浙江大学计算机科学与技术学院

硕士学位论文

中医药多源搜索引擎推荐系统研究及其实现

姓名：施少敏

申请学位级别：硕士

专业：计算机应用

指导教师：魏宝刚;吴江琴

20100128

摘要

中医药文化博大精深，通过阅读数字图书馆中的中医著作或者浏览相关电子资料是一种了解、学习中医药知识的有效方式，而这些著作和电子资料是通过搜索引擎呈现给用户的。虽然传统中医药图书推荐系统会在用户搜索的同时提供字面相关的推荐词条或者图书推荐页面，但是这些推荐词条都是搜索关键词的字面衍生，推荐页面也仅仅含有搜索文本，而没有先后之分。实际上用户关心的不仅是搜索文本字面相关的信息，也可能是与搜索文本语义、语境等相关的知识术语、不同媒介的知识载体。因此传统推荐系统满足不了这种学习认知模式。

本文中设计的中医药多源搜索引擎推荐系统整合了中医药知识的各方面信息，包括中药、方剂、病证、图像、视频、药物产地、名家名医、著名医馆等多源信息，推荐项是各方面信息的聚合，力求帮助用户快速有条理地学习中医知识，多角度探触搜索项的方方面面。

该推荐系统以数字图书馆 CADAL 项目为背景，利用自生成正则表达式的信息抽取技术从 OCR 书籍和网络中提取信息，使用双数组 Trie 树算法分离药物、方剂主治信息，使用文中提出的方剂分离算法对方剂组成信息进行细化，之后在中医药辞典的基础之上，使用了“辞典学习”的学习算法，得出与用户搜索语义相关的关键词，并结合网络信息以及用户关注度协同过滤算法得到推荐词条集合，然后对推荐词条集合采用反向索引的方式找到索引页面并对页面进行评估，最后对页面重新排序再推荐给用户。

关键词：数字图书馆，信息抽取，机器学习，推荐算法，多源知识聚合

Abstract

Chinese Traditional Medicine culture has a long history and also has a profundity for its variety. By visiting digital library, users can read masterpicces or other electronic records to discover the knowledge in Traditional Chinese Medicine. All these materials can be provided by using search engine exists in digital library. The search engine recommender can give us a lot of literal-related recommended words and book pages. However, the recommended words generated maybe not semantically associated and the book pages would be without sorted. Actually, readers want to learn the detail description of a query words from the recommended pages, even more, they may concern with other terminologies and other carriers of knowledge which are intrinsic related with the search term, so current search engine recommender system does not satisfy the reading cognitive model.

The multi-source recommended system designed in this thesis integrates the various aspects knowledge of Chinese Traditional Medicine including medicine, prescription, illness, picture, video and etc. The aim of this system sought after is to help users learn traditional Chinese medicine knowledge more methodically and probe kinds of aspects of search items in multi-angle.

First of all, the system uses auto-generated regular expression to extract information from OCR books and web. Secondly it uses Double Array Trie Tree to detach the efficacy filed of medicine and prescription, and then it uses an algorithm proposed in this thesis to treat with the components of prescription. At last it uses a dictionary learning algorithm presented in this thesis to get semantic words items prevalent with search item, combined with additional recommend items from network and logs formed as the recommended set.

Keywords: Digital Library, Information Extraction, Machine Learning, Recommended Algorithm, Multi-Source Knowledge Aggregation

图目录

表目录

浙江大学研究生学位论文独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得浙江大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：　　签字日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解浙江大学有权保留并向国家有关部门或机构送交本论文的复印件和磁盘，允许论文被查阅和借阅。本人授权浙江大学可以将学位论文的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月　　日

第1章 绪论

1.1 数字图书馆和图书推荐系统

数字图书馆是用数字技术处理存储各种文献的图书馆，其本质上是一种聚合各种多媒体技术的分布式信息系统。数字图书馆将处于不同载体、位于不同地理位置的信息资源用数字化技术加以存贮，以便于跨越区域、面向对象的网络查询和传播。它涉及信息资源加工、存储、检索、传输和利用的全过程[1]。

由于数字图书馆是互联网上的信息聚合中心，故各国都投入巨资建设数字图书馆，力图繁荣国民教育，提升民族素质。美国最早开展数字图书馆建设，相继启动了数字图书馆先导计划 I 期和 II 期。欧洲数字图书馆 Europeana 于 2008 年 11 月 20 日正式开放服务，提供了 200 万册数字图书、录音、图片、档案以及电影资料。2009 年 4 月 21 日，世界数字图书馆正式启用，由教科文组织及 19 个国家的 30 余个公共团体合作建设，提供 7 种语言服务，全球读者可免费使用其中的图书、地图、手抄本、影片与照片等。近年来，我国也相继开展数字图书馆建设，即将于 2010 年完工的国家数字图书馆工程将提供 2 亿条以上结构化元数据和 1 亿页全文数据。

图书推荐系统对图书馆数据库保留的大量用户来源 IP 地址、点击次数、图书全文、目录内容信息进行挖掘，分析得出用户的阅读模式，进而将用户最希望看到的图书通过网页链接的方式呈现在读者面前，达到辅助学习的目的。图书推荐系统可以对数据进行整合，改变过去被动服务模式，主动地将合适的数字图书在合适的时间推送给合适的读者，解决海量信息带来的信息过载、信息迷失等问题。

1.2 课题背景及其意义

业已建设的高等学校中英文图书数字化国际合作（CADAL）一期项目中，收

录了上万种中医药类图书，不乏珍本善本，里面蕴含的内容需要通过现代信息化的手段加以表现。书籍中蕴含的中医药信息包括中医的辨证施治的理论、各名家的学说、中药的种植、中医科研等领域的信息，如果能够将其有机的结合起来，形成一个知识网络，不但能够提供一个中医药相关信息的检索平台，而且能够给读者提供有效的学习帮助。中医学作为世界医学的一部分，数千年来为华夏各民族的繁衍生息做出了不可磨灭的贡献。先人留下的文化遗产—中医药图书是中华民族的文化瑰宝，蕴含于其中的朴素唯物主义思想。其辨证施治的原理、望闻问切的诊断方法，已经或正在对世界医学体系的发展、诊疗手段的提高产生深远的影响，中医中体现的“整体观念，以人为本”的思想逐渐被人们所接受。

目前，中医药研究已经进入了信息时代，研究的方法及其手段都随着计算机技术的发展而不断地更新。将数据挖掘、信息检索等技术运用于中医药古籍、科学研究、加工、炮制等文献的推荐已经成为可能。中医药在历史发展过程中形成了自己独特的理论和专业术语，在一定程度上保持了中医药理论的特色，但也成为中医药图书推荐系统的障碍之一。应用现代科学技术对中医药理论和实践进行科学阐释，特别是进行信息化、数字化和知识化的研究是促进中医药国际化和现代化进程的重要内容，对中医药信息进行整合并进行推荐是促进人们了解并掌握中医药知识的途径之一。

当前中医药类检索平台提供的服务仅仅是依照由用户提供的书籍名称，检索出与条件相关的书籍，并在详细介绍中提供作者和出版社等元信息。但是实际上读者所需要的并不仅仅是这些，他们需要更深层次的服务。对于研究人员来说，更加需要的是从书籍中提取主要的医学信息为之所用。否则依据检索出来的书籍再去查看就跟普通的阅览没有实质上的区别。如果新的中医药推荐系统能解决上述问题，那么必将使中医药信息化建设迈上新的台阶。

系统收录的中医书籍和多媒体资料中，包括中药、方剂、诊断以及中医经典理论书籍、图像和视频。这些资料较为全面地涵盖了经典中医科学中所描述的知识。

中药是在中医理论的指导下，用于预防、治疗、诊断疾病并具有康复和保健

作用的物质。中药主要来源于天然药及其加工品，包括植物药、动物药、矿物药及部分化学、生物制品类药物[2]。因此中药信息的聚类，图片的检索、产地的GIS定位、内容的跨媒体检索都成为了良好的信息推荐研究方向。推荐系统既可以为用户提供药物产地方面的信息，也可以在不同的中药之间建立一定的关系，比如主治，植物形态等。

方剂学是研究和阐明方剂治法和理论及其临床运用的学科，是中医学重要的基础学科之一。方剂是中医临床用药的主要形式和手段，其配伍规律有着深刻的科学内涵。一首方剂的确立，要经过审证求因、据证立法、择药定量、合理配伍等一系列的抽象思维过程，因而中医方剂又是祖国医学辨证论治精髓的集中体现[3]。近几年来，随着对方剂配伍科学内涵的逐步揭示，方剂学在中医药现代化研究中重要地位的认识不断深入，形成了一个前所未有的方剂学研究热潮。数据挖掘是目前有效处理和利用海量方剂数字信息的主要手段，是解决中医方剂信息过载但知识缺乏等问题的主体方法。目前建立的大多数中医药方剂数据库，只能提供检索、统计等一般性的服务，其包含在这些数据中的大量隐含知识尚未得到充分挖掘和利用。

中医诊断学是根据中医学的理论，研究诊察病情、判断病种、辨别证候的基础理论、基本知识和基本技能的一门学科[4]。中医的“辨证论治”是一种典型的个体化诊疗方法。长期临床实践中所积累起来的海量病例信息为中医临床疗效评价提供了最为准确、详实的依据，所以它是中医疗效评价体系的信息基础。各种病证直接或者间接的联系可以通过文本挖掘技术加以呈现，各种诊断经验可以通过决策理论加以支持。

中医古籍是中医名家留下的从医精华，是广大中医工作者的行医指南。里面的实践理论，行医经验需要通过网络加以推荐和弘扬。可以利用流行的协同推荐技术将医学工作者认可的书籍页面呈现给更多的读者。

CADAL 中医药推荐系统是一种基于辞典、网络文档、网络热门推荐词、页面推荐价值、用户兴趣度评估等多源混合信息推荐系统，在基于后台辞典文本数据、网络文档的机器学习基础上，结合网络用户搜索关注度，经过相似度权值矩

阵叠加、阀值聚类、网络热门推荐词语过滤排序等一系列步骤，综合给出初步的推荐词条集合，最后用反向索引机制查找含有搜索文本的页面，并通过页面价值评估函数对搜索结果页面进行合理排序。在推荐集合中周期性地考虑用户点击流，以达到推荐协同过滤、更新和优化的目的，这种多源推荐设计给予了用户学习传统中医新的体验。

1.3 本文所做的主要工作

首先，本文研究了现有的图书搜索推荐相关算法以及系统相关架构，研究了Web文本挖掘技术，自然语言理解理论在图书推荐中的应用。这里包括词语的信息处理、分词技术、句法分析技术、语料库结构、语义网络的构建等。利用网络爬虫技术，爬取数万条中医名词术语、中药植物学术语及其解释。利用 Web2.0技术对信息检索、个性化推荐等应用做了实现。主要工作如下：

（1）信息获取

利用 OCR 书籍、网络资源构建中医药信息资源数据库，利用爬虫技术构造URL请求，使用异步多线程获取中医药解释文本和中药图片，采用正则文法技术提取书籍非结构化信息。

（2）信息抽取及其关联

由于抽取的文本为非结构化文本，需要进行结构化管理存储。首先生成正则表达式对文本进行粗分，而后参考中医药辞典，分别构建了中医、中药、方剂、病证、产地、植物科属等术语词典，在此基础之上构建双数组 Trie 树（前缀树），抽取文本中的相关词语。对中医药书籍目录进行信息抽取，过滤掉无关页码，使术语和中医药图书页面直接对应。

（3）构建推荐框架

整个搜索推荐框架分为机器学习、网络协同更新、页面价值评估三大模块，当用户输入搜索文本时，将得到推荐词条和推荐页面。其中推荐词条来自机器学习模块，推荐页面来自页面价值评估模块。机器学习模块包含辞典学习、网络文档学习和网络词条推荐学习三个部分。网络协同更新模块主要有两大任务：其一

是维护用户推荐词条点击日记，该日记对推荐词条点击量进行统计，经过一段时间后，根据点击量的大小，后备推荐词条将获得特定替换概率，替换点击量较少的推荐词条，最后返回更新后的推荐结果。其二是周期性更新网络流行词条和网络文档数据，由于网络流行词条和网络文档不断变化，故必须周期性更新学习数据，使得学习结果集动态更新。

（4）用户 Web2.0 动态界面设计

整个系统建立在 Java EE 的 Web 架构之上，采用 MVC 设计模式构建业务逻辑，页面呈现中采用 Ajax 技术异步获取后台服务器数据，达到区域刷新的效果。区域采用 jQuery 技术动态显现、隐藏技术，同时实现动画效果，加深用户体验。

1.4 论文结构安排

本文第一章给出了课题背景，然后给出了课题研究意义和本文所做的主要工作。第二章中给出了 CADAL 推荐系统相关技术分析。此章中介绍了 CADAL 项目的现状，数字图书馆涉及的一些技术。详细介绍了信息推荐的相关技术，包括信息抽取、全文检索、排序、信息压缩和传送、Web 信息挖掘、协同过滤、基于知识的推荐技术等，同时阐述了如何评估推荐系统的优劣，最后给出了推荐系统的发展趋势。在第三章中，给出了 CADAL 中医药推荐系统总体设计的细节。包括系统框架以及相关任务，紧接着给出了中医药数据的采集方法，以及相关数据的集成和变换。再者陈述了搜索引擎界面的设计原则，推荐的产生反馈处理以及用户数据的获取处理工作。论文的第四章涉及系统的一些技术细节，从基于正则表达式的实体抽取算法、基于双数组 Trie 树的信息分离算法、语义推荐算法、页面评估算法到最后的实验结果分析，论文中给出了具体的实现公式和相关数据结构。第五章是系统的实现情况，介绍了系统底层软件的支撑平台，接着介绍了元数据管理模块的实现和数据处理模块，最后给出了流程控制模块和页面显示模块的介绍，在第六章给出了论文的总结和展望。

第2章 CADAL推荐系统相关开发技术分析

2.1 CADAL项目介绍

2002年9月，国家计划委员会、教育部、财政部将“中英文图书数字化国际合作计划（简称CADAL）”列为“十五”期间211工程公共服务体系建设的重要组成部分[5]。CADAL一期项目已经数字化100万册中英文图书。二期项目将进一步扩大数字资源建设的范围和数量，完成300万册文献资源的数字化制作，使项目的数字资源总量达到400万册，继续保持国内外公益性数字图书馆规模的领先地位。项目在保证高校用户资源获得率和服务满意率同时，还需满足大部分高校对于特定数字资源的需求。建立跨学科、跨媒体的立体知识结构框架，提供具有知识创新能力的数字资源集成服务。

CADAL数字图书馆中不仅包含百万册数字图书，还含有其他媒体形式的资源，如图像、视频、音频和物件模型等，以及书法、文物等其它传统资源，存储容量达到数百TB。目前项目建设重点已经从先前“大规模数字化”转移到“服务与资源并重”。与传统实体图书馆相比，CADAL数字图书馆在服务形态上面有着巨大的优势和创新空间：在图书全文、元数据、目录章节上应用搜索技术，能够帮助用户快速准确地定位到其所需要层面的信息；个性化推荐技术的应用能够改变传统的被动服务模式，将合适的图书在第一时间推送给相关读者。与传统图书馆相比，CADAL数字图书馆力图使用个性化推荐技术解决海量信息所带来的“信息过载”、“信息迷失”问题[6]。CADAL个性化技术研究与应用当前还处于初级阶段，高级个性化技术在图书检索和个性化推荐服务中还未普遍应用，海量数字图书环境下应用个性化技术时所面对的高可伸缩性等问题也未被充分研究，中医药搜索推荐技术研究刚刚开始。

2.2 信息搜索推荐相关技术

信息搜索推荐技术涉及自然语言理解、信息抽取技术、检索技术、信息排序、信息压缩与传送、分布式信息处理、安全防范措施、Web数据挖掘、数据仓库与联机分析处理、海量规模图书存储、海量规模数据库访问技术、跨媒体呈现技术等[7]。信息检索的核心就是快速检索出和用户查询表达式相关的文档或各种多媒体信息。如果直接在数据库中简单匹配查询请求，往往会得到数量庞大无序结果集。用户所关心的可能只是此结果集的子集，如何赋予结果集中的元素不同的权重成为排序算法的核心。在检索之前，查询表达式词干的提取，无意词的停用、查询相关扩展等查询表达式理解技术也成为检索结果好坏的重要影响因素。

2.2.1 信息抽取技术

中医信息系统的数据来源通常是手工输入或购买已有的商用数据库。手工输入涉及数据的收集、整理和录入，费时费力、周期长。购买商用数据库则费用高、不易扩展。新的解决方案可以建立在信息抽取技术之上。信息抽取技术的主要目的是从非结构化文本中抽取出特定的事实信息。比如，可从中药书籍中抽取药物的药性、功效、应用、用法用量、使用注意、临床应用等信息；从方剂书籍中抽取出方剂的组成、功用、方解、适应症等信息。迄今为止，从自然文本、网页中提取生物、医学、分子学信息的技术可大致分为四类。

（1）基于自然语言处理的抽取方式。许多自然语言处理工具像MEDLEE[8]、UMLS[9]和GENIES[10]用于提取西医特定领域的信息。其中MEDLEE用于提取病人诊断报告中的诊疗信息。UMLS是一个统一医学术语系统，用于文本间医学信息的发现和医学问题的检索。GENIES是一个从文本中提取基因和蛋白质信息的工具。

（2）综合利用自然语言处理技术和文本结构抽取方式。目前采用这种原理的典型算法和系统有Rapier[11]、Srv[12]和Whisk[13]，其中最著名的是Whisk系统。Whisk主要是根据语义项的上下文实现敏感信息的定位，该系统首先根据分隔符（如标点符号、段落标记等）将源文档分隔成多个语义相关的文本块。在交互式

环境下，系统每一次呈现给用户一组语义相关的文本块。用户根据系统提供的文本，标记出感兴趣的信息并定义模式。系统使用语法分析器和语义类（如人名、机构名）分析器，分析出用户标记信息的语法成分和对应的语义类，生成基于语法标记和语义类标记的抽取规则，实现信息抽取。

（3）包装器方式。信息抽取根据事先由用户标记的样本实例，应用机器学习归纳算法，生成基于定界符的抽取规则。其中定界符是对感兴趣语义项上下文的一种描述，即根据其左右边界来定位语义项。该类方式和基于自然语言理解方式的最大不同在于：仅仅使用语义项的上下文来定位信息，并没有使用语言的语法约束。采用该原理的系统有 Stalker[14]，Softmealy[15]等。包装器方式主要运用在提取信息粒度较大的场合，比如提取某一段说明，且要求全文具有一定的书写模式。

（4）嵌入式分类树。此提取模式一般应用于网页信息提取。抽取的方法是根据用户定义的页面结构嵌套模式，生成一个表述逻辑结构、模式信息和语义信息的 XML 树。采用该类技术的系统有 Lixto[16]、Xwrap[17]、RouRunner[18]等。这类方法用于已经存在逻辑结构的信息提取，不能适用于本身逻辑结构无定义的文本。

这些抽取技术不能直接运用于本系统的原因在于：所有信息提取方法都跟特定的自然语言、问题域、信息来源相关，不能直接适用于中医药信息的提取。

2.2.2 全文检索技术

全文检索从最初的关键词匹配，到目前的基于语义本体的检索技术，经历了数十年的时间。各大型搜索引擎公司、数据库公司和开源组织是目前使用全文检索技术的推动和发展者。常用的全文检索策略如图 2.1 所示：网络和图书馆中的文本和各种多媒体信息通过结构化存储技术存储于数据库中，接着通过分类器技术将文本做初步分类产生分类文档，在使用各类词库和停用词库的基础上，利用索引工具建立反向索性倒排表，然后以 web 页面方式提供用户查询接口。

上述分类器常采用的分类算法为基于 TFIDF（词频文档权重算法）的 Rocchio 算法[19]、决策树、朴素贝叶斯模型、KNN 算法以及简单向量距离分类法等几种。

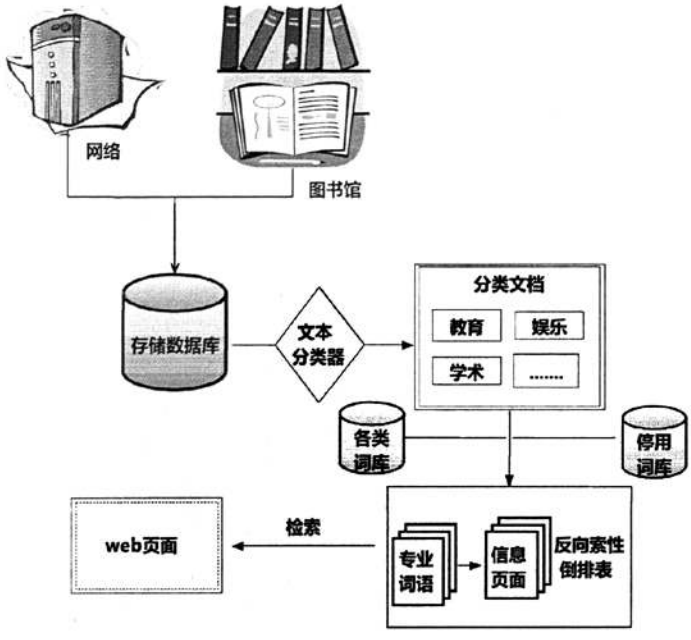

图 2.1 全文检索策略图

Rocchio 算法将文本分为训练文本和测试文本，通过计算训练文本的中心向量和测试文本向量的欧式距离，确定其所属的分类。算法优点是速度较快，但是对于类间距离较小，类内距离较大效果不佳。

决策树将文档向量和已经建立的决策树分枝进行匹配，判断文档的类别。通过训练样例构造决策树，以及修剪对数据的约束来进行进一步分类，由分支代表测试输出，叶节点代表类别。决策树最大的缺点在于当应对大规模数据处理时，构造决策树效率十分低下。

朴素贝叶斯模型采用统计分类的方法对文本进行分类。其理论基础是贝叶斯理论，分类时根据样本归属类别概率的最大值进行文本分类，贝叶斯分类优点在于分类准确且速度较快。

对于一个测试文本，KNN 算法找出和训练样本集中与该测试文本最相似的 K

篇训练文本，根据这些文档所属的类别计算文档所在类的权重。KNN 文本分类先抽取训练样本中关键词集合，建立关键词词典，通过词频得到分类器，而后将文本向量化，然后不断修正分类过程，期间，分类效果不断提高。KNN 算法对于类间距离偏小的测试样本性能有待提高。

2.2.3 排序技术

目前信息页面内容排序算法主要采用文本信息权重向量方法，其基本思想是通过计算页面中的信息和查询词相匹配的反向索引，对匹配向量的得分进行排序。排序算法排序效果优劣往往和查询时间呈一定正向关系，即，查询时间越长，则排序效果越佳。

基于链接分析的排序算法和基于机器学习的排序算法是网站排序经常采用的两类算法。

链接分析基于如下假设：链接表明了用户对网站的关注，外部链接到网站的数量越多，表明该网站越受用户关注，因此在检索的时候可以增大其排序权重。基于链接的思想，最著名的是 Google 采用的 PageRank 算法[20]以及 Klienberg 等提出的 HITS 算法[21]。

PageRank 算法的核心思想通过计算待检索网页的 rank 值，获得排序的优先次序。设 R(x)为网页 x 的 rank 值，则其计算公式为：

$$R(x) = \sum_{v \in B_x} \frac{R(v)}{N_v} \qquad \text{（公式2.1）}$$

B_x 是所有指向网页 x 的网页集合，N_v 代表网页 x 指向的网页集合的大小。因此，一张网页的 rank 值由其本身指向外部的链接递归得到。每一张反向网页链接贡献的值各不相同，如果网页 v 中指向外部的链接数越多，那么它对网页 x 的贡献就越小，x 的反向链接越多，那么其自身的 rank 值也随之增高。

机器学习排序算法一般采用通过对已知数据的训练得到排序模型，如早期的最大熵模型算法[22]、RankProp 算法[23]，流行的 RankNet 算法[24]、MFoM 算法[25]，RLR 算法[26]等。最大熵模型算法检索将检索相关性问题简化成为二值分类问题，

但是效果不佳。MFoM 算法基于最大化 ROC 曲线面积的思想，实现对物品的排序。Rankprop 本身是一种无监督蛋白质排序算法，适用于以蛋白质构建的相似度网络。可以用其他物品替代蛋白质将此算法加以推广。替换后的网络结构中，节点是物品，边代表物品序对的相似度，每个节点被赋予了初始值。对于某项查询，算法提供一个传播过程对结点的数值进行迭代，一个节点的值通过邻近节点加权合并和一个初始静态量相加获得，重复迭代此过程可以使得查询的值不断的通过网络进行传播，直到达到稳定状态。算法的输出目标是以查询项排序的蛋物品列表。RankNet 算法基于有序对数据，提出了基于交叉熵的损失函数，并且用神经网络进行优化。RLR 算法引入了常用的亏损函数到排序函数中，利用梯度下降法对风险函数计算得到风险最小估计量，得到最终的检索函数，由于它平衡了数据分布，即平衡了正误差和负误差的整体罚值，根据比例 M_D^- / M_D^+（即查询不相关文档数量和查询相关文档数量的比值）对正误差做出惩罚，从而为算法的有效性做出了保证。

2.2.4 信息压缩与传送技术

假若 Web 信息传送的数据量较大，则可以对 Web 页面进行数据压缩，在不影响页面内容的同时，提高网络信息传送速度。若使用 Apache 服务器，可以采用 mod_gzip，HTTP 压缩对于纯文本内容可压缩至原大小的 40%以下，从而提供 60%以上的数据传输节约[27]，虽然 WEB 服务器会因为压缩导致 CPU 占用的略微上升，但是可以减少大量网络 IO。对于数据压缩带来的用户浏览速度提升，使其符合 8 秒定律。mod_gzip 的工作原理是，当客户端请求浏览某网页后，Apache 服务器将所请求的网页文件进行压缩，具体压缩是利用标准的 zlib 压缩。服务器将压缩的文件下发给客户端浏览器，由客户端的浏览器负责解压缩并浏览，所以解压效果和用户的浏览器有一定关系。

在连接方面，JAVA WEB 服务器通过 TCP 连接和 SERVLET 容器连接。为了减少进程生成 Socket 的花费，WEB 服务器和 SERVLET 容器之间尝试保持持久性的 TCP 连接，对多个请求、回复循环重用一个连接。一旦连接分配给一个特

定的请求，在请求处理循环结束之前不会在分配。换句话说，在连接上，请求不是多元的，虽然这导致在某一时刻会有很多连接，但这使得连接两端的编码变得相对容易。因此本系统采用了 AJP 连接技术，AJP 是 Apache 提供的完成与其它服务器通讯的一种协议。在 Apache 中通过 mod_proxy_ajp 模块发送 AJP 数据，另外一端的服务器需要实现 AJP 协议，能够接受 mod_proxy_ajp 模块发送的 AJP 协议数据，在接受到 AJP 协议数据后做适当处理，并能够将处理结果以 AJP 协议方式发送回给 mod_proxy_ajp 模块[28]。

2.2.5 Web 信息挖掘技术

Web 数据挖掘就是从网络资源和用户行为日志中抽取特定模式以及隐含信息。挖掘形式大致可分为三类：Web 内容挖掘、Web 结构挖掘和 Web 使用挖掘。

Web 挖掘重在从网络数据中挖掘出潜在并符合用户兴趣的信息和行为，其挖掘过程可以看作为知识发现的网页扩展。

在数字图书馆应用中，Web 使用挖掘显得格外重要。用户对书籍或其他媒体的点击行为可以记录到系统日志中，可以将此日志作为基于用户行为推荐信息的训练数据，以此产生一个正向激励，不断改善推荐结果。Web 使用挖掘中可以通过路径分析技术确定某特定资源的访问路径，对访问频繁的路径进行优化，如使用“web 标签控件”、“导航页面”等减少页面切换。Web 使用挖掘中还包括序列模式挖掘技术，用于确定用户访问网页之间的序列关系。如有 90%的用户访问页面 A 后，又紧接着访问网页 B，那么说明页面 A 和 B 中的内容是序列紧密相关的，这类信息对导引整个用户的行为显得格外重要。

2.2.6 协同过滤技术

协同过滤是目前运用最成功的个性化推荐技术，其核心思想是，返回给用户的推荐信息来自于其他相同兴趣用户历史请求。协同过滤技术被成功的运用到个性化推荐系统中，但是随着系统规模扩大，用户和资源的数量急剧增长，系统的性能会急剧下降。

协同过滤核心是两大问题的解决，第一是用户模型的建立，第二个是用户相似度的计算。主要步骤如下：

（1）数据预处理。主要是指数据清理，去除无效噪声数据，将有效数据根据模型量化成归一化向量以备后续处理。

（2）可以有选择的利用多种聚类算法，如 FCM[29]、K-Mean[30]、AP[31]等产生类别的聚类中心，并获得相关用户在类别中的隶属度矩阵。FCM 算法应用广泛，但是该算法对初始化数据特别敏感，容易陷入局部极小值或者鞍点，而得不到全局最优解，必须事先指定数据集的聚类数，然而聚类个数一般很难预知。K-Mean 算法是硬聚类算法，它将数据点到原型的欧式距离作为优化的目标函数，利用函数求极值的方法得到迭代运算的调整规则，其初始类聚类中心点的选取对结果具有较大的影响。AP 算法不需要事先指定聚类数目，它将所有的数据点都作为潜在的聚类中心，算法通过迭代过程不断更新每一个点的吸引度和归属度值,直到产生若干个高质量的聚类中心，同时将其余的数据点分配到相应的聚类中。

（3）确定关联邻居集合，产生推荐集合。用户 a 和用户 b 之间的相似度 sim(a, b)可以通过公式 2.2 计算：

$$sim(a,b)=\frac{\sum_{u\in U_{ab}}(R_{u,a}-\overline{R_a})(R_{u,b}-\overline{R_b})}{\sqrt{\sum_{u\in U_{ab}}(R_{u,a}-\overline{R_a})^2}\sqrt{\sum_{u\in U_{ab}}(R_{u,b}-\overline{R_b})^2}} \quad \text{（公式 2.2）}$$

其中 $R_{x,y}$ 代表用户 x 和 y 间的距离，$\overline{R}$ 代表用户在类别中和其他用户的平均距离，U_{ab} 代表 a 和 b 所属的共同类别。

2.2.7 基于知识的推荐技术

基于知识的推荐技术在搜索系统中应用广泛。常使用本体论中的层次概念来代表用户和推荐品，并利用相关的兴趣传播算法来推导用户的兴趣集合。Y. Blanco-Fernandez 等人利用语义网上的语义推理技术挖掘出用户和物品之间的内在关联，这些关联提供用户额外的一些偏好倾向，可以帮助系统更加合理和有效地处理用户和推荐品之间的关系[32]。

语义网技术现已经被数字电视内容推荐系统加以采纳，并使用了TV-Anytime、OWL 标准。在艺术推荐系统 CHIP Demonstrator[33]中，利用语义网解决了三个主要问题：（1）冷启动处理问题；（2）推荐多种复杂关系的处理问题；（3）结合历史时间轴、地理博物地图、多角度浏览展示问题。

K. C. Lee 等人提出了一种基于因果图的在线推荐算法[34]，将定性因素如界面设计、购买意向等融入传统协同过滤技术中，评估了推荐系统对用户行为的影响。提升了用户对推荐系统的满意度。

2.3 推荐系统的评估

多数推荐系统采用 J. L. Herlocker 等人提出的推荐精度及推荐覆盖率公式来评估算法的有效性[35]。推荐精度的计算可以基于两种方式：基于统计的和基于决策支持的。基于统计的方法主要是比较物品的预测得分和真实得分之间的偏离，通常的计算方法有 MAE（平均绝对偏差），RMSE（均方根差）、相关系数等。决策支持度量则是判定一个推荐物品有无可能的判断，具体包括常用的查准率、查全率、F-measure、以及 ROC 曲线。覆盖率指的是推荐物品占所有物品的百分比。

2.4 推荐系统的发展趋势

越来越多的推荐系统将物品的具体上下文加以考虑，很多电子商务网站推荐商品的时候会考虑到物品的产地或者促销活动等上下文信息。

利用多维度物品的评分，渐渐形成主流。系统的推荐数据来自用户对物品多维度的评分。比如用户购买商品时，对商品的种类、材质、颜色、价格等做出相应的评价，推荐系统考虑这几个方面各自的权重，合成最终的推荐结果，这一过程可以利用评分生成工具 CollaFis[36]来完成。

现有的推荐系统都需要用户提供显示的反馈，比如给物品评分或者留言，然而实际中，由于用户天生的“惰性”，他们并不积极参与到推荐系统中来，系统获取的数据十分稀疏，因此，推荐系统的发展趋势是转向非侵入式获取用户偏好的技术上去。推荐系统采用用户的点击习惯，以及浏览行为，挖掘从他们对物品

的兴趣度，从而给出相关的推荐物品。

随着 Web2.0 技术不断地普及深入，互联网用户的数量呈几何级增长，且分布分散。为了考虑不同地理位置、偏好的用户不同需求，分布式推荐系统应运而生。在分布式推荐系统中，对于每一项查询请求，系统将请求转发到位于不同地理或者偏好的子查询系统，并给出相应的推荐结果。E. Diaz-Aviles 等人采用一种类似于代理机制的框架，将推荐系统和点对点网络相互关联，并采用积极复制转发传播形式，自组织起一个具有相似兴趣结点的邻居结点[37]。

基于信任的推荐系统也逐渐受到人们的重视。系统建立的理论基础是：具有同一风格、爱好兴趣的用户具有较高的信任的关联度，因此可以通过聚合信任网络中的用户观点来生成相关推荐。目前基于信任的推荐系统已经在社区系统如 LibraryThings[38]和一些拍卖信用推荐系统得到了应用。

2.5 本章小结

本章给出了 CADAL 项目相关介绍，给出了其背景、概括和特点。简略介绍了数字图书馆推荐系统相关开发技术，主要介绍了信息搜索排序相关技术，涵盖了信息抽取技术、全文检索技术、排序技术、信息压缩与传送技术、Web 信息挖掘技术、协同过滤技术、基于知识的推荐技术等，接着介绍了推荐系统的评估方式以及未来推荐系统的发展趋势。

第3章 CADAL 中医药推荐系统总体设计

3.1 推荐系统框架设计

信息搜索推荐模式历经了两个时代。第一代推荐模式是基于关键词的推荐，提供搜索词文字相关的链接，主要以 Google、Baidu 等搜索引擎为代表。第一代搜索推荐方式可以为用户提供字面相关的推荐资料，但是存在信息迷失，推荐信息并非语义关联、信息媒体单一等问题。因此新型的智能搜索已成为需要，第二代搜索推荐模式是基于用户偏好、语义、跨媒体的搜索推荐，此推荐模式更加关注用户的搜索行为，因此也更加智能。

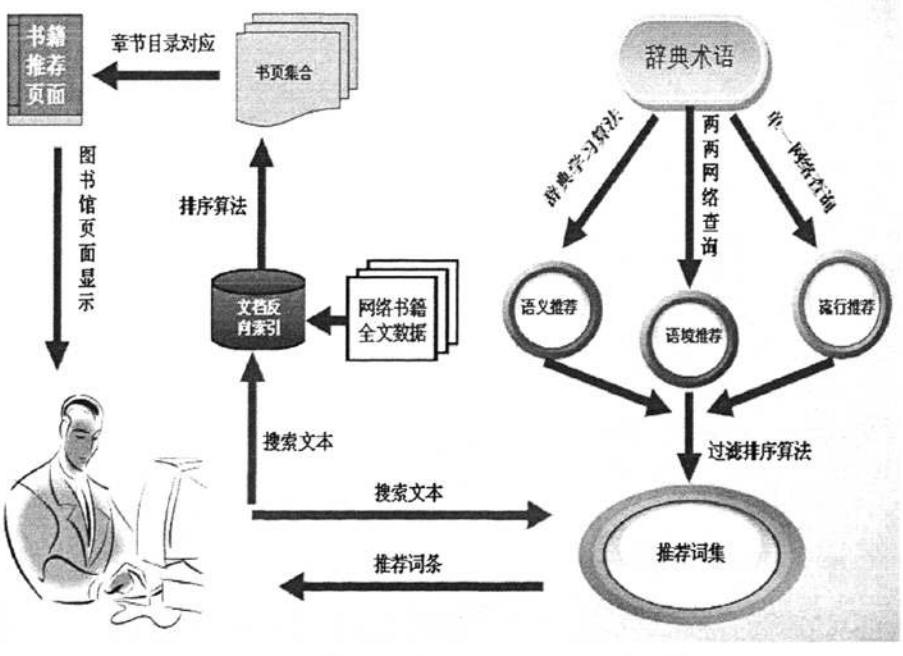

图 3.1 搜索词的推荐书籍页面及其推荐词条实现框架

传统图书推荐系统会在搜索的同时提供字面相关的推荐词语或者图书推荐页面，但是这些推荐词语都是搜索关键词的字面衍生，推荐页面仅仅含有搜索文本而没有先后之分。实际上用户关心的可能是与搜索文本语义相关的知识术语、关

联媒体、或者希望通过图书推荐页面学习到搜索文本的相关描述。因此传统推荐系统满足不了这种学习认知模式。

图3.1是CADAL中医药推荐系统中推荐书籍页面及其推荐词条实现框架，它是一种基于辞典、页面推荐价值评估的中医药术语、页面的推荐框架。在基于后台辞典文本数据的机器学习基础上，结合图书馆用户搜索关注度，经过相似度权值矩阵叠加、阀值聚类、网络热门推荐词语过滤排序等一系列步骤，综合给出初步的推荐词语集合，最后用反向索引机制查找含有搜索文本的页面，并通过页面价值评估函数对搜索结果页面进行合理排序。在推荐集合中考虑用户访问日志，以达到推荐协同过滤、更新和优化的目的。

图3.2描述了方剂、中药、病证、视频之间的关联框架。由于方剂本身是由中药组成，故方剂和中药之间的联系可以通过组成关联予以建立，每一方剂、中药都可以治疗一些病证，故方剂、中药和病证之间的关联可以通过它们的主治关联加以连接。由于视频中含有方剂、中药以及病证的内容，故它们之间的联系可以通过视频语音识别技术加以实现。它们之间的关联一旦建立，无论用户查询哪一方面的信息，系统都可以给出其他各方面的推荐信息。

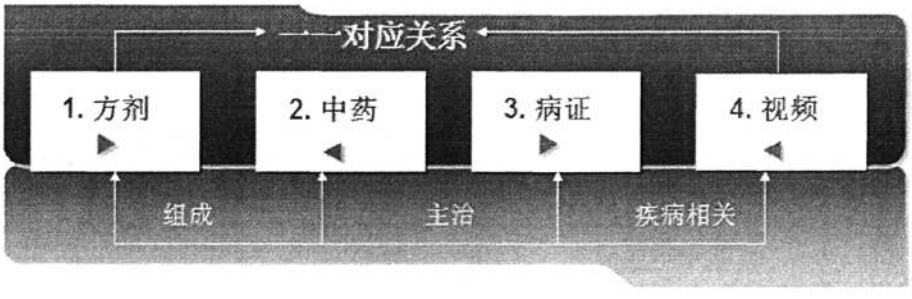

图3.2 方剂、中药、病证、视频关联框架

图3.3描述了系统其他辅助关联推荐框架。该框架包含气候、产地、地图、图片、中药、医馆、名医和病证的关联。多数中药具有其关联图片、产地以及经常使用它们的名医。在某一产地，用户可以找到当地的医馆以及医馆中的历史名医，在地图推荐中，你可以通过具体的地图图片信息查看到当地的地理概况以及风水气候，还具有的联系是一个名医可能擅长治疗某种病证或者使用某种药物。

结合图3.1和图3.2，所有它们之间的联系，可以构成一个推荐知识网络。

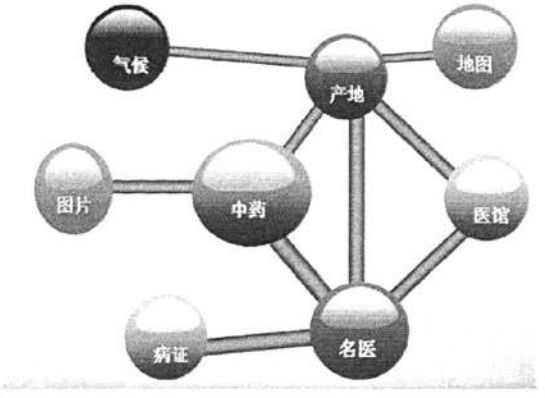

图 3.3 其他辅助关联推荐框架

3.2 建立系统所需任务

为了建立CADAL中医药推荐系统，需要完成以下几项任务：

（1）外部数据采集。本文的数据主要来自两方面：中医药图书的OCR文本数据以及网络多媒体数据。系统从中医药图书OCR文本中对目录和正文进行处理，不仅提取出药物、方剂、病证对应的页码，而且提取出药物、方剂、病证对应的解释内容。本系统设计的网页爬虫程序从专业的中医药网站爬取相关的中医药术语、中医名家、植物学名词、中药图片、视频等相关信息。

（2）数据预处理。采集的数据可能存在信息不一致，不完整等特点。对于空缺的属性值，可能需要人工填写，或者通过使用最可能的值来填充空缺值。方法有基于回归、基于推导的贝叶斯方法、工具或判定树归纳确定。

（3）数据集成和变换。集成将多个数据源中的数据结合在一致的数据存储中，并处理冗余以及相关的数据冲突检测和处理。变换将数据转换成适合于挖掘的形式，可能涉及平滑、聚集、规范化和属性构造等过程。

（4）用户界面设计。友好美观的用户界面有助于用户浏览兴趣的提高以及整个网站点击量的增加。界面设计需要既要遵循简洁又要考虑到信息展示的充实性，既要考虑到界面的美观又不能过于绚丽。因此整体的界面设计需要采用折中的原则，达到查询操作简单，页面展示美观等特点。

（5）推荐的产生及其反馈。通过系统机器模块学习，系统中每一中医药术语都将得到其最初的推荐集合，此集合通过文档的更新、用户行为数据的反馈不断迭代获得更新。

（6）用户数据获取及处理。用户数据来自用户推荐书籍点击日志以及用户浏览日志。日记对推荐词条、页面点击量进行统计，经过一段时间后，根据点击量和浏览量的大小，后备推荐词条、页面将获得特定替换概率，替换点击量较少的推荐词条，最后返回更新后的推荐结果。

3.3 中医药数据采集方法

整个数据采集分为书籍数据采集和专业网站数据采集。书籍数据采集又可以分为书籍目录信息采集和 OCR 文本信息采集。专业网站数据采集指的是从特定中医药网站上获取有关中医药术语、多媒体信息。

3.3.1 书籍数据采集

CADAL 一期项目中，纸质书籍的全文通过扫描设备扫描后，扫描成 djvu 格式的电子文档，此电子文档再通过 OCR 识别软件识别后，成为 txt 格式的文本文档，待人工校对后，就可以作为书籍的全文数据加以利用。书籍目录的录入工作已在一期的项目完成，形成一个目录 XML 文档存放在服务器中。目录中包含关键词的页码所在的页码信息，可以作为页码索引加以利用。

推荐系统的目标是依据中医数字图书知识库，提供中草药、方剂和病证资料的结构化信息、交叉信息检索推荐和多媒体检索推荐服务。基于这一目标，系统主要可以分为四个步骤：第一步，进行数字图书目录的处理，实现实体信息定位。主要需要处理的是第二步，对已经定位的数字图书内容进行基于正则模式特征粗粒度文本信息分析，根据分析出的文本特征在数字图书信息中进行特征检索，获取在语义上符合样本文本特征的文本字段。根据这些字段形成原始本体库。第三步，根据形成的结构化信息，寻找关联项，比如方剂中的组成含有中药，是一对多的关系，方剂包含主治，主治中含有各种病证，也是一对多的关系。中药里的

应用字段中包含各种病证，对病证也行成了一对多的关系。根据这些关系，生成中医信息交叉检索知识库。第四步，利用形成的本体库进行图书信息标注，标注后各种信息形成一个聚合体，这些聚合体为语义检索提供了进行语义检索支持。

图 3.4 描述了整个书籍数据采集的流程，提取内容信息在一定程度上可以通过提取目录信息来实现。中药书籍的目录往往记载药名及其归类，方剂书籍主要记载的是方剂的名称和归类，诊断、病证的书籍记录了各种诊断手段以及病症的名称和病证表象。利用这些元信息，可以十分方便的建立起药物、方剂、病症的本体数据库框架，再利用图书目录的页面信息关联到的页面信息，采用 OCR 技术识别出该页面上的文字，以各种信息抽取手段抽取文本中的实体属性，就可以填充本体数据框架的内容。

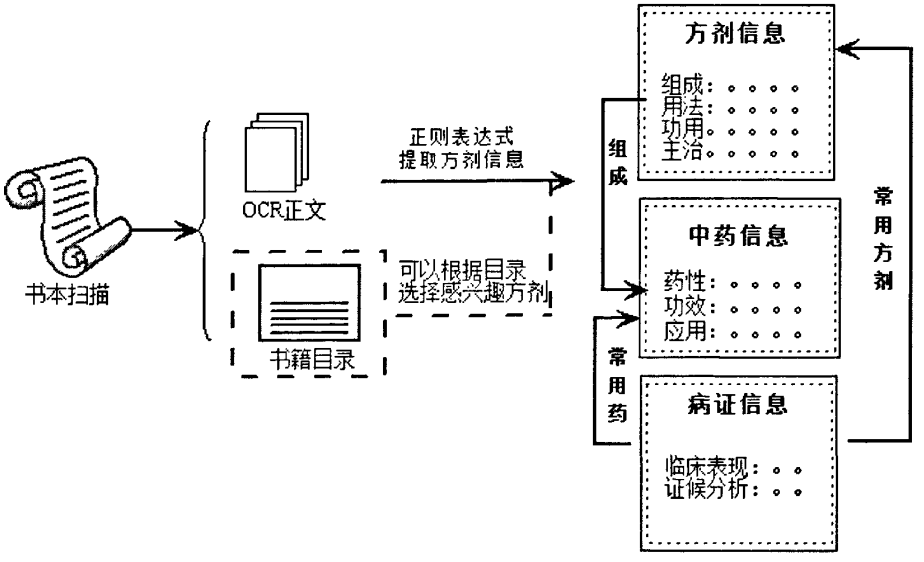

图 3.4 书籍数据采集的流程

在处理目录的时候，如果发现新的名词术语条目，需要进行页面跟踪，定位到所在的上下文相关信息，对其进行文本分析，进而自动归纳出文本特征，进行信息提取，然后记录该信息所在的书籍页号；若处理的目录条目和以往的相同，只需记录该条目所在的书本页号即可。

系统对典型的中草药、方剂以及病症的描述文本进行分析，获取这些文本的特征信息，利用描述文本特征信息在数字图书中搜索相匹配的页面（段落），从而获取符合特定语义特征的描述页面。

系统通过交互式的图形界面，分析中医图书的语言习惯和专业背景，得到信息实体的书写模式，然后根据此模式对海量数字图书信息中进行文本特征匹配，从中检索出具有中草药、方剂以及病症的实体信息。

对提取出来的信息，可以做进一步细化。比如方剂的组成信息，可以从方剂组成中提取出方剂组成中提取药物、制法、现代用量和古用量。系统采用基于NPOS模型分词技术，对组成文本进行分词，然后根据分词结果中药物名词、动词、量词的各种特殊情况做一一处理，进而提取出相关的结果。

还可以依据自然语言处理中的基于规则的处理方式提取中药的化学成分。如果将常见的化学成分名称存入数据库，然后对文本进行比对，如果符合特定的结构，那么提取之。但常见的化学成分名称已经十分复杂。而且即使存在这么一个常见化学式名称的数据库，对于新的、不常见的化学式来说必然提取效果不佳。必须挖掘其书写模式，然后进行提取。

3.3.2 中医药数据网络采集

许多中医药古籍正文，由于都是珍籍善本，存在OCR识别问题，故正文需要从网上间接获取。再则，中药图片、中医药视频都不能直接通过书籍获取，也需要通过设计网络爬虫采集。

针对部分网站资源丰富，且网页内容结构具有统一规范的上下文结构，故可以设计基于正则表达式的多线程爬虫来获取此类多媒体资源。可以针对特定网页制定个性化的抽取规则，支持网站跨层采集、POST采集、脚本采集、动态页面采集等功能。数据采集后，可以直接通过资源导入界面，导入到数据库或者相关的文本文件、XML文件中。图3.5描述了整个抽取流程：通过分析网站的上下文信息后，手工编写抽取规则，接着进行数据提取，然后对数据进行发布。

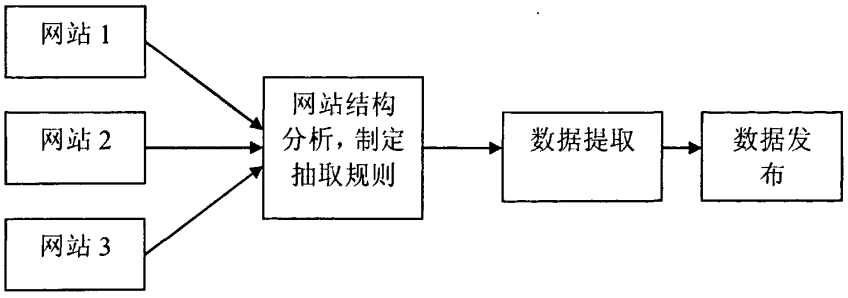

图 3.5 网站信息抽取

3.4 中医药数据集成和变换

3.4.1 数据集成

数据集成时，需要将多个异构数据库、信息源结合起来存放在一个统一的数据库中。这些异构信息源可能来自不同的地理位置的文件或者不同结构的数据库中。

集成时，需要处理实体匹配问题，远程数据库中的药物编号的字段名称可能和本地数据库中药物编号的字段名称不同，需要制定名称匹配表加以解决。数据冗余是另一个需要处理的问题。在数据抽取时，很多信息存在重复抽取的问题，需要将重复项去除。某些冗余可以通过冗余相关检测技术加以检测。属性 A 和属性 B 的相关性 $r_{A,B}$ 可以通过公式 3.1 度量：

$$r_{A,B}=\frac{\sum(A-\bar{A})(B-\bar{B})}{(n-1)\sigma_A\sigma_B} \qquad \text{（公式 3.1）}$$

其中 n 是元组个数，$\bar{A}$是 A 的平均值，$\bar{B}$是属性 B 的平均值。σ_A、σ_B代表属性 A 和 B 的标准差，如果 $r_{A,B}$ 大于 0，则代表它们是正向相关的，因此 A 和 B 有可能存在冗余，需要去除其中之一。如果 $r_{A,B}$ 等于 0，那么可以说明 A 和 B 是独立的，如果 $r_{A,B}$ 小于 0，则代表负相关，一个值随着另一个值减少而增加。

集成时还需要考虑数据值的冲突及其处理，比如同一方剂名称，可能不同书籍具有不同的描述，在统一存储时就可能存在冲突，药物的剂量如果在单位上以

不同的方式表达，可能导致属性标注的不一致性，因此在数据导入时需要通过程序脚本外加人工方式加以校对。

3.4.2 数据变换

数据变换是对数据的一次适应性转换，即转化成更加适合挖掘或者展示的形式。转换可能涉及到以下过程：

（1）平滑：收集的中医药文本、多媒体数据以及从用户采集的评分数据可能存在噪声，需要加以去除，需要用分箱、聚类或者回归的方式加以解决。

（2）聚集：数据收集后，需要进行汇总和聚集。比如对每个月的点击量日志进行分析，计算月平均和日平均点击量，由此可以为多粒度数据分析构造数据立方体，进一步进行挖掘和回馈改进。

（3）数据概化：可使用概念分层对数据进行从低层次到高层次之间的转化。比如“风寒感冒”和“风热感冒”可以统一转换成“感冒”等。

（4）规范化：将某些数值进行按比例缩放，使之落入到一个规范化的空间内，一般是将其落入单位化空间[-1.0~1.0]之间。

（5）属性构造：可以在已有属性集的基础上添加新的属性，比如说将药性量化，将量化的矢量作为新的属性加入到药物的属性集合中。

3.5 搜索引擎界面设计

界面的设计需要考虑如何在展示中医药信息基础的同时，提供对于中草药、方剂以及病证的交叉检索和多媒体检索推荐服务。通过数据关联和知识关联，将中医语义信息知识库同已有的中医研究数据库以及动植物研究材料相关联，开发具有系统性、完备性的多层面个性化的中医信息检索方式。

在此基础之上，界面应首先考虑到各种用户的需求，采用快速原型的开发方法，给出一个基本界面，而后在此基础上不断改进和扩充。

系统将整个界面设计分为结构设计、交互设计、视觉设计三个部分。结构设计也称为概念设计，是界面设计的骨架。通过对用户研究和任务分析，制定出产

品的整体架构。整体架构可以通过Adobe公司开发的Dreamweaver软件完成，系统采用简洁的层次结构，主界面如图3.6所示：用户可以选择中药、方剂、病证三个方面的其中之一进行信息搜索，系统色调以蓝色为主，与操作系统软件界面的默认色彩相吻合，若使用与之差别过大的色彩，比如玫瑰红、土黄等，那么色彩的强烈变化会影响用户的使用情绪，甚至会引起反感。

交互设计的目的是让产品功能更加简单使用。由于引擎功能的实现都是通过人和界面的交互来完成，因此，以下交互设计的原则在界面中需得以体现。

（1）有清楚的错误提示。误操作后，系统提供有针对性的提示。当输入词为空或者无效时，系统会给出输入词无效的提示信息框。

图 3.6 检索主界面

（2）让用户控制界面。“下一页”、“返回”，面对不同层次提供多种选择，给不同层次的用户提供多种可能性。检索界面给出的结果采用分页的形式给用户跳转的机会。并在次级页面设置导航条功能，使其快速回到需要的页面。

（3）允许兼用鼠标和键盘。同一种功能，同时可以用鼠标和键盘。提供多种可能性。在此原则上，界面提供了“TAB”键按钮对页面控件进行切换，提供了回车键对表单提交和链接转发功能的支持。

系统开发了中药检索条目的横向检索功能，由中药可以检索到相关的产地，由产地可以检索到产地的气候，由气候可以检索到相关的药物生长条件，制药方法，由产地可以检索当地所在的中医院、馆所、博物馆、当地历史名医等。由历史名医可以检索出名医著作，学术论文和学术观点，进而可以构成一个知识网络。

高级用户可以定制检索引擎，比如对用户偏好进行引擎结果筛选，只返回某特定属性的结果集。根据用户的不同层次的需求提供各种个性化的信息服务，采用相应的搜索匹配引擎。用户可以对结果集进行反馈，从而改进结果集，做到协同过滤的效果。

在视觉设计上，系统使用了一些常见的比喻，力求符合用户的习惯，比如页面的载入就采用了类似进度条的 gif 动态图片来展示页面载入需要一定时间的事实。当鼠标滑过选定区域时，界面将使其高亮显示，符合人的直觉并在客观上给予用户一定的提示。

3.6 推荐的产生以及反馈

推荐的最初结果集是由机器学习模块产生的。该模块作为整个推荐框架的后台数据获取源，负责计算辞典中词条语义相似度和词条文档关联度，以及抓取网络热门推荐词条等工作。综上所述，机器学习模块包括辞典学习、网络文档学习、网络词条学习等 3 个子模块，各模块的主要作用如表 3.1 所示。

表 3.1 机器学习各子模块的作用

子模块	作用
辞典学习	通过比对词语的解释字段得到词语间的语义相似度。
网络文档学习	网络文档学习关注术语间的语境信息，语境信息反映了两个词语同时在文档中出现的频率相关性。
网络词条学习	在以某一词语作为搜索词时，将该搜索词的字面衍生词作为推荐词语。

辞典学习基于以下事实：通过词语字符比对衡量词语间的语义相似度一般并不准确，表面相似的词语可能意义完全不同，如“鱼”和“木鱼”。本文提出的辞典学习方法建立在辞典术语含义的基础之上，通过比对词语的解释字段得到词语间的语义相似度。

网络文档学习关注术语词条间的语境信息。语境信息反映了两个词语同时在文档中出现的频率。本文采用了三大搜索引擎中的数据，具体实施方法是：对辞典中的所有名称字段在谷歌、百度、雅虎分别做 C_n^2+n 次搜索操作，记录查询结

果总数，其中C_n^2次为两个词语共同查询操作，n 次为单个词语查询操作。仿照向量余弦计算公式得到其相似程度。

网络词条学习是指在以某一词语作为搜索词时，将该搜索词的字面衍生词作为推荐词语，此方法是目前各大搜索引擎采用的方法。该方法考虑了网络用户基于关键词扩展搜索项的关注程度。具体实施细节是：将用户搜索过的所有搜索词记录在日志中。推荐过程实际上是一次日志的排序过程，对搜索记录日志中含有当前搜索词的词语，按搜索次数进行排序，并设置阀值提取排序靠前的记录作为推荐词集合。本文将辞典中的每一术语作为关键词在三大引擎中进行搜索，记录各自的推荐集合为 A1，A2，A3。滤去 A1，A2，A3 那些在中医图书搜索引擎中搜索结果为 0 的词语。对 A1、A2、A3 中的推荐词语进行迭代，如果推荐词语∈A1 ∩ A2 ∩ A3，则设置其排序值为 3；如果推荐词语∈（A1∩A2 | A2∩A3 | A1∩A3），则设置其排序值为 2；其他情况设置推荐词语的排序值为 1。设 A4=A1 U A2 U A3，则将 A4 中的推荐词语按排序值从高到低排序后便得到推荐列表。

3.7 用户数据获取及处理

在此模块中，本文通过分析用户对推荐词的点击流日记，获取用户对推荐词语、页面的满意度。在公式 3.2、3.3 定义满意度函数为：

$$f(word)=\sum_{i=1}^{5天}(i天word点击次数+i天word点击IP总数) \qquad （公式 3.2）$$

$$f(page)=\sum_{i=1}^{10天}(i天page点击次数+i天点击IP总数+i天停留时间) \qquad （公式 3.3）$$

其中 word 代表推荐词，page 代表推荐页面，经过相关的天数天后，统计各推荐词 f(word)和 f(page)的值。在公式 3.4 定义过滤算子 filter 为：

$$filter(item_i)=\frac{f(item_i)}{\sum_{i}^{n}f(item_i)} \qquad （公式3.4）$$

其中 n 代表推荐词的总数，若 filter(item)<$1/n^2$，则说明该推荐项的点击率比平均点击率的 1/n 还要小，不受用户关注，需要用后备集合中推荐项加以替换，

在公式3.5定义替换概率P(word)为：

$$P(item)=\frac{n-m+1}{\sum_{i=1}^{n} i}(1-\varepsilon) \qquad \text{（公式3.5）}$$

其中n代表后备推荐词集合搜索词语word和推荐页面page总数，m代表推荐项item在推荐集合中的排名名次。其中1-ε代表此替换概率为一个松弛参数，系统可以对此概率进行调整。机器学习模块后备推荐词、页面集合中的元素以此替换概率替代filter<$1/n^2$的词语和页面。更新搜索词的推荐列表后，并将f(word)值清零。被替换掉的词语置放在各自推荐词列表的末尾以供再次使用。由于含有搜索关键词的网络文档数、网络热门推荐词语不断变化，所以网络文档学习和网络词条推荐学习部分的结果需要不断更新。更新实质是一次重新机器学习过程。

3.8 本章小结

本章讨论了CADAL中医药推荐系统的总体设计，首先介绍了中医药数据采集的方法，其中涵盖书籍数据的采集以及中医药数据的网络采集。接着描述了中医药数据集成和变换的一些注意事项和一些细节。然后讨论了系统界面设计必须遵循的一些设计原则以及界面设计的禁忌，接着本章详细叙述了推荐是如何产生以及其反馈机制，最后陈述了用户数据获取和处理的一些理论公式设计情况。

第4章 CADAL中医药推荐系统技术设计

4.1 基于正则表达式的实体抽取算法

正则表达式是用一个“字符串”来描述一个特征，然后去验证另一“字符串”是否符合这个特征。比如表达式“ab*”描述的特征是：一个‘a’和任意个‘b’。例如‘a’、‘ab’、‘abbbb’都符合这个特征。

正则表达式信息提取方法广泛适用于格式良好的信息提取中，由于选择的书籍是中医药教科书和由循环语句生成的网页，故符合这种提取要求。在提取上，可利用书籍、网站的编排格式，例如书中方剂信息的书写形式为：方剂名（一段中文）+组成介绍（组成关键字+分隔符+中文）+功用介绍（功用关键字+分隔符+中文）+其他信息（其他书写格式），那么上述书写规则就可以用正则表达式加以表述，自动提取方剂所在页面的方剂组成信息。

图 4.1 对应于提取书籍信息模式，该模式是细粒度下的书籍信息提取：用户只需要选择组成和功用作为关键词，并选择方剂名作为主题词，系统分析文本模式后，生成相关的正则模式做全文单元划分，然后并可以提取方剂名、组成和功用。在图中，所有提取出来的信息呈现绿色。系统根据用户交互的内容，分析出主题句的模式以及需要提取的关键字，并根据主题句和关键字组成一个提取模式，这种提取模式本质上是一个正则表达式，这种提取模式将在全文中做模式匹配，先根据主题句做全文分块，若主题句之间没有匹配的关键字，则寻找下一匹配句。匹配后根据关键字，做关键字内容信息提取。所有匹配的内容都将用高亮颜色加以表示。提取信息可以存放在关系数据库的表中。

信息一旦提取出来，就可以作为结构化的实体存储于数据库中，以便以后相关的处理。

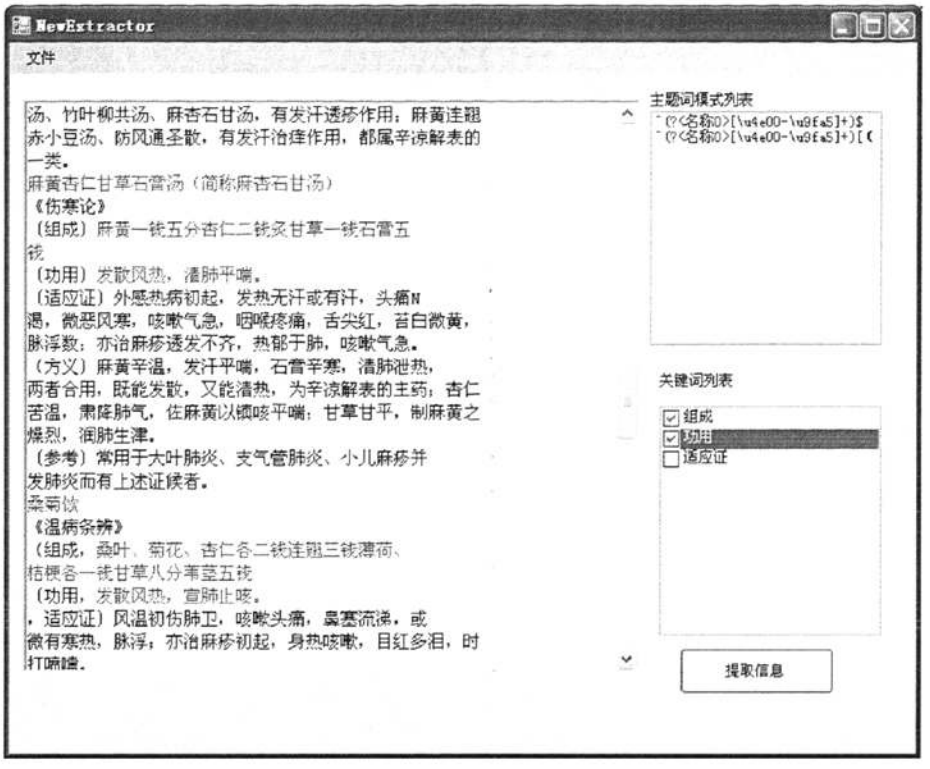

图 4.1 高效提取模式提取方剂图

4.2 基于双数组 Trie 树的信息分离算法

信息抽取后，不同实体之间的关系尚需建立。比如方剂和药物之间的关联，药物和病证之间的关联等。因此，需要对书籍和网络中抽取的信息进行进一步分离提取。

系统采用中医药辞典建立中药、方剂、病证词库，辞典的内容包括近 3000 条中医术语及其解释、11000 条中药名词及其功效和应用说明、2200 多条方剂名称及其组成与功能主治信息、2100 多条病证名称。例如中药麻黄，其主治字段字符串为：“1. 风寒感冒。本品味辛发散，性温散寒，主入肺与膀胱经，善于宣肺气、开腠理、透毛窍而发汗解表，发汗力强，为发汗解表之要药……”，如果该字段中含有辞典中的病证术语，那么就简单认为药物和该病证相关。但是由于主治长度平均具有 500 以上的字符，再者辞典规模过大，如果简单地采用和主治辞

典术语一一匹配的策略，势必影响到分离效率。因此，系统采用了基于双数组Trie树分词算法[39]对信息进行分离。

Trie树这种数据结构在信息查找时广为采用。对于给定字符串S，若S的长度为n，则查找匹配指定字符子串的最坏复杂度是O(n)，而与辞典的规模无关，但是空间利用率度过低，很多树枝没有充分利用。双数组Trie树是Trie树的变形和改进，它引入了DFA（有限状态机）到查询状态的判断中，所有的Trie结点保存在base和check两个数组中。Trie树中的节点作为DFA的状态，使用base数组来确定状态转移，并使用check的数组来检验转移的正确性，数组的长度是Trie节点的数目。

设数组下标为i，如果base[i]，check[i]均为0，表示该位置为空。如果base[i]为负值，表示该状态为词语。check[i]表示该状态的前一状态，t=base[i]+a，check[t]=i。

下面举例来说明用双数组 Trie 构造分词算法词典的过程。假定词表中只有“疫、阴虚、阴阳、阴阳转换、人参、人中”这6个词，首先对词表中所有出现的9个汉字进行编码：#=1，疫=2，阴=3，虚=4，阳=5，转=6，换=7，人=8，参=9，中=10。“#”代表终结符，其节点数为12个，故数组长度为12。对于每一个汉字，需要确定一个base值，使得对于所有以该汉字开头的词，在双数组中都能容纳。

表 4.1 base和check数组

数组下标	1	2	3	4	5	6	7	8	9	10	11	12
base[i]	1	-6	-1	1	-2	-4	0	0	1	-9	-11	0
check[i]	0	4	1	1	4	4	0	0	1	9	9	0
词缀	#	虚	#	阳	#	转	换	#	参	#	中	#

例如，现在要确定“阴”字的base值，假设以“阴”开头的词的第二个字序列码依次为a1，a2，a3……an，找到一个值i，使得base[i+a1]，check[i+a1]，base[i+a2]，check[i+a2]……base[i+an]，check[i+an]均为0。一旦找到了这个i，“阴”

的 base 值就确定为 i。用这种方法构建双数组 Trie，经过几次遍历，将所有的词语放入双数组中，然后还要遍历一遍词表，修改 base 值。因为我们用负的 base 值表示该位置为词语。如果状态 i 对应某一个词，而且 Base[i]=0，那么令 Base[i]=-i，如果 Base[i]的值不是 0，那么令 Base[i]=-Base[i]，得到双数组如表 4.1 所示。Trie 树的查询过程其实就是一个 DFA 的状态转移过程，在双数组中实现起来比较简单：只需按照状态标志进行状态转移即可。如图 4.2 所示，在查询“阴阳转换”时，先根据“阴”的序列码 b=3，找到状态“阴”的结点 4，再根据“阳”的序列码 d=5 找到“阴阳”的结点 base[b]+d=6，同时根据 check[base[b]+d]=4，表明“阴阳”可以是某个词的一部分，可以继续查询。base[4]+1=2，然后再找到状态“阴阳转换”。表明“阴阳转换”在词表中，查询完毕。

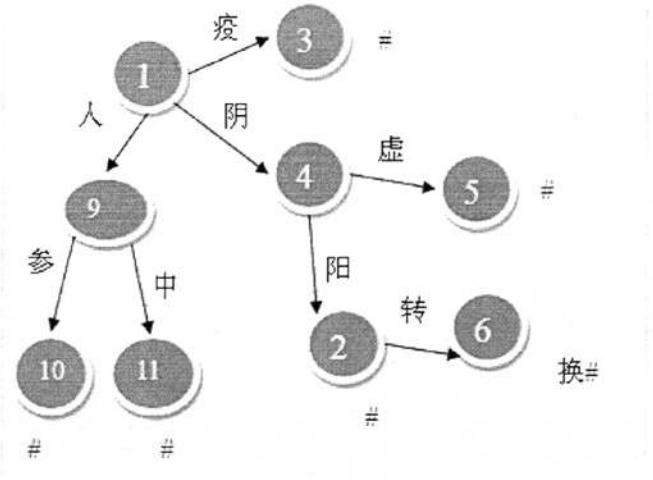

图 4.2 查询结点转换

图 4.3 是系统提取药物主治的截图，使用该算法后，系统能在数秒之内计算出上万种中药的所有主治功能。

	medicine	medicinenam	application
▶	1	麻黄	1. 风寒感冒，本品味辛发散，性温散...
	2	荆芥	1. 外感表证。本品辛散气香，长于发...
	3	防风	1. 外感表证。本品辛温发散，气味俱...
	4	柴胡	1. 表证发热及少阳证。本品辛散苦泄...
	5	升麻	1. 外感表证。本品辛甘微寒，性能升...
	6	生地	1. 热入营血，舌绛烦渴、斑疹吐衄。...
	7	丹皮	1. 温毒发斑，血热吐衄。本品苦寒，...
	8	苍术	1. 固湿中焦证。本品苦温燥固以祛湿...
	9	白术	1. 脾气虚证。本品甘苦性温，主归脾...
	10	附子	1. 亡阳证。本品能上助心阳、中温脾...
	11	干姜	1. 腹痛，呕吐，泄泻。本品辛热燥烈...
	12	当归	1. 血虚诸证。本品甘温质润，长于补...
	13	熟地	1. 血虚诸证。本品甘温质润，补阴益...
	14	火麻仁	肠燥便秘。本品甘平，质润多脂，能润...
	15	黄芩	1. 湿温、暑湿、胸闷呕恶，湿热痞满...
	16	陈皮	1. 脾胃气滞证。本品辛行温通，有行...
	17	麦冬	1. 胃阴虚证。本品味甘柔润，性偏苦...
	18	人参	1. 元气虚脱证。本品能大补元气，复...
	19	生姜	1. 风寒感冒。本品辛散温通，能发汗...
	20	大黄	1. 积滞便秘。本品有较强的泻下作用...
	21	石膏	1. 温热病气分实热证。本品性味辛甘...
	22	川芎	1. 血瘀气滞痛证。本品辛散温通，既...

	编号	药物名	主治疾病
▶	1	麻黄	风寒感冒
	2	麻黄	风寒
	3	麻黄	外感风寒
	4	麻黄	喘
	5	麻黄	咳嗽
	6	麻黄	伤寒
	7	麻黄	咳嗽气喘
	8	麻黄	肺热
	9	麻黄	风水水肿
	10	麻黄	风
	11	麻黄	肺失宣降
	12	麻黄	水肿
	13	麻黄	小便不利
	14	麻黄	脾证
	15	麻黄	阴虚
	16	麻黄	痰咳
	17	荆芥	外感表证
	18	荆芥	风
	19	荆芥	风寒
	20	荆芥	风热
	21	荆芥	风寒感冒
	22	荆芥	咳嗽

图 4.3 药物主治提取

4.3 方剂组成信息的细化

方剂的组成信息的细化可以利用以下算法实现，该算法流程如下：

（1）将方剂组成文本进行分词，标注词性。

（2）以药物名词为分界点、整理这些数据，将名词、动词、数词、量词归为一类。

（3）分析词段，将药物名词作为药物字段、动词作为制法字段、数词和量词合并称为用量字段，在剂量的单位考虑上，由于两、斤、升、钱等量词都是古代的量词，而克、千克、公斤、g、kg、l、ml等是现代量词，需要加以区别以古用量、现用量加以入库。

一个简单的例子如麻黄汤，它的组成用斜体表示为：“麻黄去节，三两(9g)，桂枝去皮，二两(6g)，杏仁去皮尖，七十个(6g)，甘草炙，一两(3g)”。

将此文本进行分词，并滤去标点符号后，可以切分为：麻黄（名词）、去节（动词）、三（数词）、两（量词）、9（数词）、g（量词）、桂枝（名词）、去皮（动词）、二（数词）、两（量词）、6（数词）、g（量词）、杏仁（名词）、去皮尖（动词）、七十（数词）、个（量词）、6（数词）、g（量词）、甘草（名词）、炙（动词）、一（数词）、两（量词）、3（数词）、g（量词）。

以名词为分界点、整理这些数据，将名词、动词、数词、量词归为一类，上述文本由此可以划归为4个语段：①麻黄（药名）、去节（制法）、三两（古用量）、9g（现用量）；②桂枝（药名）、去皮（制法）、二两（古用量）、6g（现用量）；③杏仁（药名）、去皮尖（制法）、七十个（古用量）、6g（现用量）；④甘草（药名）、炙（制法）、一两（古用量）、3g（现用量）。这样处理后就可以得到麻黄汤的组成药物和它们的用法用量了。

还有以下几种情况需要加以考虑：

（1）语段的量词后置。如方剂玉屏风散，它的组成为："防风一两(30g)　黄芪蜜炙　白术各二两(60g)"，在分词后，用分词算法标注出各个词语的词性后，发现如果直接以药物名词为分界点，则原句子分为三段："防风一两(30g)"、"黄芪蜜炙"、"白术各二两(60g)"，如果直接入库，那么黄芪的用量将不能获取，于是修改算法原始流程，以名词为分界点分割成语段后，如果发现某语段没有量词，则搜索后面的语段，如果发现"各"这个代词，在此处，缺失的量词在第三个语段中发现，复制此语段所在的用法用量，放于"黄芪蜜炙"的后面，拼接成新的语段"黄芪蜜炙二两(60g)"。

（2）某一语段中存在多个名词。在药方的提取中，这类情况也比较常见，比如方剂安宫牛黄丸，它的组成含有犀角，写为"犀角（水牛角代）"，此水牛角为犀角的替代品，那么同一语段将含有2个药物名词，提取的时候将出现误差。因此，算法采取如下的策略，在分段前，先忽略括号内的名词，然后再以括号外的名词为分段点进行分段，如果一个语段中出现2个或2个以上的名词，简单的以第一个为准入库，当然也可以调整为第二个为准。

（3）一个语段存在多个动词。比如《景岳全书》记载的方剂左归丸，某一语段为"鹿角胶敲碎，炒珠，四两（120g）"，其实里面的2个动词敲碎、炒珠都是动词，所以只要把他们都作为制法入库即可。如果动词后紧跟形容词，或者形容词后紧跟动词，比如"捣细"、"微炒"等，可以将其合并归类为动词入库。

（4）一个语段存在多个量词。比如龙胆泻肝汤，某一语段为"龙胆草一钱五分炒"，按照原始算法，"五分"将作为古用量入库。所以在发现"一钱"和"五

分”两个都作为古用量后，必须将其拼接成一个整体加入处理。

（5）语段中量词模糊。比如济川煎，某一语段为“升麻五分至七分或一钱(1.5~3g)”，这里升麻的用量不确定，算法将“五分至七分或一钱”作为整体存入古用量字段中，将“1.5~3g”作为另一整体存放如现用量中。

综上，提取算法流程如图 4.4 所示，先分词，再分语段，接着语段处理，最后提取信息入库。

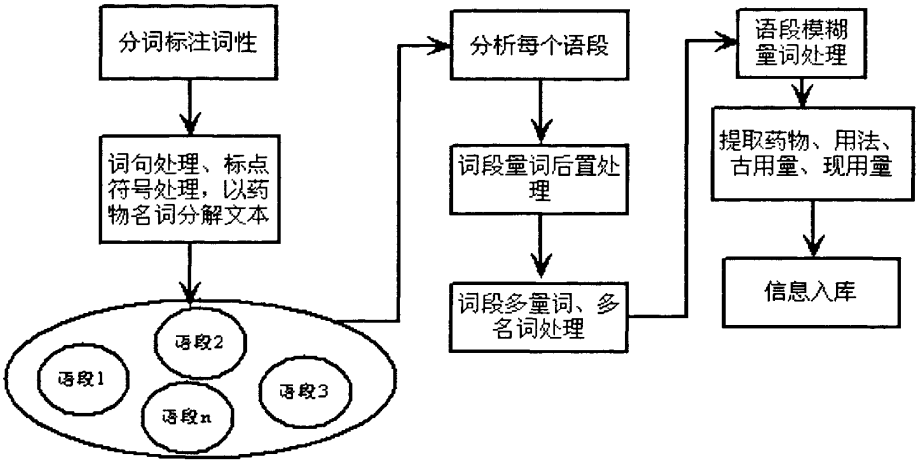

图 4.4 提取算法流程图

实验选用国内中医药大学普遍选用的普通高等教育“十五”国家级规划教材《方剂学》为例子，共 362 首，其中正方 182 首，附方 180 首。提取它们的组成字段，运行上述算法，提取药物、用量的正确率达到 99.2%，制法的正确率为 99.7%，下图 4.5 是提取结果。

id	方剂名	药物	制法	古用量	现用量
44	生脉散	人参	<NULL>	五分	9g
45	生脉散	麦门冬	<NULL>	五分	9g
46	生脉散	五味子	<NULL>	七粒	6g
47	玉屏风散	防风	<NULL>	一两	60g
48	玉屏风散	黄芪	蜜炙	二两	60g
49	玉屏风散	白术	<NULL>	二两	60g
50	茵陈蒿汤	茵陈	<NULL>	六两	180g
51	茵陈蒿汤	栀子	<NULL>	十四枚	12g
52	茵陈蒿汤	大黄	去皮	二两	60g

图 4.5 提取药物后的部分结果

4.4 语义推荐算法

本文先将中医术语的解释、中药功效和应用、方剂组成和功能主治信息进行NPOS最短路径中文分词[40]。所谓NPOS最短路径即为所有组词方案中可能性最大的组合方案，也就是各组合间权值最小的路径方案。组合的权值按公式4.1计算。

$$weight = -\log_{10} P(w_i = w \mid c) \qquad \text{（公式4.1）}$$

其中，weight代表词语在具体语境下的权重，等式右边是词语概率的负对数，由此可见，词语在语境中出现的概率越大，权值越小。在分词进行前，将所有中药名称、方剂名称及中医名词术语添加到词库，标记词性为n；将中医术语的动词如“相生”、“相克”、“相侮”等，标记词性为v；将中医名家人名词汇标记词性为nr；将书籍名词标记为nb。按照公式4.2所示的NPOS概率公式算出词语w_i在语境中的概率：

$$P(w_i = w \mid c) = \sum_{g_j \in G} P(g(w_i) \mid g(w_{i-N+1}^{i-1})) \times P(w_i = w \mid g(w_i) = g_j) \qquad \text{（公式4.2）}$$

公式中的c代表某一具体语境，w代表词语，g代表词性，G代表词性的集合。等式左边P表示词语w_i在整个语料库中的概率。右边表示一个词的词性出现的概率，条件依赖于前N-1个词的词性（$P(g(w_i) \mid g(w_{i-N+1}^{i-1}))$），而词语本身的概率又依赖于该词语所属的词性（$P(w_i = w \mid g(w_i) = g_j)$），累加号表示词语在具体语境中可能有不同的词性（$\sum_{g_j \in G}$）。

比如有以下的文本：“*大黄半夏*”，通过计算字和字，词与词的关联度，可以画出图4.6的有向图，此有向图的边上的数字代表两个字或两个词之间的权值，比如边“大半”表示词语“大黄半”，这个词语的距离是12.10，边“大黄”的距离为7.30，表示“大黄”这个词的距离是7.30，比“大黄半”这个词的可能要大些。始、末两个节点只是为了分词便利而设置的工具节点，并无具体的意义。利用NPOS最短路径算法可以得到最短路径为1，3，5，6，路径和为27.69。

对相关字段分词得到的结果只保留动词、名词、书籍、人名等主题词，原因在于中医、中药、方剂辞典术语的解释往往是主谓宾结构，并可能含有对某医学家著作的引用，或提及某医学家的名字，故句子的中心意思由这些词语决定。然后，对主题词按拼音字母进行排序，并只对词性相同的词语进行相似度计算。排序是为了方便后续词组比较算法的进行，词性相同的词语才进行相似度计算的原因在于字面相同的词语可能具有不同的词性，且比较不同词性的词语会对词条语义计算产生偏差，符合非同类不可比逻辑。

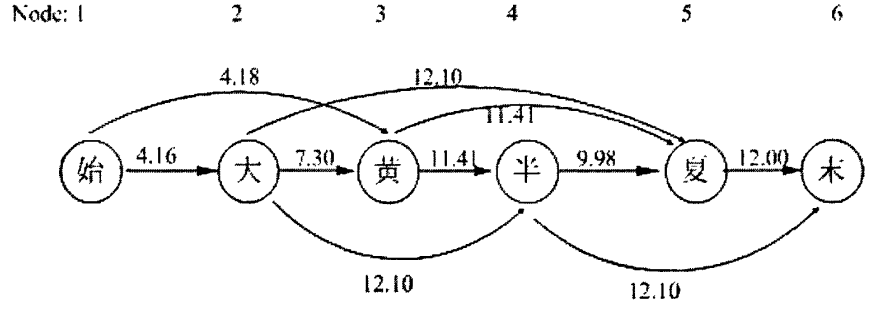

图 4.6 大黄半夏分词图

对相关字段分词得到的结果只保留动词、名词、书籍、人名等主题词，原因在于中医、中药、方剂辞典术语的解释往往是主谓宾结构，并可能含有对某医学家著作的引用，或提及某医学家的名字，故句子的中心意思由这些词语决定。然后，对主题词按拼音字母进行排序，并只对词性相同的词语进行相似度计算。排序是为了方便后续词组比较算法的进行，词性相同的词语才进行相似度计算的原因在于字面相同的词语可能具有不同的词性，且比较不同词性的词语会对词条语义计算产生偏差，符合非同类不可比逻辑。

在词语相似度计算中，如公式4.3、4.4、4.5所示，采用Aminul Islam等的最长公共子序列LCS三种修改算法[41]：

$$v_1 = NLCS(r,s) = \frac{length(LCS(r,s))^2}{length(r) \times length(s)} \quad \text{（公式 4.3）}$$

$$v_2 = NMCLCS_1(r,s) = \frac{length(MCLCS_1(r,s))^2}{length(r) \times length(s)} \quad \text{（公式 4.4）}$$

$$v_3 = NMCLCS_n(r,s) = \frac{length(MCLCS_n(r,s))^2}{length(r) \times length(s)} \quad \text{（公式 4.5）}$$

Islam 对这三个不同值 v_1，v_2 和 v_3 各赋一个权值系数 w_1、w_2、w_3，并满足约束条件 $w_1+w_2+w_3=1$，两个字符串的相似度由公式 4.6 表示为：

$$similarity = \sum_{i=1}^{3} w_i \cdot v_i \quad \text{（公式 4.6）}$$

图 4.7 为当 w_1、w_2、w_3 取不同值的时候正确率变化的曲线图。当 vextex 取最大值（0.7, 0.2, 0.94）时有 $w_1=0.7$，$w_2=0.1$，$w_3=0.2$。

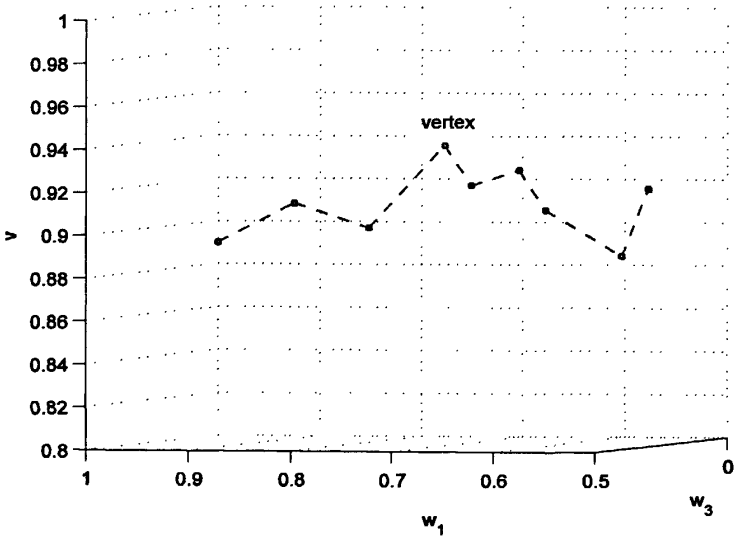

图 4.7 正确率变化的曲线图

公式 4.3、4.4、4.5 中的 r，s 代表两个不同的字符串，length 函数返回字符串的长度。式 3 中的 LCS 返回两个字符串的公共子序列。式 4.4 中的 $MCLCS_1$ 代表两个字符串从首字符开始匹配的最长连续子序列。式 4.5 中 $MCLCS_n$ 代表两个字符串从任一字符开始匹配的最长连续子序列。

本文设置 $w_1>w_3>w_2$，这样做的原因是经过分词后，词语本身长度就比较短，因此子字符串是否连续相同并非作为考虑的主要因素。设置 $w_3>w_2$ 可以使相似度增大，以利于排序。假设 r=“胸中之府”，s=“髓之府”，$w_1=0.7$，$w_2=0.1$，$w_3=0.2$，

则：

$LCS(r, s)=$“之府”

$MCLCS_1(r, s)=NULL$

$MCLCS_n(r, s)=$“之府”

$NLCS(r, s)=2^2/(4\times3)=0.33$

$NMCLCS_1(r, s)=0/(4\times3)=0$

$NMCLCS_n(r, s)=2^2/(4\times3)=0.33$

$similarity(r, s)=0.33\times0.7+0.33\times0.2=0.297$

在本文中，两个词条的相似度由其辞典解释的主题词数组决定。词组比对采用贪心策略，每个词语和另外词组中每一词性相同的词语做相似度运算，只记录最大的相似度值s，将其累加到词语相似度值r上，最终返回归一化后的r。两个主题词数组的相似度算法如下所示：

算法1：getSmilarity (获取特征数组之间的相似度)

输入: A，B /*A和B按词性排序好的词语数组，A数组词语个数小于B*/
输出: r　　/* r是两词语数组之间的相似度，0<=r<=1 */

```
1   foreach(string itema in A)
    {
       s ← itema 和itemb之间最大的相似度
          //itemb∈B，itemb的词性和itema的词性相同。
       r = r+s
    }
```

2　$length=\sum_{i}^{posset}\max(pos_i)$

/* 归一化操作，posset表示所有的词性种类集合，设A中词性为i的词语个数为a，B中词性为i的词语个数为b，比较a和b含有词性为i的词语个数，返回较多者的个数。*/

3　r=r/length

将全部中医辞典的解释字段分词后提取主题词，对其主题词语数组做C_n^2次求元素相似度运算，形成一个上三角矩阵 M_1。中药的功效字段、方剂的功能主治字段分词后，同样做C_n^2次求相似度操作，形成上三角矩阵M_2和M_3。设推荐词语

集合阀值大小为 K，则推荐词语集合就是对矩阵中查询词所在的行、列向量做一次 K 阀值聚类。由于辞典学习中可能存在用户不感兴趣的推荐词，故选取一定数量的后备推荐词语以替换此类词语。取 K=N+M，N 为显示在页面上的词语个数，M 为后备待使用的词语个数。

网络文档学习关注术语词条间的语境信息。语境信息反映了两个词语同时在文档中出现的频率相关性。本文采用了三大搜索引擎中的数据，具体实施方法是：对辞典中的所有名称字段在谷歌、百度、雅虎分别做 C_n^2+n 次搜索操作，记录查询结果总数，其中 C_n^2 次为两个词语共同查询操作，n 次为单个词语查询操作。仿照向量余弦计算公式，本文中计算词语间的语境相似度如公式 4.7 所示：

$$similarity = \frac{f^2(w_i, w_j)}{f(w_i) \cdot f(w_j)} \qquad \text{（公式 4.7）}$$

其中 $f(w)$ 代表词语在网络中出现的文章篇数，$f(x,y)$ 表示两个词语在网络中同时出现的文章篇数，由 $0 \le |\cos\theta| \le 1$ 可知相似度 similarity 的取值范围 $\in [0,1]$。类似辞典学习中的方法，三种搜索引擎可以得到不同相似度矩阵，用公式 $M = aM_1 + bM_2 + cM_3$ 计算出不同词语间的最终语境相似度矩阵 M。其中 a+b+c=1，a，b，c>0。M_1、M_2 和 M_3 分别代表百度、谷歌和雅虎的搜索矩阵。本文按搜索引擎的中文页面收录数量高低，故设 a>b>c。类似于辞典学习中的方法，同样在 M 中选取语境相似度最大的 K 个推荐词语作为推荐列表。

4.5 页面评估算法

此模块主要是工作是建立页面反向索引和对页面进行价值评估。首先对图书全文按照二元文法方式进行切分，比如：“脾开窍于口”可切分为“脾开|开窍|窍于|于口”。这样，在查询的时候，无论是查询“脾开窍”还是查询“开窍”，都将查询词语按同样的规则进行切分：“脾开开窍”、“开窍”。多个关键词之间按逻辑与的关系组合，因此能够正确地映射到相应的索引中。对用户输入的搜索文本，引擎采用反向索引机制在页面中进行寻找。反向索引就是引擎维护了一个词、短语表，对于这个表中的每个词、短语都有一个链表描述了其所在页面。用户在输

入搜索文本时，通过反向索引就能快速得到搜索页面。页面首先按照页面排序评分公式4.8[42]对搜索结果页面依次排序。

$$score(q,d)=\sum_{t\,in\,q} tf(t\,in\,d)\,idf(t)\cdot boost(t.field\,in\,d)\cdot LN(t.field\,in\,d) \quad \text{（公式4.8）}$$

评分公式每个因子含义如下：*tf(t in d)*：页面d中出现搜索项t的频率，即t出现在当前页面d中的次数，出现次数越多，d的得分越高。实现公式4.9为：

$$tf(t\,in\,d)=\sqrt{frequency)} \quad \text{（公式4.9）}$$

idf(t)：搜索项t在倒排页面中出现的频率，反映有多少个页面d包含了该索引项，页面数越多，该因子的值越小，反之越大。它意味着terms在页面中出现的次数越少，贡献分数越高，即个性特征越明显。它的实现公式4.10为：

$$idf(t)=1+\log_2\left(\frac{numDocs}{docFreq+1}\right) \quad \text{（公式4.10）}$$

其中numDocs为索引中的页面数目，docFreq代表搜索项t在页面中出现的总次数。

boost(t.field in d)：搜索域的加权因子，它的值在索引过程中进行设置，可以设置每个页面的boost值，来改变它在所有页面中的重要性。框架中设置中医古籍的boost值大于现代中医书籍。页面中的一个域代表与这个页面相关的一部分数据，每个域对应于页面中的一个数据段，可以将域理解为类似数据库中的表字段。

LN(t.field in d)：搜索域的标准化值。一个比较短的域贡献更高的分数，如果域越长那么这个因子的值就会越低，反之越高。实现公式4.11为：

$$1/\sqrt{numTerms} \quad \text{（公式4.11）}$$

然后在这些页面中进行二次搜索，所有含有“搜索词+指的是”、“搜索词+是”、“搜索词+为”、“所谓搜索词”、“乃搜索词”、“即搜索词”、“为搜索词”等陈述性句式的页面，按照页面在书中的页序提升至第一次排序结果的最前面。这样做的原因是考虑到一般用户进行搜索的目的是寻求搜索词的具体含义，而陈述性句式多包含搜索词语解释性内容，上述搜索项反映了中医书籍中常见的陈述句式。

4.6 实验结果分析

此部分以中药“当归”为样例，分析框架产生的推荐结果以及对其评估情况。表4.2给出了和当归药性、功效、语境相关以及网络推荐的前10条结果。

表 4.2 当归关联中药表

药性相关	功效相关	语境相关	网络推荐
蚕沙	川芎	人参	当归的功效
防风	白芍	五味子	当归的作用
乳香	丹参	三七	当归粉
缬草	白术	金银花	当归片
延胡索	熟地	西洋参	当归的药用价值
羊红膻	何首乌	牛蒡子	当归丸
肉桂	阿胶	黄芪	当归性能
山楂	黄芪	甘草	当归汤
麝香	肉桂	川芎	当归苦参丸
伸筋草	牛膝	茯苓	当归食用方法

药性相关的中药和当归同属辛、温类药物，且归肝、心或脾经。功效相关的药物都具有补血调经、或者活血止痛的功效。语境相关的药物可能反映了它们经常和当归一起使用的频率较其他药物为高，或者经常被人们拿来对比，可以为读者学习时提供参考。网络推荐的词语反应了网络用户当前关注的词语，可能也会引起读者的兴趣。

定义推荐项的有效率=未过滤项/推荐项总数，并将有效率作为一个随机变量X。由经验得知，X服从正态分布$N(\mu, \sigma^2)$，而μ，σ^2均未知。从辞典中随机选取10种中药作为搜索词，记录其相应的推荐项的有效率。推荐集合大小为20，经一个月后统计其推荐效果，如下表4.3所示：

表 4.3 随机采样10个样本的有效率

中药	灵芝	茯实	泽兰	五加皮	佛手	玉竹	常山	五味子	车前子	虎杖
有效率	85%	85%	75%	90%	90%	75%	85%	85%	95%	85%

在置信度为0.9的置信水平下，有理由认为推荐的平均有效率大于80%。证

明如下：

用t检验法做关于平均有效率μ的单边检验。零假设$H_0:\mu \le \mu_0 = 80\%$，备择假设$H_1:\mu > 80\%$，取显著性水平α=0.10，即

$$\alpha = P\{t > t_\alpha(n)\} = \int_{t_\alpha(n)}^{\infty} \frac{\Gamma[(n+1)/2]}{\sqrt{\pi n}\Gamma(n/2)}(1+\frac{t^2}{n})^{-(n+1)/2}dt \qquad \text{（公式 4.12）}$$

知此检验问题的拒绝域为：

$$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \ge t_\alpha(n-1) \qquad \text{（公式 4.13）}$$

现在n=10，$t_{0.10}(9)=1.383$，$\bar{x}=85\%$，s=6.2361%，故$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}=2.4053>1.383$，故接受$H_1$，即认为推荐结果的平均有效率>80%的置信度大于0.9。证毕。

4.7 本章小结

本章中，首先介绍了基于正则表达式的实体抽取算法在系统的实现，该算法利用用户选取的关键词以及主题词模式生成抽取正则模式，以此模式对全文信息进行提取。本章接着对双数组Trie树信息分离算法做了描述，该算法通过模拟有限状态机的转移，对混杂在资料中的关键字进行了快速的信息分离。本文随后对方剂组成信息如何细化做出了阐述，然后给出了关键字、页面的语义推荐算法以及相应的页面评估的实现过程，在章节的最后利用统计学中常用的t假设检验方法实验结果做了分析和证明。

第5章 CADAL中医药推荐系统的实现

5.1 系统底层软件支撑环境

系统采用Sun公司提供的JAVA EE轻量级架构，通过Tomcat 6.20版中间件容器，提供Web访问功能，整个系统的设计模式采用流行的MVC（Model，View，Control）模式进行页面流程控制。具体采用Structs1.3进行页面逻辑编程，JSP页面编写采用了Jstl、Xhtml、Ajax、JavaScript等技术，开发IDE采用MyEclipse7.5。数据库底层采用 SqlServer2000、MySql 进行数据存取，其中的存取操作采用Hibernate 3.2 API操纵。建立搜索索引的时候采用Lucene2.3的API。采用Axis2对WebService进行部署，并采用C#对后台数据信息进行抽取、标注。

5.2 元数据管理模块的实现

通过信息提取并将元数据存储到数据库后，系统通过Hibernate框架对数据库进行管理。

Hibernate是一个开放源代码的对象关系映射框架，它对JDBC进行了非常轻量级的对象封装，使得Java程序员可以随心所欲的使用对象编程思维来操纵数据库。Hibernate可以应用在任何使用JDBC的场合，既可以在Java的客户端程序使用，也可以在Servlet/JSP的Web应用中使用，最具革命意义的是，Hibernate可以在应用EJB的J2EE架构中取代CMP，完成数据持久化的重任。

通过 Hibernate 框架，系统对中医信息数据表中的数据进行对象封装。配置Hibernate框架由XML存储。操纵配置文件如图5.1所示：Database Connection Configuration 存储了数据库连接信息，包括使用数据库系统名称（如Sqlsever2000），数据库连接URL，数据库的驱动名称，及登陆时需要使用的数据库的用户名和密码。Hibernate框架通过XML映像文件把实例映像至数据库表格。Mappings 一栏显示的是各个表格实例的 XML 映像文件，如 Medicine2.hbm.xml

它的作用是将 Medicine2 实例映像至数据库表格 Medicine2。

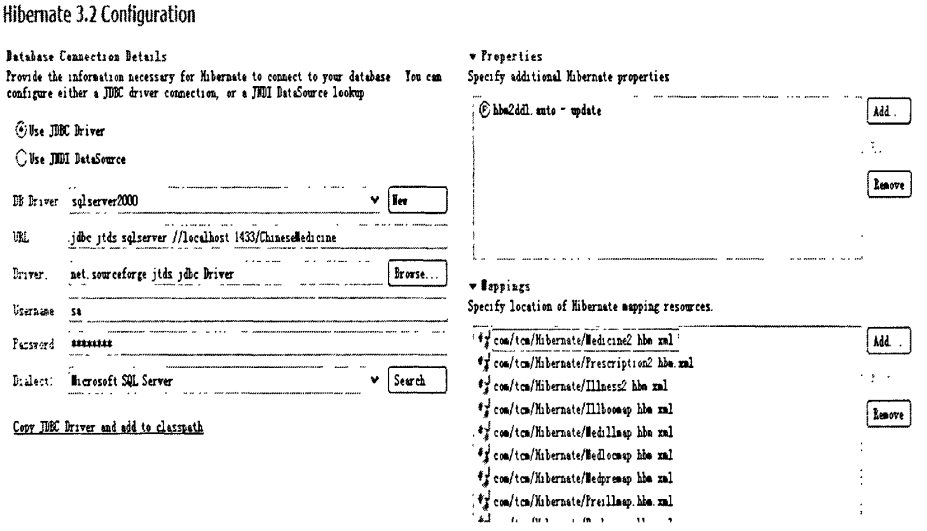

图 5.1 Hibernate 配置图

5.3 数据处理模块的实现

数据处理模块由知识获取模块，数据关联模块，数据检索模块三个模块组成。知识获取模块用于从医学书籍中提取信息，知识获得后，通过数据关联模块将大量的知识结构化，并以表格的形式存储到数据库中。用户需要检索时，通过提交检索条件，系统使用数据检索模块访问数据库，并返回相应的数据。

5.3.1 知识获取模块

利用正则表达式可以对一些格式良好的中医书籍进行信息提取，获得所需要的知识。数字图书馆将图书扫描后的 djvu 文件转化为文本文件，系统利用正则表达式对这些文本文件进行知识获取，以下蓝色字体是《中药学》书中的一页。

7．羌活

【性味归经】辛、苦，温，归膀胱、肾经。

【功效】解表散寒，祛风胜湿。止痛。

【应用】

①外感风寒，恶寒发热，头痛身痛等证。

②风寒湿邪侵袭所致的肢节疼痛、肩背酸痛，尤以上半身疼痛为适用。

【性能特点】本品辛温发散，发表力强，药势上达，其作用部位偏上偏表，善散太阳肌表风寒湿，通利关节而止痛。

【用量】3~10g。

【使用注意】①本品气味浓烈，用量过多，易致呕吐。脾胃虚弱者不宜服。

②血虚痹痛，阴虚头痛者慎用。

由于这本书结构良好，每种药物后均有关于药物的性味归经、功效、应用等药物相关描述，而且每项描述都是以“【】”开头，紧接着的是描述相关的信息，可以利用正则表达式提取知识。通过用户选定该段文本，可以分析出正则表达式“[\\d]{1,2}[．][\u4e00-\u9fa5]{1,7}(【[\u4e00-\u9fa5]*】([\u4e00-\u9fa5|，|、|；|~|\\w|\u2460-\u2470]*。)+)+”符合该段文本模式，可利用此提取信息。

5.3.2 数据关联模块

系统将从中医书籍中获得的知识进行结构化存储，在建立数据关联之前，首先必须建立以下各实体表：book（书籍表，包含书籍类型，书籍编号、书籍名称、作者名、出版时间、出版社、作者等书籍信息）、medicine（药物表，包括药物的各种信息）、illness（病证表，包括病证名称、症状名称等）、prescription（方剂表，包括方剂的各种具体信息），doctor（名医表，包括医生的描述，生活年代和活跃场所）、location（地点表，地名以及它们的经纬度信息）、dvdkeyframe（视频表，包含视频名称、视频段、关键帧以及视频时间等信息）。接着建立实体之间的关联表。而后可以建立起药物联系表 medsim（药物和药物之间的相似度表，包含药物的编号、名称以及药物之间的相似程度）；药物和方剂联系表 medpremap（由于药物属于方剂的组成，故可以建立药物和方剂间的联系）；药物和病证联系表 medillmap（药物可以主治关联的病证，故可建立药物和病证之间的联系）；药物和产地联系表 medlocmap（每一种药物都主产于某地，故可以建立药物和产地之间的关联）；药物和图片之间的联系表 medpicmap（每种药物的关联图片可能不止

一张，故里面存放的是图片的一些标准信息和图片存放地址）；药物和图书之间的关联表 medbookmap（此表包含了药物到书籍页面的直接映射）。方剂和病证也仿照药物和其他实体表关联方式建立相应关联表。

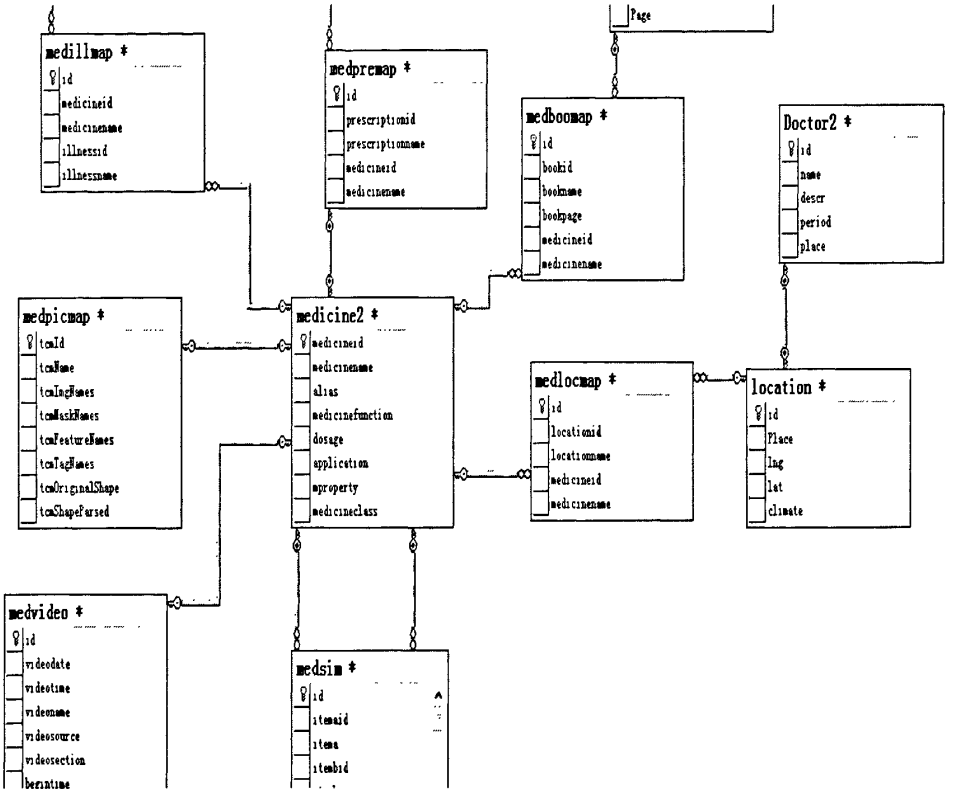

图 5.2 药物关联图

图 5.2 以药物为中心，描述了其和其他实体之间的关联，系统具体三方面关联图形成了一个复杂的关系网，在查询时可以体现出来。

5.3.3 数据检索模块

系统使用 Hibernate 框架对数据库进行管理。Hibernate 框架对数据库中的表格进行对象封装，而 SQL 数据库查询语句并不能方便地返回数据实例，而使用 HQL(Hibernate Query Language)来访问数据库时，由于其查询语句是面向对象的，它引用具体的类名以及其内部属性名称来检索数据，故可以直接返回数据实例。

HQL是面向对象的查询语言，它的查询语句和标准的SQL语句有些类似。它支持以下功能[43]：

（1） 在查询时可以动态设定查询条件，查询条件如限制字段长度，相似度等。

（2） 支持投影查询，即用户可以单独检索对象的特定属性。

（3） 支持分页查询，这在搜索系统中经常使用。

（4） 支持对象间的连接查询，此处的连接层数没有限制。

（5） 支持分组查询，允许使用having或者group by等分组关键字。

（6） 提供内置函数，如sum()求和函数，count()统计计数等。

（7） 能够调用自定义的SQL函数进行查询。

（8） 支持子查询，即嵌入式查询方式。

（9） 支持动态绑定参数，比如查询词可以从函数实参中传递。

用户可以首先通过Session中的createQuery()方法创建一个Query对象，然后在此对象中设置HQL查询语句。在语句中，可以动态绑定参数，之后可以调用list()方法进行查询。

5.4 流程控制模块

系统使用Jsp Web页面显示用户检索的结果，当用户发送请求后，系统将请求发送到相应的Action对象，对请求进行处理，并返回用户检索的结果。

Struts1.3是采用Java Servlet/JavaServer Pages技术，开发Web应用程序的开放源码的框架。采用Struts1.3能开发出基于MVC（Model-View-Controller）设计模式的应用构架。

系统使用Structs1.3框架创建Web应用程序，对应用程序的显示、表示和数据的后端代码进行了抽象。

系统通过设置structs-config.xml配置文件控制每个页面控制器的入口点。

如图5.3所示，该图展示了药物相关图片信息检索MVC模型的流程，图片上传网页duimage.jsp首先填充图片检索的查询条件，即填充模型，形成searchUploadImage表单，同时将表单模型向Action模块imageresult发送检索图

片的请求，imageresult 对请求进行处理后，将图像 url 以及图像标注数据返回给 duimage.jsp，由其负责将数据呈现给用户。

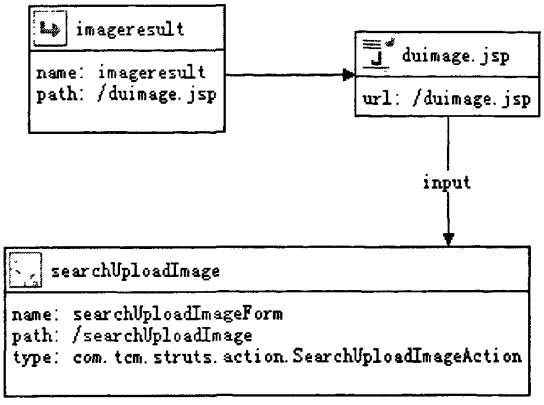

图 5.3 药物图片信息检索的 MVC 模型

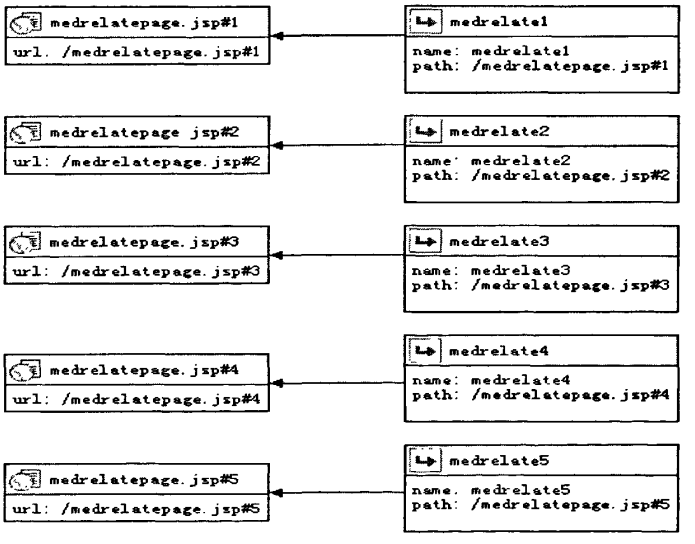

图 5.4 药物推荐部分流程控制图

图 5.4 展示了当检索药物时，相关推荐的流程控制，所有的推荐都集成在一张网页 medrelatepage.jsp 中。中药的相关推荐包括视频、图片、图书、地图、方剂、病证、推荐（类似药物）几个方面的推荐信息。如图 5.5 所示，系统采用了

标签JQuery滑动特效，以增强用户体验。用户可以通过点击标签的方式切换标签页，用户可以用鼠标点击标签控件来查看相关的推荐信息。

图 5.5 药物推荐集成页面

5.5 页面显示模块

在Web页面显示的展示层中，系统使用jsp、html加上css来展示基本页面。并使用JavaScript脚本语言处理客户端事件。

JavaScript是一种常用的网页脚本描述语言，该脚本可以无阻碍地嵌入于web文件之中或独立于web页面。系统中主要通过JavaScript做到区域显示，文档元素查找等功能，并使用Ajax技术做到页面异步、区域刷新，以减轻服务器负担。

Ajax的核心是JavaScript对象XmlHttpRequest。该对象在Internet Explorer 5中首次引入，它是一种支持异步请求的技术。简而言之，XmlHttpRequest对象使用户可以使用JavaScript向服务器提出请求并处理响应，而不阻塞用户。Ajax的工作原理相当于在用户和服务器之间加了一个中间层，使用户操作与服务器响应异步化。这样把以前的一些服务器负担的工作转嫁到客户端，利于客户端闲置的

处理能力来处理，减轻服务器和带宽的负担，从而达到节约ISP的空间及带宽租用成本的目的。

系统使用了JQuery[44]文档树操纵API库，该库是继Prototype之后又一个优秀的JavaScript框架，其特点是代码精简，但是功能强大。其1.3最高压缩率版只有21kb，具备了高轻量级特性，令其他js库过往所不及，它还兼容CSS3，以及各种主流浏览器（IE 6.0+、FF 3.5+、Safari 2.0+、Opera 9.0+）。

系统Web页面调用JQuery的Ajax API来显示药物、方剂及病证的相关信息。对于用户检索，服务器并不会重新刷新整个首页的页面，而只是通过Ajax技术，将显示检索得到的信息加载入网页特定区域进行刷新。系统中的中医药地图加载也是使用Ajax技术显示的。用户在地图推荐页面中拖动Google地图时，页面仅仅是刷新显示Google地图，对整个页面并不进行刷新。

5.6 本章小结

在本章中，先是叙述了系统底层软件支撑环境，主要是对数据存取层、业务逻辑层、Web显示层技术做了介绍。同时描述了元数据管理模块的实现细节，介绍了Hibernate存取数据实体的一些配置方式，接着说明了数据处理模块的实现环节，包括知识获取模块、数据关联模块以及数据检索模块，然后表述了流程控制的具体情况，主要是对struct框架下的MVC模式做了探讨，最后再讨论了页面显示模块的实现过程，即Ajax技术在本系统中的一些应用细节。

第6章 总结与展望

6.1 总结

本文详细介绍了多源推荐集合产生的主要步骤，内容涵盖了信息抽取、信息分离、词条相似度分析、网络文档术语协同信息计算、网络流行词条引进、用户点击流日志统计、页面搜索结果排序以及推荐词条更新等。

系统通过对中医药书籍、网络信息进行信息抽取、实体信息分离等手段获得含有丰富本体的中医药信息知识库。通过文本特征分析、用户日志分析和辞典学习聚类获得多源推荐集合。通过 CADAL 中医药推荐系统，用户可以使用多种检索条件去检索中医知识库，在获得相关的中医药书籍和信息的同时，亦可通过系统提供的推荐信息进一步了解查询项的其他内在关联的知识信息。用户在查询中药、方剂和病证三者任一方面的信息时，三方的关联可以有机的结合起来，呈现在页面之上。由于用户能够同时获得横向的关联信息，方便了各类用户进行知识发现和科学研究工作。

另外，系统还存在其他增强功能服务（文中并未做主要介绍），如中药地图频道、中医药视频频道、中药分类浏览、方剂量化对比、中医文言文图书与翻译版本对照系统等等，使用户体验得到增强，通过书籍页面的呈现，使用户能够追本溯源，获得药物、方剂、病症相关的书籍资料。

6.2 展望

新的信息提取算法有待开发，从结构不良好的书籍、OCR 识别率低的图书中提取到结构化的数据。开发新的检索算法，作为搜索引擎中的个性化搜索后备资源。可以利用 UIMA 提供的接口开发一个搜索算法共享平台，开发者只需提供具体的搜索逻辑，框架将会自动集成算法到引擎的核心中。

另外，系统服务可以做进一步的增强，使用户可以通过系统获得更多信息。

比如用户需要寻找某一个朝代中医名家以及他们的相关著作；用户需要查找关于某药物、方剂、病证的论文资料；用户需要鉴别手头的中药是否为正品；用户需要判别一个方剂在医药界的使用率；用户需要得知某中药在市场的经济走向；用户需要对自己的舌像进行中医辨证分析等等。所有这些用户需求都可以作为系统将来继续发展的目标。

系统希望可以将新的知识不断加入到这个中医知识库中，形成一个互联的知识网络。网络可以无限拓扑，具有良好的可扩展性，每当有新的知识链接加入，就增加了知识网络的联通数，随着新知识的不断加入，整个数字图书馆的中医药信息利用率将随之提高。

参考文献

[1] 数字图书馆_百度百科 [DB/OL]. http://baike.baidu.com/view/8181.htm, 2009.

[2] 高学敏. 中药学 [M]. 北京：中国中医药出版社, 2002:1.

[3] 邓中甲. 方剂学 [M]. 北京：中国中医药出版社, 2004:1.

[4] 朱文锋. 中医诊断学 [M]. 北京：中国统计出版社, 2002:1.

[5] CADAL 管理中心. 百万册书数字图书馆项目在中国的背景情况 [OL]. http://www.cadal.zju.edu.cn.

[6] 张寅. 个性化技术及其在数字图书馆中应用的研究 [D]. 浙江：浙江大学计算机学院, 2009:引言 2.

[7] 阜新. 浅谈数字图书馆建设 [DB/OL]. 辽宁工程技术大学图书馆, 2004.

[8] MedLEE-A Medical Language Extraction and Encoding System [DB/OL]. http://lucid.cpmc.columbia.edu/medlee/.

[9] Unified Medical Language System [DB/OL]. http://www.nlm.nih.gov/research/umls/.

[10] Genome Information Extraction [DB/OL]. http://www.ims.uni-stuttgart.de/projekte/GenIE/.

[11] The RAPIER Information Extraction Rule Learning System [DB/OL]. http://www.cs.utexas.edu/~ml/rapier.html.

[12] Freitag, D. Information extraction from html: Application of a general machine learning approach [C]. Proceedings of the 15th Conference on Artificial Intelligence (AAAI-98), Madison USA, 7.26–7.30, Menlo Park: AAAI Press, 1998:517–523.

[13] WHISK: Learning IE Rules for Semi-structured and Free Text [DB/OL] http://www.cis.uni-muenchen.de/~yeong/Kurse/ws0809/WebDataMining/whisk.pdf.

[14] SoftMealy: Machine Learning for Web Information Extraction [DB/OL].

http://aiia.iis.sinica.edu.tw/index.php?option=com_content&task=view&id=30&Itemid=50.

[15] Robert Baumgartner, Sergio Flesca. Visual Web Information Extraction with Lixto [C]. Proceedings of the 27th VLDB Conference, Roma, Italy.2001. San Francisco USA, Morgan Kaufmann Publishers Inc, 2001:119-128.

[16] Ion Muslea, Steve Minton. STALKER: Learning Extraction Rules for Semistructured Web-based Information Sources [C]. In Proceedings on AI and Information Integration, Madison USA, 7.26–7.30, Menlo Park: AAAI Press, 1998:415-421.

[17] Liu L. Pu C. XWRAP: an XML-enabled wrapper construction system for Web information sources [C]. Proceedings of 16th International Conference on Data Engineering. San Diego, USA 2.29-3.3, 2000:611-622.

[18] Valter Crescenzi, Giansalvatore Mecca. Automatic Web Information Extraction in the ROADRUNNER System [J]. Lecture Notes In Computer Science; 2001, 2465:264-277.

[19] Rocchio J I. Relevance Feedback in Information Retrieval [C]. The SMART Retrieval System. Prentice-Hall, 1971:313-323.

[20] Sergey Brin, Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine [C] In: Proceedings of the Seventh International Conference on World Wide Web 7, Netherlands, 1998:107-117.

[21] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment [J]. Journal of ACM, 1999, 46(5):604–632.

[22] Nallapati R. Discriminative Models for Information Retrieval [C]. Proceedings of the 27th SIGIR conference. on information retrieval, 2004:64-71.

[23] Caruana R, Baluja S, Mitchell T. Using the future to "sort out" the present: RankProp and multitask learning for medical risk evaluation [J]. Advances in Neural information Processing System (NIPS), 2008:959-965.

[24] Burges C, et al. Learning to rank using gradient descent [C]. Proceedings of the 22nd intl. conf. on machine learning, 2005:89-96.

[25] T-s.Chua, S-Y. Neo, H-K. Goh, et al. Trecvid 2005 by nus pris [J]. NIST

TRECVID, Nov, 2005.

[26] Rong Yan, Alexander, Hauptmann D. Efficient Margin-Based Rank Learning Algorithms for Information Retrieval [J]. CIVR 2006, LNCS 4071:113-122.

[27] mod_gzip - serving compressed content by the Apache webserver [DB/OL] http://www.schroepl.net/projekte/mod_gzip/, 2003.

[28] ajp_百度百科 [DB/OL]. http://baike.baidu.com/view/2176620.htm?fr=ala0_1.

[29] Bezdek J C.Pattern recognition with fuzzy objective function algorithms [M]. New York: Plenum Press, 1981.

[30] Z Huang. Extensions to the K-means Algorithm for Clustering Large Data Sets with Categorical Values [J]. Data Mining and Knowledge Discovery, 1998, (2):283-304.

[31] Brendan J. Frey, Delbert Dueck. Clustering by Passing Messages Between Data Points [J] Science Express on 11 January 2007 Science 16 February 2007:Vol. 315. no. 5814, pp. 972 – 976.

[32] Y. Blanco-Fernandez, J. J. Pazos-Arias, et al. A flexible semantic inference methodology to reason about user preferences in knowledge-based recommender systems [J]. Knowledge-Based System, 2008, 21 (4): 305-320.

[33] Y. Wang, N. Stash, et al. Recommendations based on semantically enriched museum collections [J]. Journal of Web Semantics, 2008, 6 (4): 283-290.

[34] K. C. Lee, S. Kwon. Online shopping recommendation mechanism and its influence on consumer decisions and behaviors: A causal map approach [J]. Expert System Application, 2008, 35 (4): 1567-1574.

[35] J. L. Herlocker, J. A. Konstan, et al. An Algorithmic Framework for Performing Collaborative Filtering [C]. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, 1999: 230-237.

[36] N. Manouselis, C. Costopoulou. Preliminary Study of the Expected Performance of MAUT Collaborative Filtering Algorithms [C]. Proceedings of WSKS'08: First World Summit on the Knowledge Society, Athens, Greece, Springer, 2008: 527-536.

[37] E. Diaz-Aviles, L. Schmidt-Thieme, et al. Emergence of Spontaneous Order Through Neighborhood Formation in Peer-to-Peer Recommender Systems [J]. 2007, 177(6) :1349-1363.

[38] R. Schenkel, T. Crecelius, et al. Social Wisdom for Search and Recommendation [J]. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. 2008, 31 (2): 40-49.

[39] Jun-Ichi Aoe, Katsushi Morimoto. An Efficient Implementation of Trie Structures [J]. Software-Practice and Experience. 1992, 22(9):695-721.

[40] 王晓龙, 关毅. 计算机自然语言处理 [M]. 北京：清华大学出版社, 2005：50-51.

[41] AMINUL I, DIANA I. Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity [J]. ACM Transactions on Knowledge Discovery from Data, 2008, 2(2):10.

[42] Hatcher E, Gospodnetic O. Lucene in Action [M]. Greenwich: Manning Press, 2004:28-42.

[43] 孙卫琴. 精通 Hibernate：Java 对象持久化技术详解 [M]. 北京：电子工业出版社, 2006:284.

[44] jQuery: the write less, do more, JavaScript Library [DB/OL]. http://jquery.com.

攻读硕士学位期间主要的研究成果

[1] 施少敏，杨艳，魏宝刚. NPOS 最短路径分词实现方剂药物信息提取 [J]. 计算机应用与软件（录用）.

[2] Shi Shaomin, Wei Baogang, Yang Yan. Msuggest: A semantic recommender framework for Tradtional Chinese Medicine book search engine [C]. Proceeding of the 18th ACM conference on Information and knowledge management, 2009, KM: 533-542.

致谢

短短两年六个月，瞬间而过。本人在浙大攻读硕士研究生阶段也划上了句号。但是，两年多所经历的人和事并非一张白纸可以叙述完毕。非常感谢在学习、生活中给予我关怀和帮助的师长、师兄、同学、朋友、亲人以及叫不出姓名的保洁、保安人员。

首先感谢计算机学院的院长、实验室的领导人庄越挺教授，是庄老师给予了我在图书馆 CADAL 南方技术中心学习研究的机会，庄老师治学严谨，其治学之风使个人受益良多。

深深感谢我的导师魏宝刚教授，魏老师为人和蔼，对学生循循善诱，使本人从懵懵懂懂的本科毕业生成长为实验室科研的一分子。每遇其困难，魏老师必口传身教，其传业授道解惑精神令人由衷钦佩。魏老师对本人科研方向把握即时，是本人科研前进的舵手。再者感谢吴江琴副教授，吴老师为人平和，对学生礼遇有加，在生活上亦无微不至，其教导使人如沐春风。

感谢庄凌、朱文浩、鲁伟明、张寅、俞凯、袁杰师兄，感谢他们对本人学术问题的帮助，感谢杨晨醒、杨艳、张亮、周振坤同辈同学，和他们共事是人生中美好的回忆，感谢杜晨阳、叶振超、沈春辉、张振庭师弟、以及李瑞峰、谢大伟同学，没有他们，实验室的生活是不完整的。

感谢我的亲人和朋友，是他们在生活上无微不至的关怀，才能使我顺利完成学业，走向人生另外一个起点。

感谢实验室的保洁员以及图书馆的保安人员，是你们让实验室清洁和安全。

施少敏

二〇一〇年一月

作者简历

施少敏，男，浙江瑞安人。2007 年 7 月毕业于浙江工业大学软件学院，获工学学士学位，2007 年 9 月开始在浙江大学计算机学院攻读计算机应用硕士学位。

现在浙江大学数字媒体计算与设计实验室从事中英文图书数字化国际合作（简称 CADAL）项目研究工作，研究方向为数据挖掘、中医药信息检索。毕业后成为阿里巴巴数据仓库工程师。