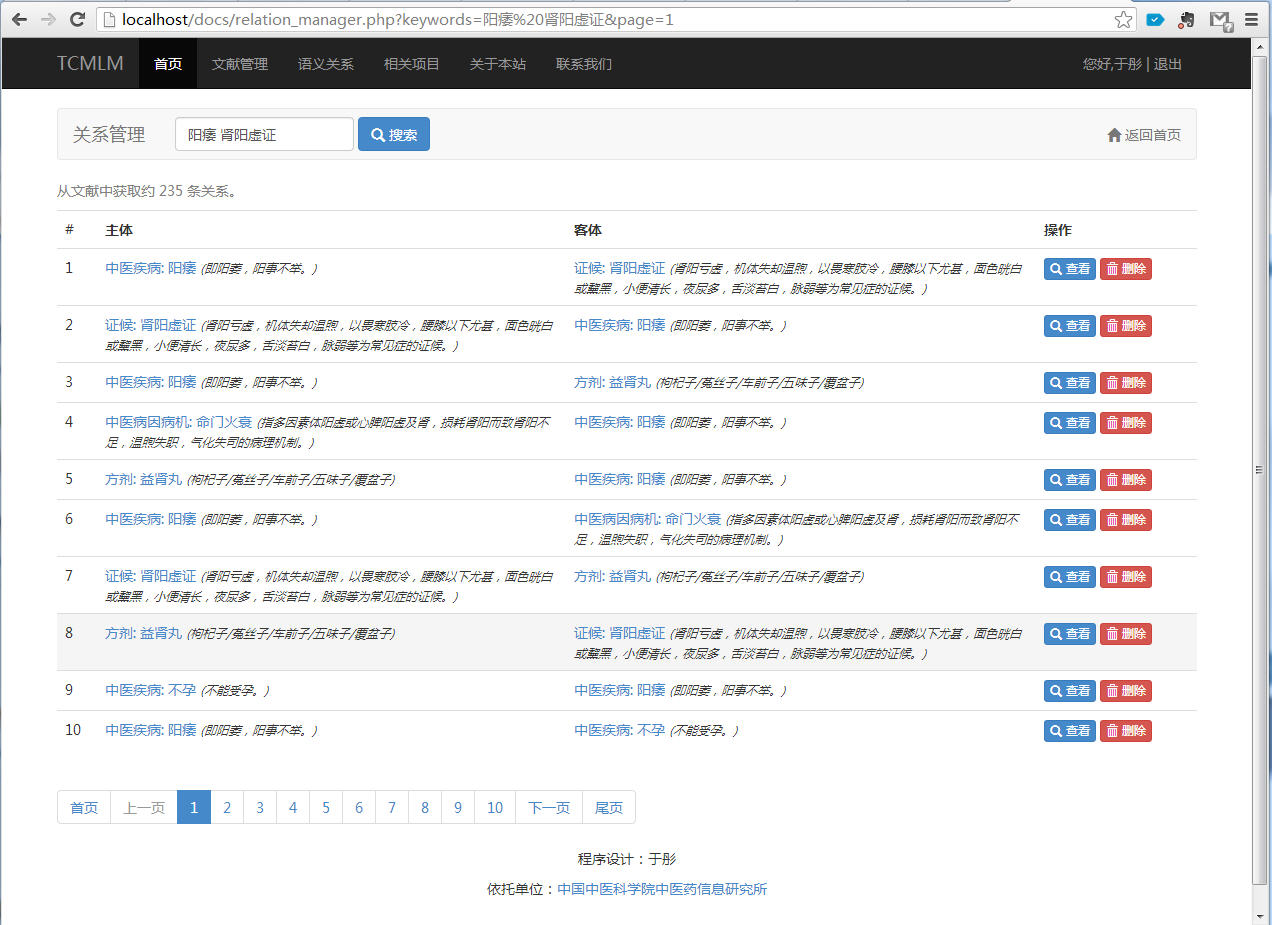


图 5‑19 文本语义关系检索界面

如所示，在语义关系检索界面中，系统会将机器发现的语义关系分页列出。用户可输入关键词（如“阳痿 肾阳虚证”）搜索个人关心的语义关系。系统列出了每条关系的主体和客体（如“益肾丸-肾阳虚证”），给出每个概念的类型、正名和定义。用户可点击查看某个概念，系统会转到这一概念的信息页面。当用户在上文提到的“语义关系检索界面”中点击查看某条关系，系统就会跳转到这条语义关系的展示和处理界面。



图 5‑20 文本语义关系展示界面

如所示，在语义关系的展示和处理界面中，用户可以查看这条关系的主体信息、候选谓词、客体信息、以及参考性参数。其中，对于主体和客体，都给出了概念的类型、正名、定义以及概念信息页面的链接。候选谓词是基于TCMLS中的用法来生成的，例如，若主体为“人参”，客体为“肾阳虚证”，则系统会推荐“治疗”作为候选谓词。



图 5‑21 文本语义关系的相关文献界面

如图 5‑21所示，用户可以点击“文献资源”，查看该语义关系所出自的文献。对于每篇文献，系统都给出了题名和摘要。用户单击选择某篇文献的题名时，系统会跳转到该文献的题录信息页面。如图 5‑22所示，用户可以点击“百度搜索”，查看该语义关系相关的百度搜索结果。



图 5‑22 文本语义关系的相关Baidu搜索结果界面

如图 5‑23所示，用户可以点击“相关陈述”，查看该语义关系在TCMLS中的相关陈述，以供语言学家参考。



图 5‑23 文本语义关系在TCMLS中的相关陈述界面

如图 5‑24所示，用户可通过如下图所示的表单，将这条语义关系加入语言系统中。系统会根据TCMLS中的相关用法，列出一些相关属性以供用户选择。用户也可以输入新的属性，例如，为添加“人参 除 邪气”这条关系，用户可输入“除”这一新属性。用户也可以添加一些注释信息。另外，当关系被录入TCMLS后，系统会记录这条关系的文献来源。



图 5‑24 文本语义关系录入界面

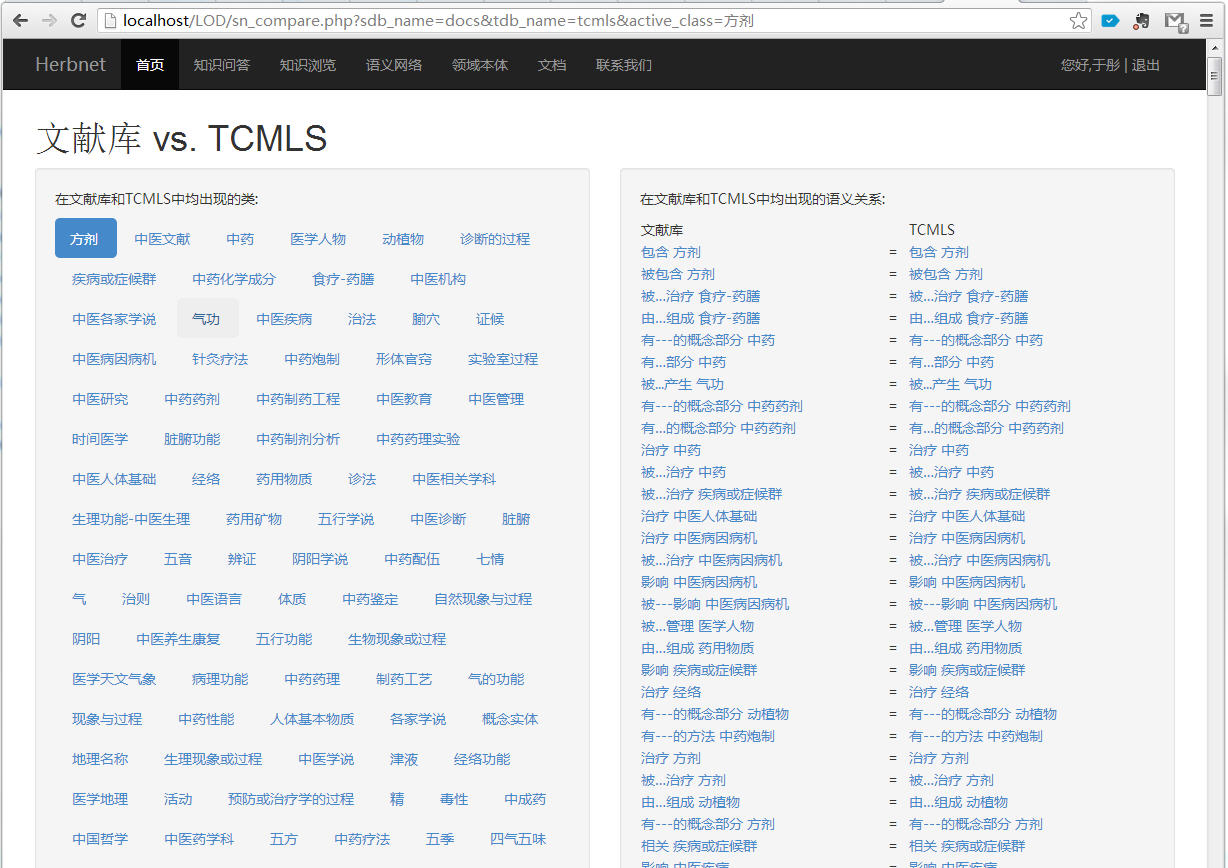


图 5‑25 语义网络比较界面

## 语义网络比较

对TCMLS的语义网络与从文献文本中实际抽取的语义网络进行了比较。如 所示，我们开发了语义网络的比较界面，它能够比较两个语义网络中类型的差异。其核心功能是在用户选定某个类后，比较该类在两个语义网络中的“用法”。系统首先列出了两个语义网络共有的语义类型，仅在文献库中出现的语义类型，以及仅在TCMLS中出现的语义类型。当用户选择一个语义类型（如：中药化学成分），系统会列出在两个语义网络中均出现的语义关系，以及仅出现在TCMLS或文献库中的语义关系。 对来自文献的语义网络与TCMLS语义网络进行了比较。

表 5.1 源自文献的语义网络与TCMLS语义网络的比较报表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 语义类型（主体） | 文献库语义类型间的语义关系 | TCMLS语义类型间的语义关系 | 共同的 | 仅在文献库里出现的 | 仅在TCMLS里出现的 |
| 方剂 | 1184 | 97 | 41 | 1143 | 56 |
| 中医文献 | 7 | 156 | 0 | 7 | 156 |
| 中药 | 730 | 122 | 31 | 699 | 91 |
| 医学人物 | 367 | 53 | 5 | 362 | 48 |
| 动植物 | 1256 | 226 | 67 | 1189 | 159 |
| 诊断的过程 | 406 | 254 | 26 | 380 | 228 |
| 疾病或症候群 | 591 | 169 | 50 | 541 | 119 |
| 中药化学成分 | 46 | 18 | 2 | 44 | 16 |
| 食疗-药膳 | 44 | 40 | 5 | 39 | 35 |
| 中医机构 | 0 | 7 | 0 | 0 | 7 |
| 中医各家学说 | 0 | 479 | 0 | 0 | 479 |
| 气功 | 95 | 114 | 5 | 90 | 109 |
| 中医疾病 | 571 | 218 | 43 | 528 | 175 |
| 治法 | 0 | 115 | 0 | 0 | 115 |
| 腧穴 | 27 | 130 | 3 | 24 | 127 |
| 证候 | 424 | 150 | 17 | 407 | 133 |
| 中医病因病机 | 211 | 386 | 19 | 192 | 367 |
| 针灸疗法 | 0 | 29 | 0 | 0 | 29 |
| 中药炮制 | 92 | 40 | 2 | 90 | 38 |
| 形体官窍 | 446 | 234 | 12 | 434 | 222 |
| 实验室过程 | 0 | 32 | 0 | 0 | 32 |
| 中医研究 | 0 | 32 | 0 | 0 | 32 |
| 中药药剂 | 22 | 33 | 3 | 19 | 30 |
| 中药制药工程 | 0 | 27 | 0 | 0 | 27 |
| 中医教育 | 0 | 12 | 0 | 0 | 12 |
| 中医管理 | 0 | 20 | 0 | 0 | 20 |
| 时间医学 | 0 | 115 | 0 | 0 | 115 |
| 脏腑功能 | 0 | 114 | 0 | 0 | 114 |
| 中药制剂分析 | 0 | 10 | 0 | 0 | 10 |
| 中药药理实验 | 0 | 19 | 0 | 0 | 19 |
| 中医人体基础 | 12 | 83 | 1 | 11 | 82 |
| 经络 | 76 | 145 | 5 | 71 | 140 |
| 药用物质 | 230 | 31 | 7 | 223 | 24 |
| 诊法 | 0 | 32 | 0 | 0 | 32 |
| 中医相关学科 | 7 | 20 | 0 | 7 | 20 |
| 生理功能-中医生理 | 0 | 98 | 0 | 0 | 98 |
| 药用矿物 | 12 | 21 | 0 | 12 | 21 |
| 五行学说 | 0 | 85 | 0 | 0 | 85 |
| 中医诊断 | 33 | 22 | 0 | 33 | 22 |
| 脏腑 | 80 | 282 | 1 | 79 | 281 |
| 中医治疗 | 0 | 22 | 0 | 0 | 22 |
| 五音 | 0 | 30 | 0 | 0 | 30 |
| 辨证 | 89 | 7 | 0 | 89 | 7 |
| 阴阳学说 | 0 | 100 | 0 | 0 | 100 |
| 中药配伍 | 112 | 16 | 1 | 111 | 15 |
| 七情 | 0 | 93 | 0 | 0 | 93 |
| 气 | 0 | 112 | 0 | 0 | 112 |
| 治则 | 0 | 35 | 0 | 0 | 35 |
| 中医语言 | 0 | 96 | 0 | 0 | 96 |
| 体质 | 0 | 51 | 0 | 0 | 51 |
| 中药鉴定 | 0 | 7 | 0 | 0 | 7 |
| 自然现象与过程 | 65 | 35 | 0 | 65 | 35 |
| 阴阳 | 0 | 102 | 0 | 0 | 102 |
| 中医养生康复 | 0 | 37 | 0 | 0 | 37 |
| 五行功能 | 0 | 57 | 0 | 0 | 57 |
| 生物现象或过程 | 0 | 34 | 0 | 0 | 34 |
| 医学天文气象 | 8 | 58 | 0 | 8 | 58 |
| 病理功能 | 0 | 18 | 0 | 0 | 18 |
| 中药药理 | 0 | 8 | 0 | 0 | 8 |
| 制药工艺 | 26 | 5 | 0 | 26 | 5 |
| 气的功能 | 0 | 58 | 0 | 0 | 58 |
| 现象与过程 | 0 | 36 | 0 | 0 | 36 |
| 中药性能 | 0 | 40 | 0 | 0 | 40 |
| 人体基本物质 | 129 | 100 | 0 | 129 | 100 |
| 各家学说 | 0 | 4 | 0 | 0 | 4 |
| 概念实体 | 55 | 50 | 0 | 55 | 50 |
| 地理名称 | 0 | 6 | 0 | 0 | 6 |
| 生理现象或过程 | 0 | 9 | 0 | 0 | 9 |
| 中医学说 | 0 | 25 | 0 | 0 | 25 |
| 津液 | 0 | 33 | 0 | 0 | 33 |
| 经络功能 | 0 | 6 | 0 | 0 | 6 |
| 医学地理 | 395 | 8 | 0 | 395 | 8 |
| 活动 | 0 | 7 | 0 | 0 | 7 |
| 预防或治疗学的过程 | 0 | 10 | 0 | 0 | 10 |
| 精 | 0 | 45 | 0 | 0 | 45 |
| 毒性 | 0 | 3 | 0 | 0 | 3 |
| 中成药 | 0 | 0 | 0 | 0 | 0 |
| 中国哲学 | 0 | 30 | 0 | 0 | 30 |
| 中医药学科 | 0 | 2 | 0 | 0 | 2 |
| 五方 | 0 | 13 | 0 | 0 | 13 |
| 中药疗法 | 0 | 3 | 0 | 0 | 3 |
| 五季 | 0 | 16 | 0 | 0 | 16 |
| 四气五味 | 0 | 51 | 0 | 0 | 51 |
| 食物 | 0 | 5 | 0 | 0 | 5 |
| 按摩 | 0 | 8 | 0 | 0 | 8 |
| 天然物质 | 0 | 10 | 0 | 0 | 10 |
| 实体 | 0 | 6 | 0 | 0 | 6 |
| 阴阳功能 | 0 | 16 | 0 | 0 | 16 |
| 中医病因病机疾病 | 0 | 24 | 0 | 0 | 24 |
| 五色 | 0 | 8 | 0 | 0 | 8 |
| 有机体 | 0 | 2 | 0 | 0 | 2 |
| 中药药理作用 | 0 | 0 | 0 | 0 | 0 |
| 津液功能 | 0 | 2 | 0 | 0 | 2 |
| 事件 | 0 | 3 | 0 | 0 | 3 |
| 针灸研究 | 0 | 0 | 0 | 0 | 0 |
| 时间概念 | 0 | 6 | 0 | 0 | 6 |
| 情志 | 0 | 1 | 0 | 0 | 1 |
| 精的功能 | 0 | 5 | 0 | 0 | 5 |
| 腧穴功能 | 0 | 0 | 0 | 0 | 0 |
| 归经 | 0 | 2 | 0 | 0 | 2 |
| 医疗设备 | 0 | 0 | 0 | 0 | 0 |
| 中医预防 | 0 | 0 | 0 | 0 | 0 |
| 中医信息文献学 | 0 | 1 | 0 | 0 | 1 |
| 中医基础理论 | 0 | 0 | 0 | 0 | 0 |
| 血 | 227 | 7 | 0 | 227 | 7 |
| 神 | 0 | 5 | 0 | 0 | 5 |
| 物理实体 | 0 | 2 | 0 | 0 | 2 |
| 中药性能与功能 | 0 | 3 | 0 | 0 | 3 |
| 辅料 | 0 | 0 | 0 | 0 | 0 |
| 五行 | 0 | 10 | 0 | 0 | 10 |
| 症状或体征 | 0 | 1 | 0 | 0 | 1 |
| 五化 | 0 | 2 | 0 | 0 | 2 |
| 中药功能 | 0 | 7 | 0 | 0 | 7 |
| 血的功能 | 0 | 3 | 0 | 0 | 3 |
| 矿物 | 0 | 0 | 0 | 0 | 0 |
| 方剂功用 | 0 | 1 | 0 | 0 | 1 |
| 针灸处方 | 0 | 0 | 0 | 0 | 0 |
| 声音 | 0 | 4 | 0 | 0 | 4 |
| 心理功能-中医心理 | 0 | 3 | 0 | 0 | 3 |
| 雌雄霹雳火 | 0 | 0 | 0 | 0 | 0 |
| 经络学说 | 0 | 1 | 0 | 0 | 1 |
| 11 | 11 | 11 | 11 | 11 | 11 |

## 实现技术

本系统采用PHP、Javascript、MySQL等主流技术进行开发，部署于国际主流的Apache服务器，利用万方提供的文献资源和中医药信息研究所开发的数据库作为文献/数据资源，采用Github作为版本控制系统。

表 5.2 系统实现技术列表

|  |  |  |
| --- | --- | --- |
| 技术名称 | 技术类型 | 说明 |
| Apache | Web服务器 | 采用Apache服务器提供Web服务。 |
| PHP | 编程语言 |  |
| Java | 编程语言 | 使用Java语言实现文本挖掘算法。Java语言拥有众多的开源软件，支持跨平台运行。 |
| MySQL | 数据库 |  |
| Github | 版本控制 | 采用GitHub进行版本控制：在Github上搭建了资源库Hamster、Docs等。 |
| 万方知识服务平台 | 网络文献资源 |  |
| MMSeg4J | 分词程序 | 采用Chih-Hao Tsai开发的MMSeg4j分词程序实现中文分词的功能。 |
| TCMLS | 术语系统 | 采用TCMLS作为MMSeg4j所使用的词库。 |

## 小结

本项目向术语专家提供新的技术能力，主要包括从语义关系中归纳“顶层语义网络”的技术能力，以及向术语专家提供从文本中发现新颖语义关系的技术能力。这项工作尚存在一些局限性。例如，我们尚缺乏判断文本语义关系准确类型的有效手段，也尚未实现发现新词的方法。另外，有些中医药领域的词汇尚未收入TCMLS之中，这影响了语义关系发现的效果。

# 项目计划

该项目预期开发周期为3个月。对研究工作作如下安排：

表 6.1 项目开发的总体计划

|  |  |  |
| --- | --- | --- |
| 时间 | 任务 | 说明 |
| 10月 | **语义关系发现** | 1. 从万方抓取10-100万篇文献摘要 2. 从文本中挖掘语义关系。 3. 由领域专家对所发现的语义关系进行检查和评估。 |
| **TCMLS数据抽取** | 1. 访问TCMLS的数据库，并从中抽取实验所需的数据。 |
| **语义网络归纳方法** | 1. 将具体的语义关系（如“人参 治疗 肾虚”）归纳为语义类型之间的语义关系（如“中药治疗 证候“）。 |
| **语义网络归纳实验** | 1. 将TCMLS的语义关系归纳为一个语义网络 2. 将从文本中挖掘的语义关系归纳为一个语义网络 |
| 11月 | **中医药语义网络比较研究** | 1. 将从TCMLS中归纳出的语义网络，与从文献中归纳出的语义网络进行比较分析。 2. 用从文献中归纳出的语义网络，补充完善TCMLS现有的语义关系网络，形成能够指导中医药学语言系统实际发展和应用的语义网络。 3. 成果汇报会 |
| 12月 | **撰写文档** | 1. 技术报告（用户手册） |

后续工作包括：申请专利和软件著作权；撰写核心期刊和SCI论文等。

# 网络资源

* <http://technology.chtsai.org/wordlist/>. Chih-Hao Tsai. A Review of Chinese Word Lists Accessible on the Internet.
* http://technology.chtsai.org/mmseg/. Tsai, Chih-Hao. "MMSEG: A word identification system for Mandarin Chinese text based on two variants of the maximum matching algorithm."

# 参考文献

Agrawal R, Srikant R. (1994). Fast algorithms for mining association rules. VLDB 1994:487–99.

Berry MW, Castellanos M. (2007). Survey of text mining II: clustering, classiﬁcation, and retrieval. 1st ed. Springer.

de Bruijn B, Martin J. 2002. Getting to the (c)ore of knowledge: mining biomedical literature. Int J Med Inform;67(1–3):7–18.

Cakmak A, Ozsoyoglu G. 2008. Discovering gene annotations in biomedical text databases. BMC Bioinform;9:143.

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. Briefings in bioinformatics, 6(1), 57-71.

Feldman R, Sanger J. (2006). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press.

Ganiz M, Pottenger WM, Janneck CD. (2006). Recent advances in literature based discovery. JASIST. Available from: http://www.dimacs.rutgers.edu/~billp/pubs/JASISTLBD.pdf. Accessed: 20 October 2009.

Hearst MA. (1999). Untangling text data mining. Proc Assoc Comput Linguist 1999;37:3–10

Hirschman L, Park JC, Tsujii J, Wong L, Wu CH. 2002. Accomplishments and challenges in literature data mining for biology. Bioinformatics;18(12):1553–61.

Hotho, Andreas, Andreas Nürnberger, and Gerhard Paaß. (2005). "A brief survey of text mining." In LDV Forum-GLDV Journal for Computational Linguistics and Language Technology, vol. 20, no. 1, pp. 19-62.

Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. 2005. Using literature-based discovery to identify disease candidate genes. Int J Med Inform;74(2–4):289–98.

Jensen LJ, Saric J, Bork P. 2006. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet;7:119–29.

Jenssen T-K, Laegreid A, Komorowski J, Hovig E. 2001. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet;28:21–8.

Jiang, T., A.-H. Tan, et al. (2007). Mining Generalized Associations of Semantic Relations from Textual Web Content. IEEE Trans. on Knowl. and Data Eng. 19(2): 164-179.

Jimeno A, Jimenez-Ruiz E, Lee V, Gaudan S, Berlanga R, Rebholz-Schuhmann D. 2008. Assessment of disease named entity recognition on a corpus of annotated sentences. BMC Bioinform;9(Suppl. 3):S3.

Krallinger M, Valencia A. 2005. Text-mining and information-retrieval services for molecular biology. Genome Biol;6:224.

Li S, Zhang ZQ, Wu LJ, Zhang XG, Li YD, Wang YY. (2007). Understanding ZHENG in traditional Chinese medicine in the context of neuro-endocrine-immune network. IET Syst Biol; (1):51–60.

马明,武夷山. Don R. Swanson的情报学学术成就的方法论意义与启示. 情报学报. 2003, 22(3): 259-266.

Peng Y, Zhang X. (2007). Integrative data mining in systems biology: from text to network mining [J]. Artif Intell Med;41(2):83–6.

Shatkay H, Feldman R. 2003. Mining the biomedical literature in the genomic era: an overview. J Comput Biol;10(6):821–55.

Simpson, M. S., & Demner-Fushman, D. (2012). Biomedical text mining: A survey of recent progress. In Mining Text Data (pp. 465-517). Springer US.

Sougata Mukherjea, Bhuvan Bamba, and Pankaj Kankar. 2005. Information Retrieval and Knowledge Discovery Utilizing a BioMedical Patent Semantic Web [J]. IEEE Transactions on Knowledge and Data Engineering 17(8): 1099-1110.

Swanson DR. (1986). Fish oil, Raynaud’s syndrome, and undiscovered public knowledge. Perspect Biol Med;30:7–18.

Swanson DR. (1991). Complementary structures in disjoint science literatures. In: Proceedings of the 14th ACM SIGIR. New York: ACM Press; p. 280–9.

Swanson DR, Smalheiser NR. (1997). An interactive system for ﬁnding complementary literatures: a stimulus to scientiﬁc discovery. Artif Intell;91(2):183–203.

陶金火. (2011). 基于语义的中医药数据采集工程及应用平台 (Master's thesis, 浙江大学).

陶金火, 陈华钧, & 胡雪琴. (2011). 中医药文献语义关系图发现. 计算机科学, 38(3), 213-217.

Tsuruoka Y, McNaught J, Ananiadou S. 2008. Normalizing biomedical terms by minimizing ambiguity and variability. BMC Bioinform;9(Suppl. 3):S2.

Weeber M, Kors JA, Mons B. (2005). Online tools to support literature-based discovery in the life sciences. Brief Bioinform;6(3):277–86.

Wu Z, Zhou X, Liu B, Chen J. (2004). Text mining for finding functional community of related genes using TCM knowledge. In: Proceedings of the 8th PKDD, LNAI 3202. Berlin: Springer-Verlag; 2004. p. 459–70.

夏龙，符建辉，王石，曹存根. 从医学文本中获取知识的方法研究 [C]. 中国中医药信息研究会中医药信息数字化专业委员会中国中西医结合学会信息专业委员会2012年度学术交流会论文集，2012.

许建阳,马明,王发强. Swanson的非相关文献知识发现法对医学发展的思考.医学与哲学.2003, 24(8):21-23.

Yandell MD, Majoros WH. (2002). Genomics and natural language processing. Nat Rev Genet;3(8):601–10.

Yetisgen-Yildiza M, Pratt W. (2006). Using statistical and knowledge-based approaches for literature-based discovery. J Biomed Inform;39(6):600–11.

张小刚.基于中医药本体的语义关系发现及验证方法[D].浙江大学计算机学院,2010.

Zhou D, He Y. (2008). Extracting interactions between proteins from the literature. J Biomed Inform;41(2):393–407.

Zhou X, Liu B, Wu Z. (2005). Text mining for clinical Chinese herbal medical knowledge discovery. Lect Notes Comput Sci;3735:396–8.

Zhou, X., Peng, Y., & Liu, B. (2010). Text mining for traditional Chinese medical knowledge discovery: a survey. Journal of biomedical informatics, 43(4), 650-660.

Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. 2007. Frontiers of biomedical text mining: current progress. Brief Bioinform;8(5):358–75.

1. 该服务接口的说明由万方公司技术人员王哲（13811901609）提供。 [↑](#footnote-ref-1)