武汉理工大学

硕士学位论文

基于语义网的博客搜索系统研究

姓名：章志龙

申请学位级别：硕士

专业：国际贸易学

指导教师：聂规划

20091101

摘 要

随着博客在中国的迅猛发展，写博客已经变成一些博客爱好者日常生活的一部分，通过在博客中分享自己独到的想法，挖掘社会中的真实一面，已获得越来越多人的关注，各大门户网站，比如搜狐、新浪、网易都推出了自己的博客栏目，并在首页占据重要位置。随着博客页面成指数级地增长，如何在数量众多的博客页面中找到自己感兴趣的博客成了一个很大的问题，光靠传统的搜索引擎或者是博客网站的站内搜索远远达不到人们的需要，急需一种针对博客的专业搜索引擎，能达到在语义层次上收集、组织和检索博客资源的目的，提高博客搜索的质量、更深层次地挖掘博客潜力和更加合理地对博客进行排序，这已变成博客发展面临的最大挑战。

本文首先介绍了语义网及搜索相关技术，包括语义网相关介绍、语义网中本体相关知识和搜索引擎相关原理和技术。其次，通过分析国内外博客搜索引擎的发展情况，找出了目前博客搜索中存在的问题，结合开源搜索工具 Lucene 和语义网中本体相关技术，提出了基于语义网的博客搜索模型的想法，并对关键子模型进行了详细的分析与构建，包括原始资源收集模块、索引建立模块、集成语义的综合博客主模块和用户检索模块。重点在索引建立和页面排序模块，提出了本体意群这一概念和集成语义的综合博客主模型，通过建立本体意群到文本的索引，极大地提高博客搜索的搜准率，集成语义的综合博客主模型对于页面排序起着相当重要的作用，能更深层次地挖掘博客的内在价值。接着对模型中涉及的关键技术与算法进行了研究，采用混合本体的方式构建博客本体，包括领域本体和语义词典，对语义词典的结构和相关功能也进行了分析。在算法方面，对博客页面排序算法和基于本体意群的索引算法进行了研究。

最后对博客营销的产生背景以及博客本体在博客营销中的应用进行了分析，构建一个基于 RSS 和本体技术的博客营销模型。还对博客营销的发展前景进行了预测，这些研究对基于语义网的博客搜索系统的实现提供了良好的理论和应用基础。

关键词：语义网，博客，博客搜索，本体，博客本体

Abstract

With the great development of blog in china, reading and writing blog is becoming a part of some blog lovers' daily lives. Through sharing special ideas in their blogs and exploring the truth hiding in the face phenomena of society, the blog is getting more and more concentration, even several portal websites, for example, sohu, sina, and netware, have established their own blog column, holding important position in the front page. While blog articles are increasing exponentially, How to find the blog which you are most interest in is becoming a tough problem, it is impossible to cater to people's needs just depending on traditional search engines or blog site's inner search engines. A professional search engine for blog is needed urgently, which can satisfy the aim of collecting, organizing and searching blog resources, enhance the quality of blog search, evacuate the blog's deep potential and sequence blogs more reasonably, which are the greatest challenges that we have to face.

Firstly, the article introduced semantic web and some search-related technologies, including semantic web knowledge、ontology-related knowledge and search engine principle and technologies. Through analyzing and comparing the development conditions of home and broad, I found some shortcomings in blog search. Secondly, combining the open-source search tool--lucene and semantic web's ontology technologies, I proposed the opinion of the blog search model based on semantic web and made adequate analysis and construction to several key modules, which include module of collecting original resources, module of establishing index, comprehensive bloger module of integrating with semantic and module of searching . The article's main points lie in the modules of index establishment and pages ordering, I conceived the concept of ontology semantic cluster and constructed a comprehensive bloger model integrated with semantics. It can largely enhance the precise ratio of blog search when the index is set from ontology semantic cluster to

blog pages, the semantic integrated blog model plays a great role in excavating the inner values of blogs. Thirdly , the thesis makes research about key technologies and algorithms in the modules, constructs blog ontology which contains domain ontology and semantic dictionary by hybrid method and analyzes the structure and related functions of semantic dictionary. In the aspect of algorithm, the thesis designs two fresh algorithms which include blog pages' sequencing algorithm and index algorithm based on ontology semantic clusters.

In the end, the thesis analyzes the background of blog marketing and the application of blog ontology in blog marketing, constructs a blog marketing model based on RSS and ontology technologies, and makes a prediction to blog marketing' future. These researches established good theory and application infrastructure for the implementation of the blog search engine based on semantic web.

Keyword: semantic web, blog, blog search, ontology, blog ontology

独 创 性 声 明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，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，也不包含为获得武汉理工大学或其它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了谢意。

研究生签名：章志龙　　日　期：2009年12月1日

关于学位论文使用授权的声明

本学位论文作者完全了解武汉理工大学有关保留、使用学位论文的规定。特授权武汉理工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，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

（保密的学位论文在解密后适用本授权说明）

研究生签名：章志龙　导师签名：　　日　期：2009-12-1

第 1 章 绪论

1.1 研究目的与意义

互联网经过多年的发展，信息的组成已经不再局限于简单的发布与共享，特别是博客的兴起与 RSS 的广泛应用。信息的发布源已经由政府，公司，机构逐渐延伸至个人。个人发布信息和信息交流成为网络信息一个庞大的资源，对于传统信息的发布，往往是以信息的广泛传播，提升知名度，增加点击率，从而直接或间接的创造价值为目的。WEB2.0 的迅猛发展，增强用户体验成了一个大趋势，利用社区化来聚集人气。

自 2002 年 8 月“博客”中文一词诞生以及“博客中国”正式开通以来，博客在中国的发展已经走过了整整 7 个年头。虽然美国的博客早在 20 世纪 90 年代后期就有萌芽，但是，可以说 7 年来，博客在中国的发展超过了世界上任何一个国家，博客深入中国社会，博客影响主流大众，已经是当今社会最重大的事件之一。可以在中国各大门户网站，比如搜狐，网易，新浪等都可以看到博客栏目与其他重大新闻同样占据重要位置。从明星开博，到国家总理在博客中表达民生关切等重大问题，这都凸显了博客在社会变革中的强大力量。

但是，在热闹和喧嚣的同时，名人博客的“生活垃圾”全面泛滥，传统媒体对于博客的“猎奇式窥探”，以及大众博客对生活隐私和道德界限的屡屡突破，使得中国博客发展状况明显呈现“虚热”现象。博客应用的肤浅化和庸俗化正在给博客的健康带来极大的“回火”。

人们对博客的认识应该更加趋于理性，随着博客的发展，内涵越来越丰富，有情感交流、技术交流、信息交流等，现在也涌现出很多专业博客，很多有个共同兴趣爱好的博友聚集在一起交流心得体会，一起共同进步，来进一步提升国内的交流氛围，比如，中国最先进的 IT 领先社区 CSDN，里面基本涵盖了 IT 行业所有的技术内容，聚集了国内广大的 IT 技术爱好了，是一个很优秀的博客社区网站。

随着博客的发展，博客页面的数量也是呈指数级上升，人们希望通过博客

发布自己的信息，也希望通过浏览别人的博客，了解到自己感兴趣的知识和话题，博客成了一个发布信息、传播信息、发现信息的重要途径。但是如何在数量众多的博客页面中找到自己感兴趣的话题成了一个很大的问题，光靠传统的搜索引擎或者是博客网站的站内搜索是远远达不到人们的需要的。这种需要促使人们对博客搜索引擎这种针对博客页面的专业搜索引擎的研究。

传统的信息搜索模式实质上采用的是一级映射模式[1]，即用户提交的关键字被直接传给搜索代理，搜索代理用机械匹配的方式到预先建好的索引文件中去检索，然后把检索到的相关结果返回给用户。这种检索方式忽略了关键字背后隐藏的意思，对于文本中的语义关系也没有进行深入的挖掘，搜索出来的很多结果都是不相关的，这既浪费了用户宝贵的时间，又造成大量的服务器资源的浪费。

面对博客的迅猛发展，博客搜索也面临着重大的挑战，它不同于传统的网页搜索，它是一个具有个性化的传播媒介，更是一个个性化的交流平台。博客里面蕴藏了许多极有价值的信息，这些信息需要不断去挖掘与分析，对搜索提出更高的要求。语义网技术的出现，依托博客的发展环境，将使得博客成为语义网技术应用的一个很好的试验平台。将语义网技术应用于博客搜索，将使得博客的发展更为迅猛，使博客的社会价值得到更大的体现，将最有价值的信息展现在人们面前，让更多的人了解到博客的神奇力量。博客们可以就同一个热点问题进行不断更新报道，提供有价值的线索，让人们了解到一件事情的整个发展过程。

针对当前博客搜索存在的问题，搜索层次低，排序不合理，没能充分挖掘博客的内在价值等问题，使我产生了将语义网技术运用于博客搜索的想法。为了提高博客搜索的质量，整合博客信息，进一步提升博客的社会影响力，使之成为一支能够与传统媒体相互并存的媒体力量，来完善传统媒体存在的弊端和不足，从而更加真实地反映社会存在的问题，为构建和谐社会发挥重要作用。将语义网技术运用于博客搜索中，能达到在语义层次上收集、组织和检索博客资源的目的，能更深层次的挖掘潜在草根博客，提高草根博客的社会认可度，从而构建一个良好的博客环境。

1.2 国内外研究现状

国内的博客搜索引擎才刚刚起步，但是已经出现了百度博客搜索、网易的有道博客搜索、搜狗搜索、新浪的爱问博客搜索等针对中文博客网站的博客搜索引擎。它们各自有自己的特色，百度是名气比较大的，有大类搜索，有道名气不大，但很实用，在博客收录数量上比百度多一个级别，更新周期和百度差不多，大有赶超百度之势，而且一直在尝试给用户一些新的东西，比如，用户可以将自己未被收录的博客加进有道搜索引擎，它还能识别拼音输入，确实很强大。通过对国内博客搜索引擎的使用和研究，我发现国内博客搜索引擎的发展还是比较快的，相关性也提高了许多，对最近更新的博客能很快收录，但在搜索相关性与国外还是有差距，有名的博客搜索引擎太少，个性化博客搜索引擎与国外差距很大（国外已在自然语言处理方面有很大的突破），有的甚至就是基于商业搜索引擎的高级搜索而已，加入了一些关键词的组合和排除，而且搜索结果中充斥着很多无用信息和不相关信息，使用者不得不在花费大量时间的情况次下才能找到自己感兴趣的博客信息。

搜索引擎界的巨人 Google 也推出了博客搜索引擎，它是搜索引擎的领导者，在博客搜索这块也很强，但在中文博客搜索这块与百度和有道相比觉得还是有点差距。

从国外的发展情况来看已经出现了很多顶级的博客搜索引擎，如 Technorati[2]、BlogStreet[3]、DayPop[4] 等，由于上述博客搜索引擎涉及一些敏感信息，现已被国家防火墙所屏蔽，在国内无法访问，现在比较知名的有 BlogPulse[5]，它能搜索出最近 24 小时内新建的博客和发表的最新文章，截止 2009 年 10 月 23 日，收录的博客数量为 121,614,971 个，过去 24 小时更新的博客数量为 977,615 个，新建博客数量为 95，528 个，该搜索引擎站点分析报道博客世界每天的热点问题、人物、博客、链接和新闻事件等。

Sphere[6]也是一个很不错的博客搜索引擎，它在博客搜索结果中提供相关内容（Related Content）按钮，点击按钮打开一个窗口，显示与当前博客日志讨论内容相关的博客和新闻信息的链接。还有一个很独特的就是它开发了一些很实用的插件，相关内容插件（Sphere Related Content Widget）：安装插件后，在博客日志结尾会显示一个 Sphere 相关内容链接，点击可以打来一个拓展窗口，在拓展窗口中显示与当前阅读的博客日志内容相关的其他博客日志和新闻。相关

内容网络书签插件（Sphere Related Content Bookmarklet）：在浏览器中安装“Sphere It”按钮，当用户使用安装了插件的浏览器阅读网站的时候，点击“Sphere it”按钮，可以打开一个拓展窗口，在拓展窗口中显示与当前网页中的内容相关的博客日志的链接和摘要。最关键之处在于这些插件可以被广泛的安装到数以千万计的博客和浏览器中，成为众多用户的常用工具，而不仅仅局限于 sphere 网站本身，一旦流行起来，名气和效果自然就不断扩张。

1.3 本文的内容与结构

本文分为 6 章，具体结构如下：

第一章是本文的研究背景，国内外发展现状，还有研究目的和意义。

第二章主要介绍了搜索引擎与语义网相关知识，包括搜索引擎发展历程、工作原理、关键技术和评价原则，对开源搜索工具 Lucene 也进行了分析。对语义网中的重要组成部分本体的相关知识也进行了详细的介绍，包括本体概念、本体描述、本体构建和本体映射，为后面的博客相关本体构建做了些铺垫。

第三章主要是对基于语义网的博客搜索模型进行了设计，首先，对博客和博客搜索相关内容进行了简要的介绍，接着对基于语义网的博客搜索模型进行了设计，分整体设计和子模块构建，提出了一个集成语义的综合博客主模型和基于本体意群的索引建立方式。此外，本章还对博客搜索的整个流程进行了分析。

第四章对博客搜索中涉及的关键技术与算法进行了研究，主要有博客本体的构建，包括博客领域本体和语义词典的构建，算法方面有基于本体意群的索引算法和基于集成语义的综合博客主模块的博客页面排序算法。

第五章分析博客本体在博客营销中的应用，对其中涉及的关键技术，RSS 技术和本体技术的运用进行了分析。

第六章为总结和展望，首先分析了本文的不足，提出了下一步的努力方向，最后对博客搜索的前景进行了展望。

第 2 章 搜索引擎与语义网理论基础

2.1 搜索引擎相关理论基础

2.1.1 搜索引擎的发展历程

互联网上的第一代搜索引擎创造了一段全新的历史。这个阶段的搜索引擎以 Altavista、YAHOO 和 Infoseek 为代表，主要依靠人工目录分类，由于人工分类难以处理海量的信息，搜索结果的好坏主要以搜索出来的数量来衡量。随后的几年，相继出现了 Lycos、Altavista 和 Inktomi 系统。其中 Lycos 和 Inktomi 系统提供了网络蜘蛛的功能，真正实现了搜索内容的动态更新，搜录的数据量也快速增加。而 Altavista 是一个支持自然语言处理搜索的搜索引擎，支持用户自己提交网站，快速上线。但是这些工具没有从真正意义上占据市场，在很长一段时间内，YAHOO 是行业公认的垄断者。

互联网上的第二代搜索引擎伴随着互联网内容的指数增长而出现。链接分析技术的引入，真正提高提高了搜索引擎的结果质量，搜索引擎真正跨入了第二代自动搜索引擎。搜索引擎系统以信息自动抓取和自动排序检索为特征。这个阶段成功的产品是 Google，Google 占据了大量的用户和市场份额。微软近期推出 Bing 搜索引擎和雅虎合作，利用雅虎的搜索平台来挑战谷歌的搜索市场霸主地位。

互联网上的第三代搜索引擎目前正在发展和形成当中。第三代搜索引擎应该具有的特征正在讨论中。个性化、分类化和智能化是比较公认应该具备的特征，现在已经出现了一些这方面的初级产品，不过还没有形成市场化，比如维基百科，powerset 等，powerset 是一个专门从事所谓"自然语言"搜索的公司，通过分析用户所输入文字的意义和目的，powerset 可以分辨出各类词义，从而为用户搜索出所需要的更为精细的内容。powerset 的网络搜索做得非常好，语义上就像维基百科一样。

中文搜索引擎目前还处于黄金的发展期。知名的中文搜索引擎包括百度搜索、谷歌搜索、雅虎中文搜索、搜狗搜索、天网搜索、网易有道搜索等。百度搜索是国内目前最成功的中文搜索引擎，与谷歌之间的竞争依然很激烈，其他

几个搜索引擎也都很有自己独到的定位，比如网易，网易认为未来的搜索无处不在，今后搜索将融入网络生活的每个细节，例如，它推出的网易有道海量词典就是一项非常不错的产品，充分考虑到了用户的需求和心里，比如有笔画输入和公式计算功能。现在的主流搜索引擎都开发了针对特定领域的搜索引擎，也称为主题搜索引擎，目标更加明确化，国内对主题搜索引擎做了许多研究，也开发了一些系统来测试这些初级的系统的性能。

在基于搜索引擎框架的研究和应用基础上，黄波[7]对搜索引擎的设计原理和特色以及评分体系进行深入的研究。是一个建立在核心之上的搜索的实现。利用易于扩展的插件机制进行二次开发。研发一个第三方工具把特殊的数据格式转化为可视化的结构，以便研发人员对索引数据进行分析查询。肖亮[8]提出了构造垂直搜索引擎时最重要的两个模块，即网页搜集模块和结构化信息抽取模块的架构设计及算法模型。在网页搜集模块中，对垂直搜索所要着力解决的“主题飘移”现象，提出了通过主题判定，主题预测和网页排序的手段来防止这种现象，并在各自的模块中提出了相应的算法模型。在结构化信息提取模块中，构造了一个基于 XML 技术的信息抽取系统的原型。将搜索模块和信息提取模块进行合理的组合配置，形成了垂直搜索引擎的核心部分，为创建一个完整的垂直搜索引擎打下了良好的基础。

这些研究对搜索引擎的发展起到了一定的作用，但当研究遇到了瓶颈时，就要寻求突破，不断变换新的思维，随着语义网技术的成熟，将语义网技术运用于传统网络来进行优化就是大势所趋。

2.1.2 搜索引擎的工作原理

搜索引擎系统通常是指互联网网页信息检索系统。搜索引擎收集了成百上千至几十亿的网页。搜索引擎程序自动对网页内容进行分析，通过分词程序得到的语素关键词，经过索引加载建立索引数据库。用户通过 Web 页面查询索引数据库，返回的结果中包含了所有匹配检索关键词的网页。得到的搜索结果利用网页的重要性和关键词匹配的算法进行排序，最后展示给用户，这就是整个搜索过程。

搜索引擎的实现原理，可以看作有四步从互联网上抓取网页—建立索引数据库—在索引数据库中搜索—对搜索结果进行处理和排序。

（1）从互联网上抓取网页

搜索引擎向互联网派出能够自动收集网页的网络蜘蛛程序自动访问互联网，并沿着网页中的所有爬到其它网页，重复这过程，并把爬过的所有网页收集到服务器的原始数据库中。

（2）建立索引数据库

索引系统程序会对这些收集回来的网页进行分析，提取相关网页信息，根据一定的相关度算法进行大量复杂计算，得到每一个网页针对页面内容中及超链接中每一个关键词的相关度或重要性，然后用这些相关信息建立网页索引数据库。

（3）在索引数据库中搜索

当通过用户界面输入关键词进行搜索时，服务器会分解搜索请求，由检索器从网页索引数据库中找到符合该关键词的所有相关网页。

（4）对搜索结果进行处理排序

所有相关网页针对该关键词的相关信息在索引库中都有记录，只需综合相关信息和网页级别形成相关度数值，然后通过检索器进行排序，相关度越高，排名越靠前。最后由页面生成系统将搜索结果的链接地址和页面内容摘要等内容组织起来返回给用户。

2.1.3 搜索引擎关键技术

（1）网络机器人技术

网络机器人(Robot)又被称作 Spider、Worm 或 Random，核心目的是为获取 Internet 上的信息。一般定义为“一个在网络上检索文件且自动跟踪该文件的超文本结构并循环检索被参照的所有文件的软件”。机器人利用主页中的超文本链接遍历互联网，通过地址引用从一个 HTML 页面爬行到另一个 HTML 页面。网上机器人收集到的信息可有多种用途，如建立索引、HIML 文件合法性的验证、URL 链接点验证与确认、监控与获取更新信息、站点镜像等。

机器人安在网上爬行，因此需要建立一个 URL 列表来记录访问的轨迹。它使用超文本，指向其他文档的 URL 是隐藏在文档中，需要从中分析提取 URL，

机器人一般都用于生成索引数据库。所有 WWW 的搜索程序都有如下的工作步骤：

1)机器人从起始 URL 列表中取出 URL 并从网上读取其指向的内容；

2)从每一个文档中提取某些信息(如关键字)并放入索引数据库中；

3)从文档中提取指向其他文档的 URL，并加入到 URL 列表中；

4)重复上述 3 个步骤，直到再没有新的 URL 出现或超出了某些限制(时间或磁盘空间)；

5)给索引数据库加上检索接口，向网上用户发布或提供给用户检索。

搜索算法一般有深度优先和广度优先两种基本的搜索策略。机器人以 URL 列表存取的方式决定搜索策略：先进先出，则形成广度优先搜索，当起始列表包含有大量的 WWW 服务器地址时，广度优先搜索将产生一个很好的初始结果，但很难深入到服务器中去；先进后出，则形成深度优先搜索，这样能产生较好的文档分布，更容易发现文档的结构，即找到最大数目的交叉引用。

（2）索引技术

索引技术是搜索引擎的核心技术之一。搜索引擎要对所收集到的信息进行整理、分类、索引以产生索引库，而中文搜索引擎的核心是分词技术。分词技术是利用一定的规则和词库，切分出一个句子中的词，为自动索引做好准备。目前的索引多采用 Non—clustered 方法，该技术和语言文字的学问有很大的关系，具体有如下几点：

1)存储语法库，结合词汇库分出句子中的词汇；

2)存储词汇库，要同时存储词汇的使用频率和常见搭配方式；

3)词汇宽，应可划分为不同的专业库，以便于处理专业文献；

4)对无法分词的句子，把每个字当作词来处理。

索引器生成从关键词到 URL 的关系索引表。索引表一般使用某种形式的倒排表，即由索引项查找相应的 URL。从实现方式上看，倒排索引是典型的为满足实际应用需要而设计的一种数据存储结构。这种数据结构中的每一个元素是一个索引项，每个索引项是有关键字属性值和关键字关联记录，或记录的存放地址组成。倒排索引是利用关键字直接确定文档列表，最后确定希望找到的文档列表。与传统的顺序查找和记录组织方式相反，因此称之为倒排索引。索引表也要记录索引项在文档中出现的位置，以便检索器计算索引项之间的相邻关系或接近关系，并以特定的数据结构存储在硬盘上。

具体的结构图如下：

Lexicon 索引词表

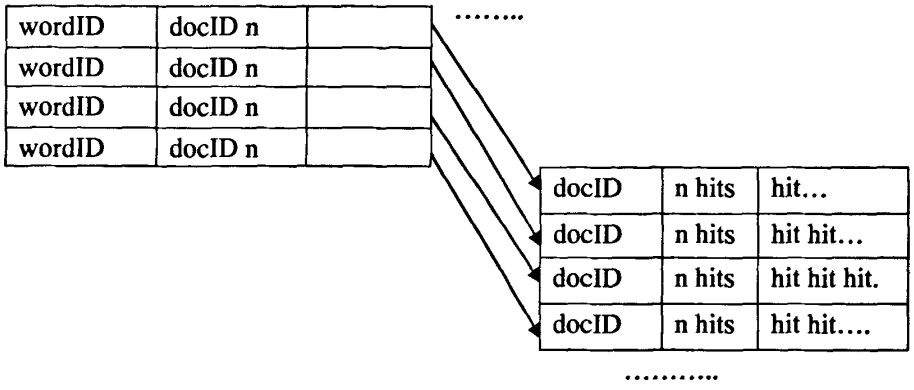


图 2-1 倒排索引文档结构图

按照索引建立和更新方式可以把索引划分为静态索引和动态索引。具体应用要看索引主要看在什么场合。如果索引的数据比较稳定、很少改变，可以采用静态索引。在要求数据实时性比较高的应用中，可以采用动态索引。静态和动态也不是绝对的，静态索引也会在适当的时机重新加载更新数据。动态索引也会有一定的更新时间延迟。

不同的搜索引擎系统可能采用不尽相同的标引方法。例如 Webcrawler 利用全文检索技术，对网页中每一个单词进行索引；Lycos 只对页名、标题以及最重要的 100 个注释词等选择性词语进行索引；Google 则提供概念检索和词组检索，支持 and、or、near、not 等布尔运算。检索引擎的索引方法大致可分为自动索引、手工索引和用户登录三类。

（3）检索器与结果排序技术

检索器的主要功能是根据用户输入的关键词在索引器形成的倒排表中进行检索，同时完成页面与检索之间的相关度评价，对将要输出的结果进行排序，并实现某种用户相关性反馈机制。

通过搜索引擎获得的检索结果往往成百上千，为了得到有用的信息，常用的方法是按网页的重要性或相关性给网页评级，进行相关性排序。这里的相关度是指搜索关键字在文档中出现的额度。当额度越高时，则认为该文档的相关程度越高。能见度也是常用的衡量标准之一。一个网页的能见度是指该网页入口超级链接的数目。能见度方法是基于这样的观点：一个网页被其他网页引用

得越多，则该网页就越有价值。特别地，一个网页被越重要的网页所引用，则该网页的重要程度也就越高。结果处理技术可归纳为：

1) 基于链接评价

基于链接评价的搜索引擎的优秀代表是 Google，它独创的“链接评价体系”是基于这样一种认识，一个网页的重要性取决于它被其它网页链接的数量，特别是一些已经被认定是“重要”的网页的链接数量。这种评价体制与《科技引文索引》的思路非常相似，但是由于互联网是在一个商业化的环境中发展起来的，一个网站的被链接数量还与它的商业推广有着密切的联系，因此这种评价体制在某种程度上缺乏客观性。

2) 基于访问大众性

基于访问大众性的搜索引擎的代表是 direct hit，它的基本理念是多数人选择访问的网站就是最重要的网站。根据以前成千上万的网络用户在检索结果中实际所挑选并访问的网站和他们在这些网站上花费的时间来统计确定有关网站的重要性排名，并以此来确定哪些网站最符合用户的检索要求。因此具有典型的趋众性特点。这种评价体制与基于链接评价的搜索引擎有着同样的缺点。

3)二次检索

进一步净化结果，按照一定的条件对搜索结果进行优化，可以再选择类别、相关词进行二次搜索等。过多的附加信息加重了用户的信息负担，为了去掉这些过多的附加信息，可以采用用户定制、内容过滤等检索技术。

由于目前的搜索引擎还不具备智能，除非知道要查找的文档的标题，否则排列第一的结果未必是“最好”的结果。所以有些文档尽管相关程度高，但并不一定是用户最需要的文档。

这些结果排序已经不能满足现在网络发展的需要，很多排序算法太简单，设计不合理，就会被一些人找到漏洞，进行投机来提升网站排名，扰乱搜索引擎的市场。这样下去会极大地影响人们对搜索引擎公正性的怀疑，使搜索引擎发挥的作用大打折扣。

现在搜索引擎排序技术不断在更新，各大搜索引擎公司都在做积极的调整，比如百度最近的重大调整有几下几点：

1）对新网站的收录变快，24 小时内就收录的网站并不稀奇；

2）对于网站的原创性要求更高，层次等级很明显的得到了改进；

3）重点提升了自身产品百科、图片、贴吧、知道、词典、有啊等相关内容

页面的权重，在百度这些产品的内容在第一页都有体现，特别是百度百科和百度图片；

4）针对论坛和博客站点导入链接降权，论坛签名已经对百度失去效应了；

5）对门户站的权重比较看重，这是算法调整最直接的表现；

6）网站上出现弹窗广告、很多 JS 代码、双向友情链接过多的网站进行降权，导出单向链接过多予以降权(是百度为了防止最近链接交易站点作弊)；

7）限制了新站的关键词排名。这一策略导致大量 1-3 个月的新站关键词排名浮动较大；

8）关键词中，区域性的网站排名靠前；

据国外媒体报道，Google 推出实验室版网站管理员工具的同时，悄然撤下了爬虫统计中的 PageRank 部分。虽然此时谷歌工具条中尚未正式撤销 PageRank，但此传闻在业界掀起轩然大波，难怪 PR 应该九月底十月初调整，而至今仍无音信。谷歌为什么欲拿 PageRank 来开刀？确实有太多的人将网站的 PR 值看得过于重要，认为 PR 值高网站的关键词排名就自然会向前排，从而忽略了真正重要的因素，那就是内容，原创内容，有规律的原创内容更新。

谷歌一直考虑将网站实用性列入判断网站好坏的标准，此次撤下了爬虫统计中的 PageRank 部分，也无疑验证了这一点，当一个以 PR 值高而出现在搜索引擎前面的网站无法给访客提供有使用价值的内容时，这个网站对搜索引擎来说也将失去意义，而搜索引擎的最大价值就在于为用户提供有价值的信息。Google 的排序算法有 100 多种，有很多也正在设计当中，它的原则是根据内容来进行排序，人为干预的因数很少，可以说是目前最好的全文搜索引擎。

2.1.4 开源搜索引擎 Lucene 介绍

Lucene 不是一个完整的全文索引应用,而是一个用 Java 写的全文索引引擎工具包,它提供了多个 API 函数与灵活的数据存储结构(可以定制),可以方便地嵌入到各种应用中实现针对应用的全文索引/检索功能。它是 APACHE 基金会 jakarta 的一个子项目。

作为一个广泛流传的全文搜索引擎，以其开放源代码、高效的索引结构、良好的系统架构赢得了诸多的开发者。从技术角度来看具有如下突出的优点：

1）整个系统开放代码，便于开发者根据自己的需要进行定制修改和调试，

开发者不仅可以充分的利用 Lucene 所提供的功能接口进行开发，甚至可以研究出搜索引擎的架构原理，写出自己的搜索引擎内核。

2）采用纯 java 实现，具有良好的跨平台和系统兼容能力。系统使用一个开发性的架构，便于设计一个合理且极具扩展能力的面向对象架构，依托于 APACHE 软件基金会的网络平台，程序员可以在 Lucene 的基础上扩展各种功能，发布到网上，并提供给其他开发者使用。

3）在核心技术和功能上，改进了传统全文检索引擎的倒排索引，实现了分块索引，提高了小文件索引速度，并提供索引优化机制，便于动态更新。提供了独立的文本分析接口和强大的查询分析接口。

（1） Lucene 系统的结构组织

Lucene 作为一个优秀的全文检索引擎，其系统结构具有明显的面向对象特征。首先是定义了一个与平台无关的文件索引格式，其次通过抽象与接口将系统的核心组成部分设计为抽象类，经过面向对象式的架构处理，运用多种设计模式，最终形成了一个松散藕合、高效率、适合二次开发的检索引擎。

Lucene 共有七个程序包组成，对于外部应用来说，主要有索引和检索两个模块。七个程序包如下：

org.apache.Lucene.search：检索入口，提供了根据索引进行检索的类。

org.apache.Lucene.index：索引入口，提供了用于访问与维护索引的类。

org.apache.Lucene.analysis：语言分析器，提供了将文本转化为可索引的词的类。

org.apache.Lucene.queryParser：查询分析器。用户可以定制。

org.apache.lucene.document：存储结构，文档的抽象描述。

org.apache.lucene.store：底层输入输出存储结构。

org.apache.lucene.util：一些公用的数据结构。

其系统结构及源码组织结构如下图所示：

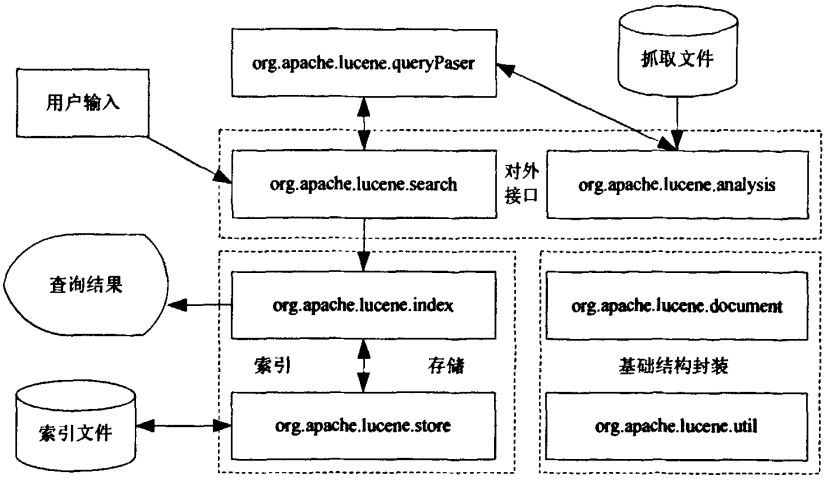
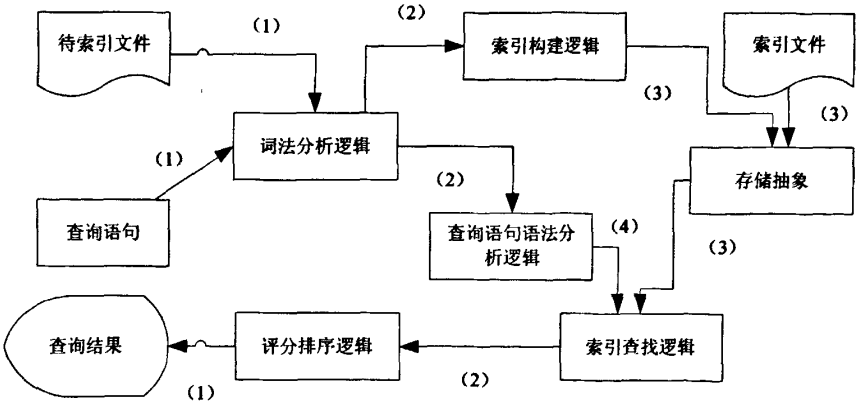


图 2-2 Lucene 系统组织结构图

（2） Lucene 数据流分析

通过探讨 Lucene 中数据流的走向，也可以理解 Lucene 系统结构，并以此摸清楚 Lucene 系统内部的调用时序。在此基础上，研究人员能够更加深入的理解 Lucene 的系统结构组织，方便以后在 Lucene 系统上的开发工作。这部分的分析，是深入 Lucene 系统的钥匙，也是进行重写的基础。图 2-3 表示的是 Lucene 系统中的主要的数据流以及它们之间的关系：



（1）文本流 （2）Token流 （3）字节流 （4）查询语句对象流

图 2-3 Lucene 系统中主要的数据流

图 2-3 表明了 Lucene 在内部的数据流组织情况，对 Lucene 内部的执行时序进行了详细的描述。

图中共存在 4 种数据流，分别是文本流、Token 流、字节流与查询语句对象流。文本流表示了对于索引目标和交互控制的抽象，即用文本流表示了将要索引的文件，用文本流向用户输出信息；在实际的实现中，Lucene 中的文本流采用了 UCS-2 作为编码，以达到适应多种语言文字的处理的目的。Token 流是 Lucene 内部所使用的概念，是对传统文字中的词的概念的抽象，也是 Lucene 在建立索引时直接处理的最小单位。字节流则是对文件抽象的直接操作的体现，通过固定长度的字节（Lucene 定义为 8bit 位长）流的处理，将文件操作解脱出来，也做到了与平台文件系统的无关性。查询语句对象流则是仅仅在查询语句解析时用到的概念，它对查询语句抽象，通过类的继承结构反映查询语句的结构，将之传送到查找逻辑来进行查找的操作。

词法分析逻辑对应于 org.apache.lucene.analysis 部分。查询语句语法分析逻辑对应org.apache.lucene.queryParser部分，调用了org. apache. lucene. analysis 的代码。查询结束之后向评分排序逻辑输出 Token 流，继而由评分排序逻辑处理之后给出文本流的结果，这一部分的实现也包含在了 org.apache.lucene.search 中。索引构建逻辑对应于 org.apache.lucene.index 部分。索引查找逻辑则主要是 org.apache.lucene.search，但是也大量的使用了 org.apache.lucene.index 部分的代码和接口定义。存储抽象对应于 org.apache.lucene.store。org.apache.lucene.util 作为系统公共基础设施存在。

（3） Lucene 的存储结构

Lucene 存放索引信息的是文件。它的索引存储文件设计的比较通用,输入输出结构都很像数据库的表→记录→字段,所以很多传统应用的文件、数据库等都可以比较方便地映射到 Lucene 的存储结构/接口中。它的索引存储文件结构描述如下:

Lucene 的基本概念为:索引(Index)、段(Segment)、文档(Document)、域(Field)、术语(term)。Lucene 每个索引由一个或多个段组成，每个段包含一个或多个文档，没个文档管理一个或多个域，每个域由一个或多个索引项组成，每个索引项是一个索引数据。它的结构可以对比数据库集合、数据库、表、表字段、字段值来理解。

1)索引（Index）

Lucene 每个索引的结构由一个或者多个段组成，索引最终结构体现到特定

格式的磁盘文件来存储。磁盘文件包括当前活跃索引段和新建的索引文件，通过工具整理可以把分段合并为统一的索引段。

2)索引段（Segment）

每次创建索引过程中，文档都是添加到特定的段里，然后索引段会根据参数进行合并。一个索引中只有一个没有后缀的 Segments_*文件，它记录当前索引的所有的 Segment 情况。它的后缀通常根据包含段的不同的变化。索引段相当于子索引，新建的索引通常以一个新段形式出现，在合并操作后每个索引体系通常只包含一个段。

3)索引文档(Document）

Document 是索引器可以直接添加的对象。每个索引可以包含多个不同的文档，每个文档又管理了数目不等的域集合。实际应用中，很多时候传统的文档会作为一个 Lucene 文档来添加。比如搜索引擎需要索引一系列网页，每个网页的文本内容作为一个 Lucene 域，创建时间、文件大小、文件来源、文件类型甚至文档摘要等附属信息也分别作为一个 Lucene 域，所有的域合在一起构成一个 Lucene 文档。

4)索引域（Field）

Lucene 的域是 Document 对象的基本组成单位。在实现中每个域对应 Field 类的实例来实现。每个域内存储了实际的索引文本数据，这些文本数据在内部调用了分析器 Analyzer 的索引项结果。域内数据的检索查询是以索引项为基本单位的，比索引项更小的单位无法检索到。

5)索引项（Term）

索引项是索引管理的最小单位，在程序中没有显示的调用，它是利用分析器，在后台自动把一个域的值进行分割，得到的每个独立元素作为一个索引项，用来建立索引。

2.2 语义网理论基础

2.2.1 语义网的产生

随着网络数据成爆炸式增长趋势，以及 Web 2.0 的出现，网络体验方式发生了重大的转变，网民成了网络信息产生与传播的一股新生力量，与传统网络内容提供商大有分庭抗礼之势。现有网络已不能满足日益增长的网络数据处理的需求，加上处理成本也大大增加，急需一种更加高效的网络来适应这些新的变化，让网络更加智能化，处理效率更高。

目前，人们十分迫切地要求网络具备一定的知识处理能力，因为网络只有实现数据的共享和自动处理，才能够发挥它的全部潜能。因特网技术的研究人员正在积极地研究新的技术，其中最令人瞩目的是语义网技术。语义网是因特网研究者对下一代因特网的称谓[9]。通过扩展现有因特网，在信息中加入表示其含义的内容，使计算机可以自动与人协同工作。即语义网中的各种资源不再只是各种相连的信息，还包括其信息的真正含义，从而提高计算机处理信息的自动化和智能化。而计算机并不具有真正的智能，语义网的建立需要研究者们对信息进行有效的表示，制定统一的标准，使计算机可以对信息进行有效的自动处理。

语义网[10]，是指在语义的基础上构建的网络。它是人的认知网络，或者说是一个巨大的知识库或概念图，存放的是人的知识，包括概念以及概念之间的各种关系。语义网中的知识表示可以粗略的分为三个层次：

（1）语言层次。反映语言表面现象的知识，如一个词的多语种形式，它的同义词、反义词、习惯用语、词的层次关系等。

（2）本体论层次。对概念的本体论的定义与解释、概念之间复杂的语义关系。

（3）常识层次。表述知识中常识上的关联。

可以看出，语义网处理的核心是语言，因为语言是知识的载体。在信息检索中，用户查询、系统查询的结果都是用语言表达的。语义网更强调的是自然语言处理技术在搜索技术中的应用。语义网的应用研究主要集中在以下几个方面:Web-services、基于代理的分布式计算、基于语义的网页搜索和基于语义的数字图书馆[11]。

Tim Berners-Lee 提出的语义网模型只是一个理想化的模型，其中的一个重要思想就是以本体来表示语义信息，通过在语义网中引入本体层来实现语义信

息的共享，从而提高网络信息服务的智能化与自动化。这一思想得到了众多语义网研究人员的认同，并在许多项目中致力于将本体论引入语义网的研究，其中代表性的项目有 On-To-Knowledge、KAON 和 COHSE[12].

语义网不同于传统 Internet 的最大地方就是它能让机器读懂人的意图，从而提高机器处理数据的能力。但并不是语义网取代传统网络，建设一个全新的网络，而是将语义网与现有网络进行融合，在某些方面应用语义网的技术来改造第一代网络。

2.2.2 语义网的体系结构

在 2000 年的世界 XML (Extensible Markup Language)大会上，万维网创始人蒂姆.伯纳斯.李对语义网的概念进行了解释，并提出了语义网的体系结构[13]。为了实现语义网信息服务智能化与自动化的目标，语义网研究者们开发了许多新技术并提出了一系列的技术标准。

语义网的体系结构共分七层，自下而上分别是编码定位层（Unicode+URI）、XML 结构层（XML+NS+xmlschema）、资源描述层（RDF+rdfschema）、本体层（Ontology vocabulary）、逻辑层（Logic）、证明层（Proof）和信任层（Trust）。各层之间相互联系，通过自下而上的逐层拓展形成了一个功能逐渐增强的体系。它不仅展示了语义网的基本框架，而且以现有的为基础，通过逐层的功能扩展，为实现语义网构想提供了基本的思路与方法。下面详细介绍一下该体系结构各层的含义、功能以及它们之间的逻辑关系。

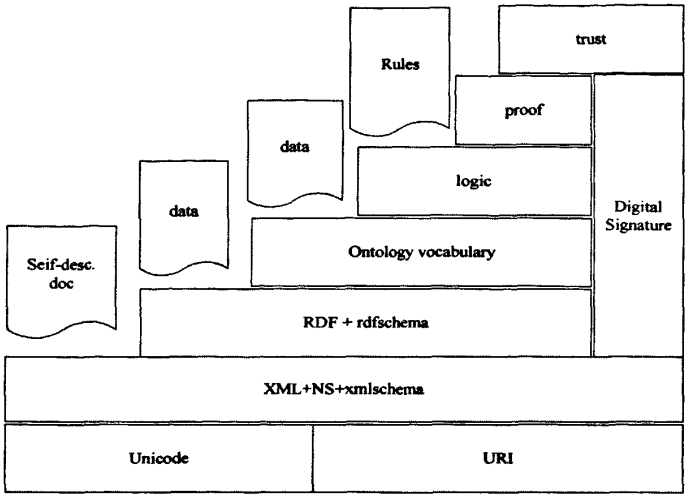


图 2-1 语义网体系结构图

（1） 编码定位层（Unicode+URI）

就像人与人之间的交流需要共同的语言一样，语义网要实现机器之间的相互交流与合作也需要使用双方认可的“语言”。语言需要对信息进行统一编码，编码是实现语言共通的前提，只有编码相同才能保证语言相通。在现有的网络中存在着种类繁多的语言及相应的字符集，要实现不同计算机之间的交流与合作，必须建立统一的编码格式。

Unicode 是一个字符集，可以表示 65535 个字符，基本上涵盖了世界上所有语言的字符。它的好处就是它支持世界上所有主要语言的混合，并且可以同时进行检索，能识别不同的语言的相同表达意思。可见，它为语义网提供了统一的字符编码格式，这种统一的编码格式不仅方便语义网上字符的表示，也有利于不同国家、不同民族的不同字符集在语义网上的统一处理、存储和检索。在现实生活中，不能仅仅通过一个简单的名字来唯一确定某个人。为了区别不同的网络资源，必须为它们确定不同的“社会关系”。对于网络资源来说，其“社会关系”就是 URI。

在语义网的体系结构中，编码定位层(Unicode + URI)处于最底层，是整个语义网的基础，其中 Unicode 负责处理资源的编码，URI 负责资源的标识。只有在对资源进行编码与标识的基础上才能对资源进行进一步的处理。

（2）XML 结构层(XML 十 NS+xmlschema)

XML 提供了一个标准，利用这个标准，可以根据实际需要定义自己的置标

语言，并为这个置标语言定义它特有的一套标签。因此准确地说，XML 是一种元标记语言，即定义标记语言的语言。

NS (Name Space)即命名空间，由 URI 索引确定，目的是为了简化 URI 的书写。例如 URI“http://www.w3.org/2001/02/22-rdf-syntax-ns#”就可以简写为“RDF"。通过在命名前加上 URI 索引前缀，即使具有相同命名的两个事物，只要它们的 URI 索引前缀不同，它们的命名空间就不一样，二者就不会被混淆。

XML Schema 实际上是 XML 的一种应用，它本身采用 XML 语法，所以 XML 文档是一种自描述文档。XML Schema 是 DTD (Document Type Definition) 的替代品，但比 DTD 更加灵活。它不仅提供了一套完整的机制以约束 XML 文档中标签的使用，而且支持更多的数据类型，能更好地为有效的 XML 文档服务并提供数据校验机制。

由于 XML 灵活的结构性、由 URI 索引的命名空间而带来的数据可确定性以及由 XML Schema 所提供的多种数据类型及检验机制，使 XML 结构层(XML+NS+xmlschema)成为语义网体系结构的重要组成部分。该层主要负责从语法上表示数据的内容和结构，通过使用标准的置标语言将网络信息的表现形式、数据结构和信息内容相分离。但 XML 在描述数据元上缺乏一定的灵活性；而且 XML 所表达的语义是隐含在文档的标记和结构中的，它只能被了解其标签含义的专业人员所使用。因此，XML 只能表达数据的语法，而不能表达机器可理解的形式化的语义，为此语义网引入了 RDF 的概念。

（3）资源描述层(RDF+rdfschema)

RDF (Resource Description Framework)，即资源描述框架，是 W3C 推荐的用来描述网络上的信息资源及其之间关系的语言规范。

RDF 非常适合描述表达 Web 资源的元数据信息，如题名、作者、修改日期以及版权信息等，具有简单、开放、易扩展、易交换和易综合等特点。由于它们都被称为 Web 资源，所以 RDF 实际以描述任何可以以在网络上标识的信息。因此在资源描述上，RDF 更像是一个数据模型。该模型以“资源-属性-属性值”的形式描述网络信息资源。资源、属性和属性值在 RDF 中分别用术语主语(Subject)、谓语(Predicated) ,宾语(Object)表示，由主语、谓语、宾语构成的子元组(Triple)称为 RDF 陈述或陈述(Statement)。如果把主语和宾语看作是节点，属性看成是一条边，则一个简单的 RDF 陈述就可以表示成一个 RDF 有向图(Graph)。

RDF 定义了一套用来描述资源类型及其之间相互关系的词汇集，称为 RDF Schema(RDFS)。在用 RDF 描述资源时，首先使用 RDF Schema 提供的建模原语构建被描述资源的 Schema 信息，然后再利用此 Schema 描述目标信息资源。通过 RDF Schema 可以定义资源的类型、属性并显式地揭示它们之间丰富的语义关系。

（4）本体层（Ontology vocabulary）

本体的概念最初起源于哲学领域，用于研究客观世界的本质。在语义网范畴内，本体是关于领域知识的概念化、形式化的明确规范。在语义网体系结构中，本体的作用主要表现在:

1)概念描述:即通过概念描述模型来揭示领域知识。

2)语义揭示:本体具有比 RDF 更强的语义表达能力，可以揭示更为丰富的语义关系。

3)一致性:本体作为领域知识的明确规范，可以保证语义的一致性，从而彻底解决一词多义、多词一义和词义含糊现象。

4)推理支持:本体在概念描述上的确定性及其强大的语义揭示能力在数据层面有力地保证了推理的有效性。

(5) 逻辑层(Logic)、证明层(Proof)和信任层(Trust)

在语义网体系结构中，本体层以上的各层统称为规则层。规则层中各层的具体含义是不同的。逻辑层主要描述推理规则，因为它是代理对用户任务进行分解、定位、协调、验证乃至最后建立信任关系的基础，所以它位于规则的最底层。证明层是为保证代理工作的可靠性而提供的一种验证机制，它应用逻辑层的规则以及本体层的数据表达逻辑推理，子任务和代理之间通过交换“证明”而为数据或结论提供可靠性保证。其基本思想是:我所提供的数据和推理是正确的，因为有多个信用好的信息源都认为我是可以信赖的，它们包括在 Proof 数据段中。信任层位于体系结构的最顶层，同时也处在规则层的最上层。通过“证明”交换和数字签名(Digital Signature)技术，可以建立信任关系，保证语义网的可靠性。

2.3 本体理论基础

（1）本体概念

本体，最著名并被广泛引用的定义是由 Gruber 提出的“本体是概念模型的明确的规范说明”[14]。通俗地讲，本体是用来描述某个领域甚至更广范围内的概念以及概念之间的关系，使得这些概念和关系在共享的范围内具有大家共同认可的、明确的、唯一的定义，这样，人机之间以及机器之间就可以进行交流。目前，本体已经被广泛应用于语义、智能信息检索、信息集成、数字图书馆等领域[15]。

本体的理论研究包括概念及概念分类、本体上的代数等，其中最有代表性的是 Guarino 等人对概念及其分类进行的研究工作。本体的本质是概念模型，表达的是概念及概念之间的关系。本体本来是哲学中物理学形而上学的一个分支[16]。从哲学的角度来说，逻辑是抽象的形式，而本体研究事物存在的方式，是具体的内容，因此在哲学上，如果没有本体，则逻辑关于任何东西都只是空洞的抽象，无法进行具体的描述而没有逻辑，本体就只能进行分析、表达和讨论，在抽象上的通性模糊不清[17]。自二十世纪九十年代初，本体概念被广泛地引用到计算机领域，特别是人工智能和知识工程研究中，因为知识工程需要开发一个领域共享的、公共的概念，实现知识共享和重用。在工业领域，本体通常被称为领域模型(Domain model)或概念模型(Conceptual model)[18]，是关于特定知识领域内各种对象、对象特性以及对象之间可能存在的关系的内容理论。通过对应用领域的概念和术语进行抽象，本体形成了应用领域中共享和公共的领域概念，可以描述应用领域的知识或建立一种关于知识的描述。本体的抽象可能是高层次的抽象，也可能针对特定领域的概念抽象。本体已经成为知识工程、自然语言处理、协同信息系统、智能信息集成、Internet 上智能信息获取、知识管理等各方面普遍研究的热点[19]。因此，随着高度结构化的知识库在和面向对象系统中的出现，对于实际应用和理沦研究，本体的标准都变得日益重要。

最近十年以来，各种研究机构和知识工程研究者提出了多种面向、具有细微差别的本体定义。例如 Gruber 的定义[20]强调了本体是知识表示的元级描述:Wielinga 和 schreiber 的定义[21]强调了本体在知识级的形式化，表示应用于可知识化的 Agent 中的知识。其中的定义被引用最多，也是迄今为止最准确的本体的定义“本体是概念模型的明确的规范说明”。概念模型指通过抽象出客观世界中一些现象的相关概念而得到的模型。概念模型所表现的含义独立于具体的环境状态，明确是指所使用的概念及使用这些概念的约束都有明确的定义，形式化指本体是计算机可读的即能被计算机处理，共享指本体中体现的是共同认

可的知识，反映的是相关领域中公认的概念集，即本体针对的是团体而非个体的共识。

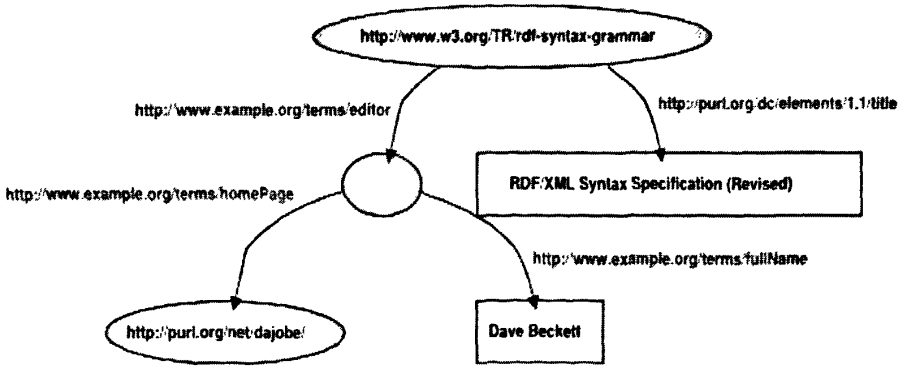
从根本上说，本体的作用是为了构建领域模型，例如，在知识工程过程中，一个本体提供了关于术语概念和关系的词汇集，通过该词汇集可以对一个领域进行建模。虽然不同的本体之间存在一些差异，但它们之间存在普遍的一致性。针对应用领域中一些特殊的任务，知识表达可能还需要一种在很高的普遍性层次上的本体抽象概念。

（2） 本体描述语言

1） RDF 与 RDFS

RDF(Resource Description Framework)、RDF-S(RDF Schema)是 W3C 在 XML 的基础上推荐的一种标准，用于表达网络资源信息。RDF 提出了一个简单的模型用于表达任意类型的数据。这个数据类型由节点和节点之间的带标记的连接弧组成。节点用于表达 Web 上的资源，弧用来表示这些资源的属性。因此， RDF 的数据模型可以方便的描述对象以及它们的关系。RDF 的数据模型实质上是一种二元关系的表达，由于任何复杂的关系都可以分解为多个二元关系，因此 RDF 的数据模型可以作为其他任何复杂关系模型的基础模型。RDF Schema 为 RDF 资源的属性和类型提供了定义良好的词汇表。W3C 推荐以 RDF/RDFS 标准来解决 XML 的语义局限问题。RDF 提供了一个简单但功能强大的模型(Model)来描述资源。RDF 模型定义为:包含一系列节点；包含一系列属性类；模型是一个三元组:{属性类， 节点， 节点或者原生值}；每个数据模型可以看成一个实体——关系图。模型中所有的资源以及用来描述资源的属性都可以被看成“节点”。

下面举一个 RDF 的例子来进行说明。



2-2: Graph for RDF/XML Example

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dc="http://purl.org/dc/elements/1.1/"
         xmlns:ex="http://example.org/stuff/1.0/">
  <rdf:Description rdf:about="http://www.w3.org/TR/rdf-syntax-grammar"
           dc:title="RDF/XML Syntax Specification (Revised)">
    <ex:editor>
      <rdf:Description ex:fullName="Dave Beckett">
    <ex:homePage rdf:resource="http://purl.org/net/dajobe/" />
      </rdf:Description>
    </ex:editor>
  </rdf:Description>
</rdf:RDF>
```

如果从资源描述的角度来看，RDF 已经能够完全胜任，但其提供的建模原语依然过于简单，不能表达丰富的语义关系。比如，RDF 没有提供类(Class)的定义，也没提供机制来约束属性（Property）之间的关系。RDFS 作为 RDF 的扩展，定义了资源的属性类、语法、属性值的类型，定义了资源类以及属性所应用到的资源类，也就是定义了 RDF 的大体框架和原语，声明了由某些标准机构定义的元数据标准的属性类。

RDF Schema 使用一种机器可理解的体系来定义描述资源的词汇。RDF Schema 在语法上使用了 XML Namespace(命名空间)机制。RDF Schema 和 XML Schema 二者在名字上十分相似，根本区别在于 XML Schema 表述的是一个 XML 文档中所使用的标签(tag)的顺序和组合；而 RDF Schema 提供的是对 RDF 模型表示的声明进行解释说明的语义信息。RDF Schema 在 RDF 基础上增加了许多语义原语，用来更进一步增加对资源语义的描述能力，比如类、属性、以及类和属性间的从属关系。常用的 RDF Schema 原语包括：rdfs:Resource、rdfs:Class、rdfs:Literal、rdf:Property、rdfs:domain、rdfs:range、rdf:type、rdfs:subClassOf、rdfs:subPropertyOf、rdfs:seeAlso 等。下面为一个人与车的本体例子，里面包含了许多属性及属性间关系的设置。

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<!DOCTYPE rdf:RDF [
    <!ENTITY rdf 'http://www.w3.org/1999/02/22-rdf-syntax-ns#'>
    <!ENTITY a 'http://protege.stanford.edu/system#'>
    <!ENTITY mv 'http://protege.stanford.edu/mv#'>
    <!ENTITY rdfs 'http://www.w3.org/TR/1999/PR-rdf-schema-19990303#'>
]>
<rdf:RDF xmlns:rdf="&rdf;"
    xmlns:a="&a;"
    xmlns:mv="&mv;"
    xmlns:rdfs="&rdfs;">
<rdfs:Class rdf:about="&mv;MiniVan">
    <rdfs:subClassOf rdf:resource="&mv;PassengerVehicle"/>
    <rdfs:subClassOf rdf:resource="&mv;Van"/>
</rdfs:Class>
<rdfs:Class rdf:about="&mv;MotorVehicle">
    <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdfs:Class rdf:about="&mv;PassengerVehicle">
    <rdfs:subClassOf rdf:resource="&mv;MotorVehicle"/>
</rdfs:Class>
<rdfs:Class rdf:about="&mv;Person">
    <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdfs:Class rdf:about="&mv;Truck">
    <rdfs:subClassOf rdf:resource="&mv;MotorVehicle"/>
</rdfs:Class>
<rdfs:Class rdf:about="&mv;Van">
    <rdfs:subClassOf rdf:resource="&mv;MotorVehicle"/>
</rdfs:Class>
<rdf:Property rdf:about="&mv;name"
```

```
        a:maxCardinality="1">
      <rdfs:domain rdf:resource="&mv;Person"/>
      <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&mv;rearSeatLegRoom"
        a:maxCardinality="1"
        a:range="integer">
      <rdfs:domain rdf:resource="&mv;MotorVehicle"/>
      <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&mv;registeredTo"
        a:maxCardinality="1">
      <rdfs:domain rdf:resource="&mv;MotorVehicle"/>
      <rdfs:range rdf:resource="&mv;Person"/>
</rdf:Property>
</rdf:RDF>
```

2） OWL

正是由于本体语言的要求，加之 RDF(S)表达能力的不足，对于更加丰富的语义信息，比如属性的局部性质(Local scope of properties)、不交类(Disjointness of classes)、类的布尔组合(Boolean combinations of classes)、基数约束(Cardinality restrictions)以及属性的特殊性质(Special characteristics of properties,比如传递性)等，RDF(S)显得无能为力。

这时出现了本体语言 OWL[22]。OWL 在 RDF 和 RDFS 的基础上增加了更多的建模原语来描述性质、类以及它们之间的关系，并更好的支持自动推理。根据不同的需求，OWL 又可以分为三个子语言，OWL Lite、OWL DL、OWL FULL。

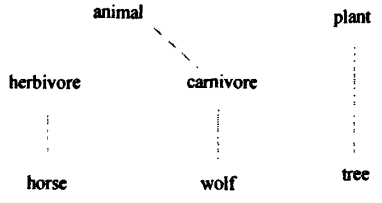
OWL Lite：用于提供给那些只需要一个分类层次和简单属性约束的用户，是 OWL DL 中相对容易实现部分的子集合。

OWL DL： 支持那些需要在推理系统上进行最大程度表达的用户，这里的推理系统能够保证计算完全性（computational completeness，即所有地结论都

能够保证被计算出来）和可决定性（decidability，即所有的计算都在有限的时间内完成）。它包括了 OWL 语言的所有约束，但是可以被仅仅置于特定的约束下,其在语义上等同于描述逻辑 DL。

OWL FULL: 支持那些需要在没有计算保证的语法自由的 RDF 上进行最大程度表达的用户。它允许在一个 Ontology 在预定义的（RDF、OWL）词汇表上增加词汇，过于复杂，对可计算性方面的要求较高，所以所有的推理软件都不能支持 OWL FULL 的所有特性.

下图表示了生物群落中类与子类的层次结构图。根据这个基本模型，按照属性的定义、类的定义所示的本体进行描述。



3-3 一个生物群落的层次结构图

首先定义一个 eat 属性和它的逆属性：

```
<owl:objectProperty rdf:ID="eat">
<rdfs:domain rdf:resource="#animal"/>
</owl:objectProperty>
```

“eat”属性的逆属性“eaten-by”定义为:

```
<owl:objectProperty rdf:ID="eaten-by">
<owl:inverseOf rdf:resource="#eat"/>
</owl:objectProperty>
```

根据图的层次结构，类的定义如下：

```
<owl:class rdf:ID="animal">
<rdfs:comment>animals form a class</rdfs:comment>
</owl:class>
<owl:class rdf:ID="plant">
<rdfs:comment>plants form a class disjoint from
```

animals</rdfs:comment>

<owl:disjointWith="#animal"/>

</owl:class>

这里 owl:disjointWith 用来说明 plant 和 animal 是不相交的。

<owl:clas rdf:ID="tree">

<rdfs:comment> trees are a type of plants</rdfs:comment>

<rdfs:subClassOf rdf:resource="#plant"/>

</owl:class>

树(tree)是一种植物(plant),rdfs:subClassOf 刻画了这种“is-a”关系。

<owl:class rdf:ID="herbivore">

<rdfs:comment>Herbivores are exactly those animals that eat only plants</rdfs:comment>

<owl:intersectionOf rdf:parsetype="collection">

<owl:class rdf:about="#animal"/>

<owl:restriction>

<owl:onProperty rdf:resource="#eats"/>

<owl:allValuesFrom>

<owl:class rdf:about="#plant"/>

</owl:allValuesFrom>

</owl:restriction>

</owl:intersectionOf>

</owl:class>

<owl:class rdf:ID="carnivore">

<rdfs:comment> Carnivores are exactly those animals that eat also animals</rdfs:comment>

<owl:intersectionOf rdf:parsetype="collection">

<owl:class rdf:about="#animal"/>

<owl:restriction>

<owl:onProperty rdf:resource="#eats"/>

<owl:allValuesFrom rdf:resource="#animal"/>

</owl:restriction>

```
</owl:intersectionOf>
</owl:class>
```

在这两个类中，根据它们的语义，分别把它们(herbivore 和 carnivore)定义为动物(animal)和具备特定属性(owl:restriction)的交集(intersectionOf)，herbivore 为草食动物，carnivore 为肉食动物。马(horse)和狼(wolf)的定义为：

```
<owl:class rdf:ID="horse">
<rdfs:comment>Giraffes are herbivores<rdfs:comment>
<rdfs:subClassOf rdf:resource="#herbivore"/>
</owl:class>
<owl:class rdf:ID="wolf">
 <rdfs:comment>wolfs are animals are carnivore </rdfs:comment>
<rdf:subClassOf rdf:resource="#carnivore"/>
</owl:class>
```

从上面用 OWL 定义的本体，它具有更强的语义表达能力，能在一个更高的层次上反映领域的知识结构。而且它还具备一定的规则定义，有利于用推理机进行推理。

（3） 本体构建

在构建本体之前，要先明确目标，即决定构建的本体类型和本体的构建方式。从描述范围来看，本体包括领域本体和公共本体。领域本体和特定的应用相关，描述了现实世界内小范围的一个模型；相反，公共本体包含公共的概念和关系，可用于不同的应用之中。

公共本体作为本体构建的基石，便于扩展、添加新的概念和关系。再次，必须确定是用自顶向下的方式构建本体，还是使用自底向上的方式。自顶向下的方式，从“is-a”继承关系的顶端开始，往下扩展。许多人工构造就是采用这种办法。而在自底向上的方式下，概念和关系是在发现概念、关系时逐步加入的。这种方式更适于自动构建。

构建本体的主要工程思想有骨架法、循环获取法、企业建模法和三种，分别如下：

骨架法[23]是 Mike Ushold 和 Micheal Gruninger 等人提出的，提出的背景是在企业本体的基础之上，也是目前最为大众所接受的方法。

骨架法建设本体的步骤如下:

1) 确定目的和范围。该阶段需要明确本体建设的目的、范围。

2) 本体构建。本体构建可细化为以下几个步骤:

a. 本体获取，即知识获取的过程，发现领域内关键的概念及关系，给出概念和关系的定义描述，识别出用来表达这些概念和关系的术语。

b. 本体译码，决定领域知识在概念模型中的结构，把上一阶段获取的概念用形式化的语言明确表达出来。先要确定本体中用到的基本词汇，比如类(class)、属性(property)、字面值(facet)等，再用合适的知识模型表达这些词汇。

c. 本体集成，重用已有本体，加速本体的开发。这一步比较困难，因为这里涉及到本体异构性的问题，因为针对不同的本体甚至是结构完全不同、概念抽取方式迥异的本体。不过目前有些这方面的研究，也提出了一些思路对冲突进行检测，协调和可视化处理。

3）本体评价。没有具体的说明本体的评价方法，但是认为本体评价是整个本体建设方法论的重要环节。

4）文档化。文档缺乏是知识共享的严重障碍。文档应该包括本体中定义的主要概念、元本体等。某些编辑器可以自动生成文档。

5）各个阶段的指导方针。即设计标准，包括清楚、一致、可扩展、最小本体承诺、最小编码偏差等要求。

循环获取法[24]采用一种环状结构的开发思路，类似于软件工程中的原型法思想，是由 Alexander Maedche 等人提出的。

基本流程如下:

1) 数据源选择:环形的起点，是一个通用的核心本体的选择。任何大型的通用本体(如 Cyc, Dahlgren 的本体)、词汇语义网(WordNet 等)、以及领域相关的本体(TOVE 等)都可以作为这个过程的开始。选定基础本体后，用户必须确定用于抽取领域相关实体的文本。

2）概念学习:从选择的文本中获取领域相关的概念，并建立概念间的分类关系。

3）领域聚焦:去除和该领域无关的概念，建立目标本体的概念结构。

4) 关系学习:除了和基础本体中继承的一些关系，其他的关系只有通过学习的方法从文本中进行获取。

5) 评价:对建设的本体进行评估。然后，可重复迭代上述过程。

企业建模法[25]是 Micheal Gruninger 和 Mark. S Fox 在 TOVE 项目中提出并加以运用的，TOVE 项目的目标是建立一套为商业和公共企业建模的集成本体。

1) 激发场景：应用领域的某些场景可以激发本体的建设，因而给出一个场景有助于理解本体建设的动机，对场景内的问题有全面的认识。

2) 非形式化的能力问题：提出一个本体应该能够回答的各种问题，作为需求，明确所建本体所能提供的功能。通过指明能力问题和场景之间的关系，可以对新扩展的本体进行一定的非形式化的判断。

3) 规范的术语：从非形式化的能力问题中提取出非形式化的术语，然后利用本体形式化语言加以定义。

4) 形式化的能力问题：把非形式化的能力问题用形式化的术语定义出来。

5) 形式化的公理：构成了本体的规格说明。

6) 完备性定理：把本体形式化表述之后，定义了本体完备性的条件。

还有一些方法学比如 IDEF-5、Methodology。其中 IDEF-5 是在结构化分析的思想上发展起来的，而 Methodology 则是结合骨架法和企业建模法的产品，IDEF[26]的概念是在 70 年代提出的结构化分析方法的基础上发展起来的。本体描述获取方法 IDEFS (Ontology Description Capture Method)[27]提供了两种语言形式，即图表语言和细化说明语言来获取某个领域的本体论。DEF-S 提出的本体建设方法包括以下五个活动:(1)组织和范围的确定，包括本体建设项目的目标、观点和语境等;(2)数据收集，主要是对本体建设需要的原始数据的收集;(3)数据分析，主要是对收集到的数据进行分析，为抽取本体做准备;(4)初始化的本体建立:(5)本体的精炼与确认。

（4） 本体映射

给定两个本体，本体映射就是要在他们之间找出语义之间的映射关系。这是一个简单的本体映射的概念。

本体映射能够通过多种方法实现，包括条件规则[28]、函数映射[29]、映射表和程序[30]等。具体过程大体上又可以分为三类:(1)一对一的本体映射，如 OBSERVER[31]，为每一对映射的本体建立一个映射程序，专门处理这两个本体之间的映射。这种方法不需要中间本体，针对性强，但是计算复杂性比较高。(2)通过一个共享本体实现映射，这种方法简单易用，但是建立一个能够被公认和广泛接受的共享本体确实非常困难的。(3)本体簇集，即通过相似度计算将资

源进行簇集，形成本体的概念层次结构。

关于本体映射方法的研究成果已有了一些总结，有斯坦福大学的本体代数方法[32]；基于概念实例 GLUE 方法[33]；本体比较方法[34]；基于统计学的 WordNet 和 EDR 映射方法[35] 等。需要说明的是，每个映射方法往往是多种技术和多种参照对象的结合。

ONION 是斯坦福大学提出的一个本体映射系统。为了解决本体之间的术语的不一致问题，提出了一个半自动的算法在它们之中建立连接规则。ONION 系统拥有自动的连接方式产生器，它通过启发式的匹配器在本体的概念节点之间建立连接。启发式的匹配器可以分为两种类型--迭代的和非迭代的。无论是迭代的还是非迭代的方法，都采用了多种策略。ONION 的主要目标就是在本体的实体中建立连接规则，即通过实体的名称建立映射规则。ONION 中的本体是用概念图来表示的，因此，本体的映射就是基于图的映射。ONION 的主要创新就是，它使用本体的关联关系来进行本体间的互操作，而且它用图形化表示本体。

在基于本体的数据集成中，多个应用本体间不可避免地出现语义冲突，本题映射的研究内容可以在一定程度上解决这些冲突。

第 3 章 基于语义网的博客搜索模型设计

3.1 博客及博客搜索

博客是一个个人性与公共性相结合的媒介，它充分利用了网络双向互动、超文本链接、动态更新、覆盖范围广的特点，其精髓不是表达个人思想或是记录个人的日常经历，而是从个人的角度，来精选和链接互联网上最具有价值的信息、知识和资源[36]。

《华尔街日报》记者佩姬·努南(Peggy Noonan)将博客解释为:“博客是每周 7 天，每天 24 小时运转的言论网站，这种网站以其率真、野性、无保留、富于思想而奇怪的方式提供无拘无束的言论。”

硅谷最著名的 IT 博客专栏作家丹·吉尔默提出:“博客代表着新闻媒体 3.0”，1.0 是指传统媒体或说旧媒体(old media)，2.0 就是人们通常所说的新媒体(new media)或者叫跨媒体，而 3.0 就是以博客为趋势的(we media)的个人媒体或者叫自媒体。”[37]

人们在使用以往的网络沟通工具时，更多地是把网络仅仅作为工具来看待。由于 bbs、msn、email 客观上丧失了人的整体性，使得很多人在不同的工具上表达的内容和方式完全不同，但却无力整合。个人 blog 实现了这种整合。正是由于这种整合，使得个人 blog 上展示了一个活生生的人，原来人在 bbs、email、msn 上分别割裂存在的不同侧面终于可以在个人 blog 上实现了统一。这种统一对于书写者可能仅仅是个整合，感觉只是其它工具功能的综合罢了，但对阅读者的感觉就完全不同。

著名的学者戴夫·温纳认为“博客就是一个人未经编辑的声音”[38]。博客为个人提供了一个广阔的空间来展现个人风采，博客作者在轻松的状态下畅所欲言，将自己的兴趣爱好、行为、思想，将自己认为最有价值的东西，从自己的角度完美地展现出来，这也正是博客的魅力所在。

读者从个人 blog 中似乎可以触摸到书写者的内心世界，感到博客主在面前跟他讲述自己的所思所想、心得体会，从博客中全方位地了解博客主的性格特点，亲切感和真实感是阅读以往任何形式的网络资料和沟通工具所根本不具有的。因而，从读者角度完全可以感到书写者就是一个‘活’生生的人，他就是在聆听‘活’在网络中的书写者谈话，并把自己的评论写在上面和书写者进行

交流。这种感觉的可以从格式塔心理学强调人们认知的完形特点得到解释。

当然，如果书写者自身也认识到了自己的个人blog对阅读者的影响和意义，他也就会调整自己书写的方式和内容，更好地表达和整饰自己的所思所想，让‘活’在网络中的自己与别人进行更好的交流。 如果每个 blog 书写者都有了这样的意识，并努力扮演好自己的角色，这样就可以真正地组成以人为节点，而非以字节或知识为节点的一个‘活’的网络社会。这个网络社会有别于游戏世界和网络社区的特点，它是真正意义上现实社会在网络上的反映。这也为人们整合网络和现实提供了基础。从这个意义上讲，以 blog 为代表的真实化、公开化使用网络的趋势，将真正地为开拓网络积极的现实意义提供了广阔的可能性。

从博客世界所引发的美国多数党领袖特伦特·洛特的下台、《纽约时报》的假新闻丑闻、约翰·克里的谎言、“拉瑟门”事件和 2004 年总统竞选等重大事件中，人们体会到博客世界的真实力量。博客传播，第一次真正实现了“所有人面向所有人传播”的理想和梦想，是真正的大集市模式。未来社会的主流传播模式不是大教堂模式垄断的时代，而是大教堂模式和大集市模式双重模式主导的时代。这种全新的传播模式的变革将逐渐影响互联网、传媒、生活、政治、经济、社会、文化等各个方面和层面。知名博客作家有机会访问政府高层来洞悉社会热点，进行跟踪报道，增加政府工作的透明度。2009 年 9 月 7 号，俄罗斯总统邀请国内最聚人气博客专家鲁斯特姆·阿达加莫夫在克里姆林宫报道他与印度总统帕蒂尔的会谈。博客传播将极大弥补、纠正和疗治传统大众媒体造成的后遗症，推动人类文明的进步，促进人类社会和谐发展。

由于博客日志具有即时性 ，内容常常更新的特点 ，博客搜索引擎具有实时跟踪和监控机制以及更新自动提醒功能 ，这使得它更新周期更短 ，信息资源具有更强的时效性，有时甚至比传统新闻来得更快，分析更透彻。现在博客搜索正在飞速的发展中，搜索相关性较以前有很大的提高，发展前景很好。博客搜索的最新发展特点有如下几点：

1）分类化。博客搜索引擎将搜索到的博客信息分类，可以分别按时间、地区、语言、话题、热门程度等进行分类，这在很大程度上整合了一些博客资源，提高了检索的效率，也方便了用户的检索。

2）个性化。博客搜索引擎根据注册用户的搜索记录信息来向用户推荐一些相关博客信息，用户还可以进行个性化的定制，搜索引擎会定期以 E-mail 的方

式向用户发送用户定制的信息。

3）媒体化。博客搜索引擎将传统新闻资源与博客资源进行整合，并对博客中集中讨论的热门事件、人物、故事等进行集中收录，使话题更有针对性，正聚集庞大的用户群，对传统媒体产生极大冲击。

4）产业化。博客的流行也催生了一个重要的行业博客营销，很多企业都开始建立企业博客网站来进行网络营销，职业博客写手也凭自己独到的见解吸引了大量广告商的注意。

由于博客的快速兴起，也引发了许多对博客及博客搜索的研究，Makoto Nakatsuji, Yu Miyoshi, and Yoshihiro Otsuka[39]将用户博客项分类成服务领域本体，并采用自顶向下的方法，根据兴趣度抽取那些从语义上表达用户兴趣的领域作为类别中的一层。接下来，采用自底向上的方法，用户修改他们的兴趣本体来在更新他们更多的兴趣细节。Knud M¨oller, Uldis Boj¯ars, and John G. Breslin[40]定义了作为语义数据类型的结构和内容相关的数据，它们在博客领域是相关的，语义博客将允许博客空间的角色使用方便数据交换的新方式。在细节上展示了这两类数据的实质，讨论了以一种方便而且简洁的方式为用户生成这样的数据，怎样在 WEB 上显示，以及怎样从一个博客用户的角度去利用它。Yoonjae Jeong and Dongman Lee 提出了一种在博客空间利用权威估计的快速自我价值搜索模式[41]，采用一种快速自我中心的搜索模式，减少搜索空间给更重要的博客。在有关系的博客空间中，它们之间的引用和评论对于评价该博客的权威性很重要。Yun Chen, Flora S. Tsai *, Kap Luk Chan[42]对商业博客搜索和挖掘的机器学习技术进行研究，提出了一种可能的模型，用了两种机器学习技术，潜在语义分析和可能潜在语义分析，并在他们的商业博客数据库中实现了这种模型。Sachit Rajbhandari, Frederic Andres, Motomu Naito[43]设计了一种可视化话题地图的数据模型，利用 Google 地球技术将 blog 空间信息显示在地图，用户可以在地图上在某一地区找自己感兴趣的信息，直接连接到博客中。将 RSS 技术与这种数据模型进行结合，增强了语义表达能力和为可协作的信息提供交换支撑，用户可以通过空间，地图和多维数据等来了解某一话题博客的即时信息。这些研究丰富了博客的内涵，从多角度分析博客的应用，使我对博客及博客搜索有了更加深入的认识。

3.2 博客关键技术

3.2.1 RSS

（1）RSS 概论

RSS 的英文全名是 Really Simple Syndication(真正简单联合供稿系统)，是一种用来聚集搜集新闻标题或是提供网页内容的格式，以 XML 可延伸标记语言为基础。如今 RSS 最广泛的使用在将网站的最新头条新闻或内容有效率的整理出来，提供需求者订阅，用户可以利用 RSS 链接自己喜欢的其他网上信息，比如搜狐新闻、新浪体育、网易财经、热门小说等信息，在信息浏览上同步的，用户可以浏览到最新的更新信息，构建一个个性化的个人在线阅读站点。并且也是一种网站和网站之间共享内容的简易方式。也就是说 RSS 不但可以描述网站上的新闻格式，以及网志(Web loggers 或 bloggers)，更可以藉由 RSS 让别人更容易发现你的网站以及追踪新闻的来源。一些大的网站如 BBC, CNET, Disney, Wired 等的网站信息都是透过 RSS 来当作信息传播的媒介，提高网站的信息流量。

（2）RSS 标准

RSS 是一种起源于网景的技术，将用户订阅的内容传送给他们的通讯协同格式(Protocol)。RSS 目前广泛用于网上新闻频道，blog 和 wiki，主要的版本有 1.0 和 2.0。Blog 从一个专业群体开始，逐步成为了网络上最热门的新话题。而 RSS 成为了描述 Blog 主题和更新信息的重要方法。于是 RSS 这项技术被著名 Blogger/Geek 戴夫·温那(Dave Winner)的公司 UserLand 所接手，继续开发新的版本，以适应新的网络应用需要。新的网络应用就是 Blog，因为戴夫·温那的努力，RSS 升级到了 0.91 版，然后达到了 0.92 版，随后在各种 Blog 工具中得到了应用，并被众多的专业新闻站点所支持。在广泛的应用过程中，众多的专业人士认识到需要组织起来，把 RSS 发展成为一个通用的规范，并进一步标准化。一个联合小组根据 W3C 新一代的语义网技术 RDF 对 RSS 进行了重新定义，发布了 RSS 1.0，并把 RSS 定义为“RDF Site Summary”。这项工作并没有与戴夫·温那进行有效的沟通，而戴夫则坚持在自己设想的方向上进一步开发 RSS 的后续版本，也并不承认 RSS 1.0 的有效性。RSS 由此开始分化形成了 RSS2.0 和 RSS 1.0 两个阵营，也由此在专业人群中引起了广泛争论。

因为有着争论的存在，所以一直到今天，RSS 1.0 还没有成为标准化组织的真正标准。而戴夫·温那却在2002年9月独自把RSS升级到了2.0版本，其中的定义完全是全新的模式，并没有任何 RSS 1.0 的影子。RSS 2.0 能被人们广泛采用的一个原因就是因为它的简单易用。它将网站看作一系列频道（channel）的组合，一个频道包含多个项(item)。对 item 定义了三个必须元素[44]:<title>、<link>和<description>。其中<title>是网站或栏目的名称，一般与网站或栏目的页面 title 一致，<link>是网站或栏目的 url，<description>是对网站或栏目的简要描述。除此之外，RSS2.0 还定义了一些可选元素，比如<language>(语言)、<copyright>（版权）、<webMaster>和<managingEditor>（RSS 文件提供者）<lastBuildDate>（最后一次修改时间）等。下面是一个 RSS2.0 版本新闻订阅的例子：

```
<?xml version="1.0"?>
    <rss version="2.0">
    <channel>
    <title>Lift Off News</title>
    <link>http://liftoff.msfc.nasa.gov/</link>
    <description>Liftoff to Space Exploration.</description>
    <language>en-us</language>
    <pubDate>Tue, 10 Jun 2003 04:00:00 GMT</pubDate>
    <lastBuildDate>Tue, 10 Jun 2003 09:41:01 GMT</lastBuildDate>
    <docs>http://blogs.law.harvard.edu/tech/rss</docs>
    <generator>Weblog Editor 2.0</generator>
    <managingEditor>editor@example.com</managingEditor>
    <webMaster>webmaster@example.com</webMaster>
    <ttl>5</ttl>
    <item>
    <title>New York City</title>
  <link>http://liftoff.msfc.nasa.gov/news/2003/news-starcity.asp</link>
    <description>How do Americans get ready to work with Russians aboard the
    International Space Station? They take a crash course in culture, language and
    protocol at Russia's Star City.</description>
```

<pubDate>Tue, 03 Jun 2004 09:39:21</pubDate>
<category>IT</category>
<author>gates</author>
<guid>http://jeky1983.nasa.org/2004/06/03.html#item532</guid>
</item>
<item>
<title>Space Exploration</title>
<link>http://jeky1983.msfc.nasa.gov/</link>
<description>Sky watchers in Europe, Asia, and parts of Alaska and Canada will experience a partial eclipse of the Sun on Saturday, May 31st.
</description>
<pubDate>Fri, 30 May 2003 11:06:42 GMT</pubDate>
<guid>http://jeky1983.msfc.nasa.gov/2003/05/30.html#item554</guid>
</item>
<item>
<title>The Engine Doing Great</title>
<link>http://liftoff.msfc.nasa.gov/news/2003/news-VASIMR.asp</link>
<description>Before man travels to Mars, NASA hopes to design new engines that will let us fly through the Solar System more quickly..</description>
<pubDate>Tue, 27 May 2003 08:37:32 GMT</pubDate>
<guid>http://www.zhangzhilong.cn/rss.xml</guid>
</item>
</channel>
</rss>

RSS1.0 与 2.0 在语法上有些区别，它是基于本体描述语言 RDF 的，能表达更丰富的语义，但设计上也更复杂，主要表现在以下几点：

1)RSS 1. 0 是基于 RDF 技术的，所以所有的内容都嵌套在<rdf:RDF></rdf:RDF>中。

2）每个<item>都有一个 rdf:about 属性，进行标识。

3）在每个<channel>中都有个子元素<items>，其下列出了 channel 中所有的 item。下面是一个 RSS1.0 的例子。

```
<!xml version="1.0"?>
<rdf:RDF xmlns:rdf=" http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns=" http://purl.org/rss/1.0/"
xmlns:dc=" http://purl.org/dc/elements/1.1/" >
<channel rdf:about=" http://example.com/news.rss" >
<title>Example Channel</title>
<link>http://example.com</link>
<description>My example channel</description>
<items>
<rdf:Seq>
<rdf:li resource=" http://example.com/2002/09/01" />
<rdf:li resource=" http://example.com/2002/09/02" />
</rdf:Seq>
<items>
<item rdf:about=" http://example.com/2002/09/01" >
<title>News for September the First</title>
<link> http://example.com/2002/09/01</link>
<description>today is a big day</description>
<dc:date>2002-09-01</dc:date>
</item>
<item rdf:about=" http://example.com/2002/09/02" >
<title>News for September the Second </title>
<link> http://example.com/2002/09/02</link>
<description>today is a great day</description>
<dc:date>2002-09-02</dc:date>
</item>
</items>
<rdf:RDF>
```

究竟让一个越来越普及的数据格式成为一个开放的标准，还是被一家公司所定义和控制，成为了争议的焦点。戴夫·温那并没有为自己辩解，他的观点是 RSS 还需要进一步发展，需要专业人士更明确的定义，不过恐怕这种轻描淡

写不能消除人们对 RSS“被一家商业公司独占”的担心。

3.2.2 Atom

由于 RSS 存在着版本号混乱，表示方法不一致，定义贫乏，不是一个真正的开放标准等问题，另一组开发人员决心通过定义新的摘要规范来与 RSS 名字的随意性决裂，以解决这种混乱的问题。这个解决方案称为原子(atom)项目。Atom 希望提供一个清晰的版本以解决每个人的需要，其设计完全不依赖于供货商，任何人都可以对之进行自由扩展。有人认为 RSS 和 Atom 之间共同点多于不同点，它们的标准竞争可能会分裂市场。还有很多 blogger 发帖表示说 RSS 是为网站内容聚合而生，而 Atom 是为博客聚合量身定做的[45]。

Atom 是基于 XML 的，下面是一个 Atom 的例子，

```
<title>Example Feed </title>
<link href="http://ecample.org"/>
<updated>2009-10-20T19:20:09Z</updated>
<author>
<name>Jeky</name>
</author>
<id>urn:uuid:123w-123aw-12qwe-qwesd2</id>
<entry>
<title>Atom-Powered Robots RunAmok</title>
<link href="http://ecample.org/2009/10/20/atom03"/>
<id> urn:uuid:qwe-dfs343-3ewr-23ewr</id>
<updated>2009-10-20T19:20:09Z</updated>
<summary>some text.....</summary>
</entry>
```

3.2.3 其他相关技术

SNS（Social Network Service），简称社会化网络软件，是 Web 2.0 体系下

的一个技术应用框架，基于“六度分隔理论”运作，放在 Web2.0 的背景下，每个用户都拥有自己的 Blog、自己维护的 Wiki、社会化书签或者 Podcast，用户通过 TAG、RSS 或者 IM、邮件等方式连接在一起，“按照六度分隔理论，每个个体的社交圈都不不断放大，最后成为一个大型网络，这就是社会化网络”。

TAG（标签）是一种更为灵活、有趣的日志分类方式，博客主可以为每篇博客日志添加一个或多个标签(Tag)，然后就可以在博客网络上看到所有和自己使用相同 Tag 的日志，并且由此和其他博客用户进行沟通和联系。同时还可以与著名的全球博客运营商 Technorati 进行合作，把博客主的标签发到全球 Blog 空间，这样人们就可以与全球的博客用户进行交流，和他们一起分享博客带来的快乐。

3.3 系统整体框架

对于博客搜索，它的流程同其他网页搜索流程差不多，只是有些处理过程不一样而已，在其中加入了语义网的相关技术，将数据源与本体相关联起来，这样在进行检索的时候通过本体的扩展和推理来对数据源进行搜索，然后对结果进行整合，最终显示在用户的面前。系统功能模块主要分为：网页信息的抓取、网页索引建立、网页索引结果排序、网页检索工具和接口。如图一：

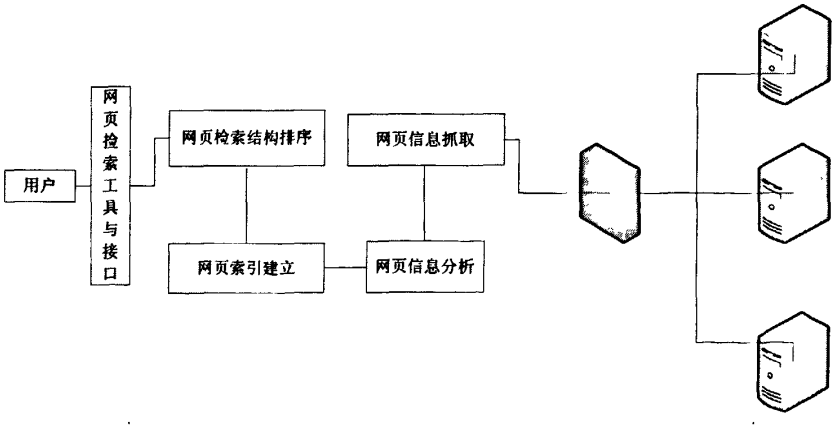


图 3-1 一般搜索引擎工作流程图

本文设计的搜索引擎与一般的搜索系统不同，它加入了语义网技术，将本体应用于搜索过程，在建立索引阶段，对抓取的博客页面进行分析，提取关键词，将关键词集合与本体进行映射，这样得到该文本与本体的映射。还可以利

用本体的语义功能对搜索范围进行适当扩展和推理，达到进一步提高搜索查准率和查全率的目的。当然在这个过程中，要引入一些算法模型来进行量化分析。经过上述分析，得出了一个系统功能模块的总体框架图。如图 3-2：

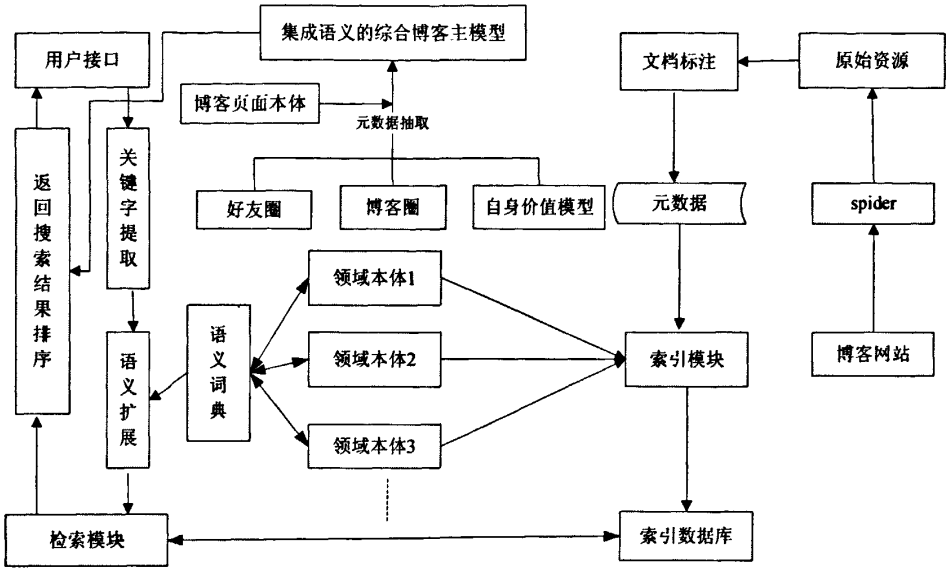


图 3-2 系统整体框架

3.4 系统子模块构建

系统共分为四个模块，分别为原始资料搜集模块、索引建立模块、集成语义的综合博客主模块和用户检索模块。它们协同工作来完成整个搜索过程，系统功能模块图如下：

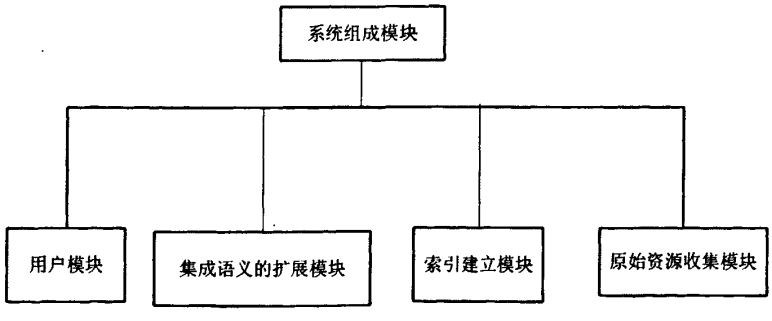


图 3-3 系统功能模块

3.4.1 原始资源收集模块

这一模块最主要的是网络蜘蛛的设计，网络蜘蛛主要负责在互联网上抓取海量数据。系统设计者必须有效地提高网络蜘蛛的性能和效率。才能满足系统数据下载需求。网络蜘蛛的设计一般包括以下一系列优化策略和原则。

1)对等待下载的 URL 进行排重，避免重复下载。

2)增加多个工作队列，提高系统并发能力。工作队列主要有：等待队列、处理队列、成功队列、失败队列。

3)利用网页 proxy 缓冲，检查是否需要从远程下载，减少不必要的传输。

4)同一站点的 URL 尽量映射到同一线程处理，避免同时访问给被访问站点带来负担。

资源搜集模块的体系结构如下：

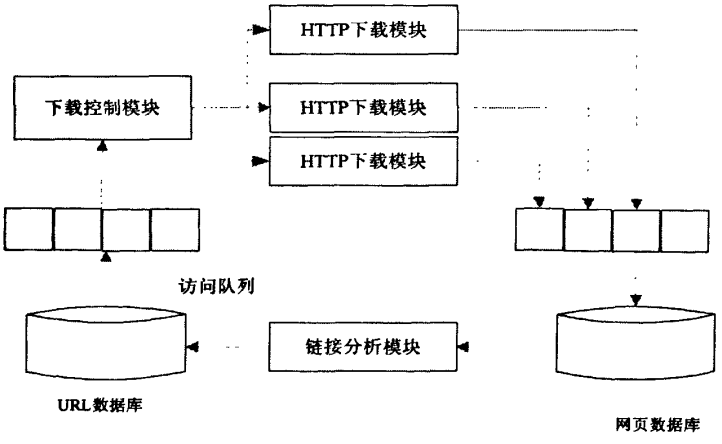


图 3-4 资源搜集体系图

网站的页面是不断更新变化的，网络蜘蛛要定期更新自己储存的数据。这种更新能解决内容陈旧、访问死链的问题。搜索引擎会根据网站的特征、搜索系统的目的来动态改变更新周期。内容重要更新频繁的网站会有比较短的更新周期。长期不修改、不维护的网站会有比较长的更新周期。

网络蜘蛛可以根据搜索的需要，设定博客搜索的范围，在几家大型门户的博客站点，比如网易、搜狐、新浪、雅虎等，还有就是一些专业性比较强的博客网站，很多有着共同兴趣爱好的人在里面建立自己的博客圈来进行更好地交流。另外，还有企业博客这块可以进行挖掘，现在很多企业也意识到博客对于

宣传公司形象和进行博客营销的巨大潜力。

网络蜘蛛的设计比较复杂，这里采用比较流行的 Nutch 网络蜘蛛，它按照一定的结构存储数据，便于索引器建立索引，一般的网络蜘蛛只负责下载页面并按照一定格式存储网页，并不关心后续系统如何处理。下面说下它的工作流程：

网页下载是由 Crawl 命令来完成的，也可以使用底层的 admin、inject、generate、fetch 和 updatedb 命令组合来完成的，其实 Crawl 也是调用底层命令的对应函数来实现的。网络蜘蛛的高效工作还需要良好的存储结构支持。这些存储结构用来存放每个 URL 的具体信息、已下载页面的信息、数据索引等。各个功能模块与数据存储不断协调工作，不断自动下载页面，直到满足系统预设的停止条件，完成整个下载任务。具体的工作流程如下：

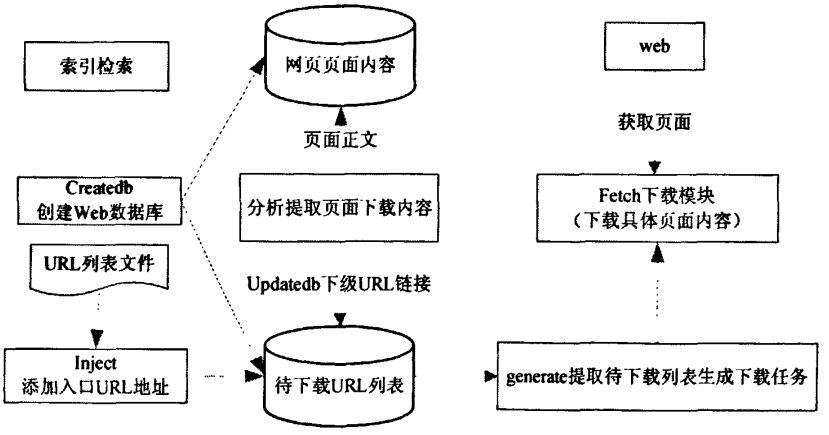


图 3-5 Nutch 网络蜘蛛结构图

Nutch 中主要有以下 5 个命令：

Admin:用来创建一个新的 Web 数据库。建成的数据库包含目录和数据存储结构，初始数据状态为空。其中的 URL 数据库，用来存放页面地址相关信息。

Inject：添加数据下载的入口链接。首先读取给定的纯文本格式文件，获取 URL 列表，作为入口地址添加到已有的 Web 数据库中。

Generate:生成待下载的 URL 列表。按照 Web 数据库格式提取未下载的 URL，以 fetchist 形式给出，为下载做好准备。

Fetch: 按照 HTTP 协议访问互联网，获取网页数据具体内容。下载过程由下载列表和操作参数控制，直到下载完毕。

Updatedb:从已下载文件中获取 URL 链接，更新 Web 数据库，添加到已有的 Web 数据库。

下载后数据存储主要以目录文件形式存放。具体内容包括 Web 数据库、数据段和数据索引。

为满足不同的类型的需求，Nutch 提供了两种工作模式。分别为局域网抓取和互联网抓取，它的设计目标是满足不同应用环境，实现搜索引擎系统对信息下载数量、更新频率、下载策略的不同要求。对于博客搜索来说，一般博客页面的层次不超过三级，属于垂直搜索的一种，比较适合采用深度优先的搜索策略。

3.4.2 索引建立模块

索引模块是搜索系统中非常关键的一个模块，直接影响到搜索的效率。索引主要用于从大量文件中快速查找某个单词或词语，完成信息索引建立、维护和管理功能的软件叫索引器。索引器实际上是一个文本信息处理系统，通常采用倒排文件索引构造索引系统，经过索引后的数据按照优化格式存放，便于快速加载和检索。

文本索引的实现方法很多，当前主流的索引技术是倒排索引、后缀词组索引和签名文件三种，后缀索引比较适合于短语查询，签名文档技术目前使用比较少。倒排索引是一种高效的索引组织方式，能够很好的支持多种检索模型，它采用字或者词作为索引项，非常适合关键词搜索。

这里要建立的本体意群到页面间的索引，它不同于传统的索引，引入了本体的概念，对索引的建立更加规范化，在语义上组织资料，这样既减少了服务器的运载负荷量，也提高了检索的准确性和检索效率。具体的流程图如下：

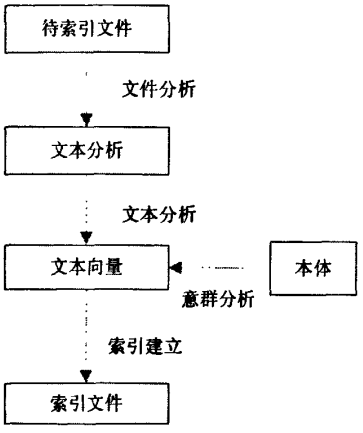


图 3-6 索引流程图

它的索引格式可以表示为如下：

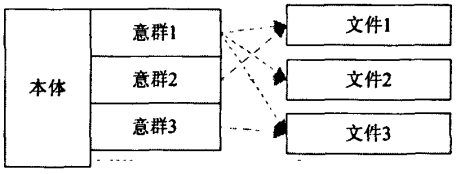


图 3-7 本体意群到文本的映射图

建立本体意群对文本的索引，意群与文本间是多对多的关系，一个文件可能包含多个意群，一个意群也可以对应多个文件。还存在一个问题，随着建立索引文件的增加，意群的数量也随之增加。意群如何界定，当意群的数量越来越多的时候该如何处理，这里采用了定期对意群进行合并的操作来限制意群的数量，对意群间进行相似度的计算，将相似度高的意群进行合并，具体操作会在第五章中的关键技术中加以详细的说明。

3.4.3 集成语义的综合博客主模型

博客是一个个性展示的平台，它的特点就是涵盖广，里面有好友动态，临时访问博友信息，还有博客所在博客圈信息，这些信息综合起来才能反映一个博客的总体质量，为博客页面进行排序提供一个更加公平合理的方式。这个模块主要由好友圈模块、博客圈模块和自身价值模块组成，这里仅列出一些主要信息，好友圈模块主要包含如下信息：好友最近访问次数、日志更新频率、分享频率等，博客圈模块包括博客圈主题、发表文章频率及平均回复率、圈内人气和圈的知名度等，自身价值模块包括发表日志频率及回复率、分享内容频率

等。

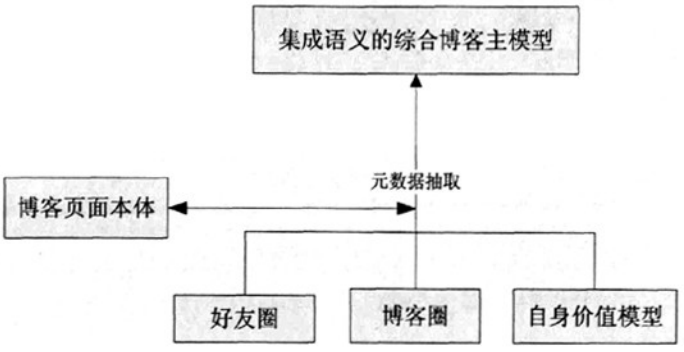


图 3-8 集成语义的综合博客主模型

这个集成语义的综合博客主模型，还需要一个博客页面本体来区别不同博客站点间概念定义的差别，为不同博客间的相似概念提供了一个解释平台，从而更大程度地集成博客间的信息，提高对博客间隐含信息的挖掘效率。

对于博客页面本体的构建，主要是收集博客页面的主要概念和属性，建立一个不同博客站点对不同概念表达方式的一种解释平台，为博客页面分析提供一个标准。下图为一个博客页面的本体概念和属性结构图。

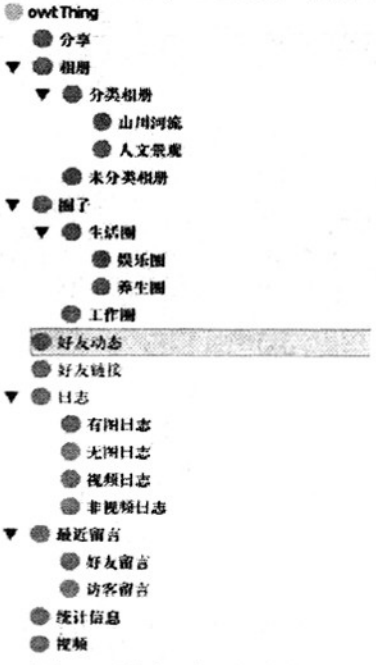


图 3-9 博客页面本体类图

属性的定义为下图：

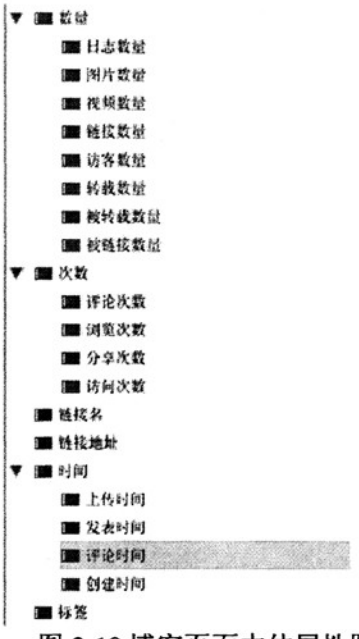


图 3-10 博客页面本体属性图

上面的设计是以简单的方式来构建的，省略了本体的同义词的添加，没有涉及一些本体间复杂的关系，只是简单本体的表现形式。

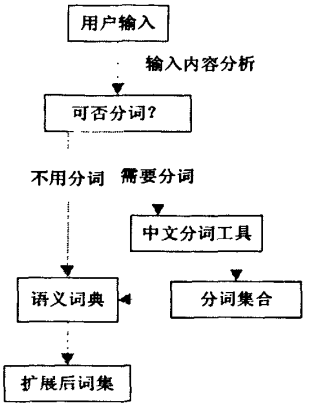
4.4.4 用户检索模块

（1）基于语义的扩展模块

对于系统的语义扩展模块，应该能对用户的输入进行分解和扩展，这里要利用语义词典的扩展功能，它包含了各领域概念的上下位、同义词、属性、整体部分等关系，可以大致确定用户的搜索意图和领域，从而找出与领域本体中的意群相对应的博客页面。

本体是一个知识领域，它能使有些概念被机器所识别，这样可以大大提高机器处理的效率。现有推理机 Racer，它能解决复杂的推理达到挖掘潜在信息的目的。通过本体对经过分词后的关键词进行扩展，找出这些词的上位词，下位词，同义词等关系词，来对这些进行扩展，找出相关的意群，这样可以搜索到一些相关的博客页面，对于搜索的全面率很有帮助。特别是当用户想了解一个热门话题的整个来龙去脉以及来自博客上最及时的各种不同的声音时，这种扩展就显得十分重要。当然还有一些情况也需要考虑到，本体的推理机制在不断

的发展，在初始阶段难免会有一些问题存在，推理的合理性方面也不可能做到完全无差错，这就需要将不确定性也考虑进去，适当对语义进行扩展，这样设计在整体上是合理的，虽然有时会包含些相关性很小的信息，但这样的情况会随着本体推理规则的不断完善而得到改善，逐渐提高用户的搜索满意度。



3-11 基于语义的关键词扩展流程图

（2）排序模块

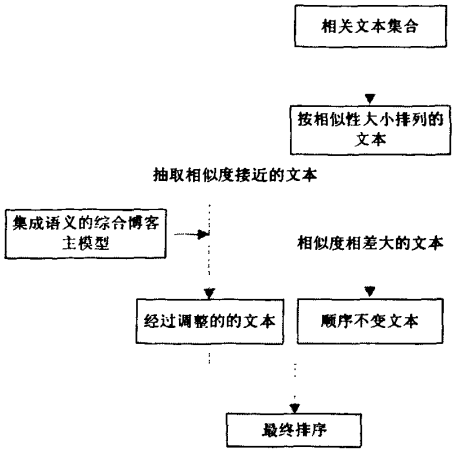
随着博客页面成爆炸式的增长，涌现出越来越多的草根博客，他们有着自己独特的视野和对社会的关注度，成为继明星博客后的一大亮点，如何去发现这些有影响力的草根博客，排序显得很重要，如何公平合理的让这些草根博客的精彩文章能被挖掘出来，是我研究的重点。排序模块的设计要利用集成语义的综合博客主模块，这里分两次排序，一次初排序，一次排序调整，具体的排序流程如下：

1）首先，得到前面根据意群相似度大小得到的文本，按照大小进行排列。

2）抽取其中相似度很接近的文本集合，其余博客页面的排序不变。由于这些集合中有些是分散的，这样就要对每个相似的文本集合所在的博客页面进行模型分析，得出该博客页面的总体价值。

3）最后将经过调整的文本集合放到原来抽取的位置，得到调整后的排序页面。

排序流程如下图：



3-12 博客页面排序流程图

3.5 系统搜索流程分析

根据前面几个子模块的构建，对于整个系统工作的流程大致清晰了，下面是搜索流程介绍。

1）首先是网络蜘蛛按照设计的策略定期地在博客站点去爬取资源，加入原始资源库。

2）原始资源有页面，PDF 文档，XML, .TXT 文件等，这些文件需要进行文档标注处理，抽取出文本信息和页面附加信息（比如标题，URL,日期，链接等，对于建立索引很有帮助）。

3）接着就需要对得到的上述信息进行分词处理，得到元数据，里面包括后续要建立索引的项，对应文本，偏移量，出现次数。

4）下面就要进入索引的建立阶段，建立领域本体意群到文本间的映射，这里要考虑多对多的关系，一个意群可以对应多个文本，一个文本也可能是跨领域的。

5）建立好索引数据库后，就可以进行信息检索了，用户输入搜索内容，经过分词模块处理，得到关键词，通过语义词典进行语义扩展，然后与本体意群进行匹配，得出相关的文本集合。

6）最后进入排序阶段，结合集成语义的综合博客主模型，对步骤 5 中的文本进行排序，显示给用户。

第 4 章 关键技术与算法研究

4.1 博客本体的构建

博客本体的构建，需要考虑到几方面的问题，第一，它必须包括博客领域常用概念及其相关概念属性；第二，它必须涵盖社会生活的各个领域，包括政治、经济、文化、体育、教育、医疗等。基于上述两点，构建一个横跨多领域的博客本体确实很困难， 这里结合骨架法，采用混合本体的方式来构建，建立一个博客领域的语义词典，形式上参考 Wordnet,，然后分别构建各领域本体，它们共享语义词典，构建分以下几个步骤来进行：

1） 确实本体应用的目的，主要是建立领域共享知识的集合。

2） 确定本体涉及的领域及专业词汇集，构建本体的大致框架。

3） 结合领域专家的建议，逐一列举各领域的主要概念、概念间关系、属性及属性特征等。

4) 对领域内的概念间关系、概念属性间关系和属性间关系进行挖掘，制定新的规则。便于后续的推荐和半自动构建本体。

5) 本体评价，构建的本体要经过领域专家的评价，并发布到网上供大家应用来测试它的完备性，人们可以对它进行改进，再发出新的版本或提出修改的建议。通过这种方式来不断完善本体的完备性和合理性。

6) 本体学习，通过抽取特定文档中的相关概念和属性进行学习，这样来扩充本体的概念范围。

以上几个步骤描述了领域本体构建的整个过程，博客本体间的联系可表示为如图 4-1：

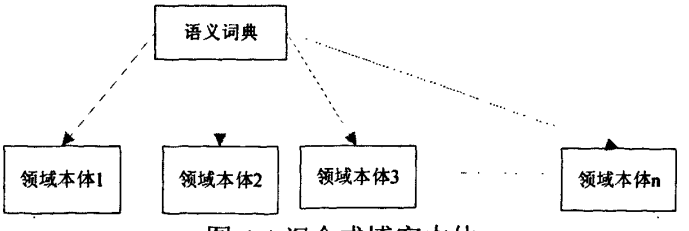


图 4-1 混合式博客本体

每个领域本体共享语义词典，这样避免了领域本体间的直接复杂的映射关系，但对语义词典的设计提出了更高的要求。构建领域本体的流程图如下：

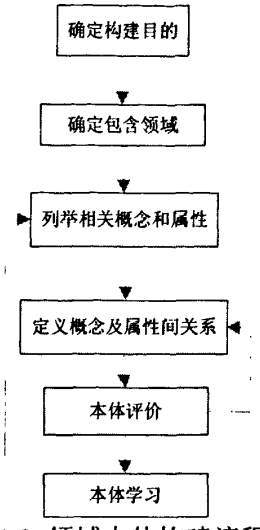


图 4-2 领域本体构建流程图

目前比较流行的是本体编辑工具是 Protégé 2000，它是由斯坦福大学的 Stanford Medical Informatics 开发的一个开放源码的本体编辑器，它是用 java 编写的，可扩展性好。该版本可以建立本体项目，既可存储本体，也可以添加实例，功能很强大。本体结构以树形的层次目录结构显示，用户可以通过点击相应的项目来增加或编辑类、子类、属性、实例等，使用户在概念层次上设计领域模型，所以本体工程师不需要了解具体的本体表示语言。并且它支持多重继承，并对新数据进行一致性检查，并且具有很强的可扩展性，对于本体的研究有很大的帮助。

语义词典是一个包含语义关系的词典，它的设计种类有很多，在这里介绍的语义词典是基于 WordNet 而设计的，它就像一棵树的形状，每一个树的一个结点都是一个同义词集合，它们共用一个 synsetid，结点和结点之间的关系就是通过这个 synsetid 来发挥作用，这里是基于词和词之间的上下位来建立总体联系（如图一），有些词有整体部分关系，属性等关系，也可以扩充这个词典，为里面的词添加关系。语义词典的结构图如下：

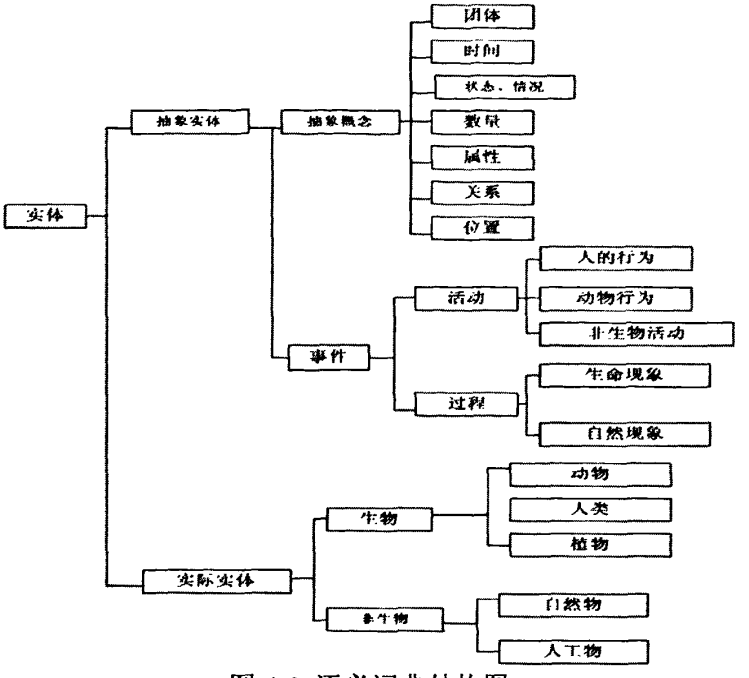


图 4-3 语义词典结构图

语义词典是一个博客相关概念的词汇集合，可以向里面添加新的概念和定义新的关系来扩充语义词典，来不断丰富语义词典的内涵。语义词典的结构定义好了之后，就需要收集相关领域的词汇，对它进行充实。随着词汇量的不断增加和人们表达习惯的变化，会出现越来越多的新词汇，仅凭人为的添加是很难满足要求的，这就需要语义词典具有半自动扩充词汇的能力，通过分析文本和实例，来扩充新的词汇。通过分析得出概念的属性，当从概念上不能判断一个概念时，可以从概念的属性来进行分析，比较新概念与已定义概念的属性，得出一个相似性矩阵，通过算法分析共有属性与相异属性的比例来确定该概念的最相似概念，来建立关系，这是一个离线自学习的过程。可以分为两步来进行：

1）经过文本分析的概念可以先添加进储备词汇库，经过语义词典不断的学习，来确定储备概念的属性关系，等到了一定程度，比如找到和储备概念的属性相似度在 90%上就可以确定它的关系，添加进语义词典。

2）通过学习，在一定时间内找不到合适的相似性概念时，可以新建一个新的概念，添加进语义词典。

4.2 基于本体意群的索引分析

基于本体的索引模块不同于传统的索引，它将本体技术运用于索引的建立过程，建立文档到本体的映射，分三个过程来进行。

（1）首先对抓取的网页文本资料进行分词，得出关键词列表，包括关键词在文中出现的频率、位置等信息，这样可以结合一定的算法，得出关键词的向量模型，$(k_1,w_1;k_1,w_2;.....k_n,w_n;)$ $k_i(0<i<n)$是关键词，$w_i(0<i<n)$是关键词在文档中的权重，然后建立关键词对文本的映射.

(2) 第二步是建立文本到本体的对应，通过文本关键词与本体的映射，得出文本的语义分类和领域相关度。领域本体$D_i(0<i<m)$，每个本体都是一个知识集合，它们之间也有联系，通过将文本关键词向量模型与每个领域本体进行对比，将文本关键词与领域本体概念进行相似度计算，得出相似向量矩阵，

$\begin{pmatrix} k_1d_1 & \cdots & k_1d_m \\ \vdots & \ddots & \vdots \\ k_nd_1 & \cdots & k_nd_m \end{pmatrix}$其中$k_id_j$ $(0<i<n,0<j<m)$是指文本中第i个关键词在领域本体D_i中的最大相似度。$\begin{pmatrix} c_{11} & \cdots & c_{1m} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nm} \end{pmatrix}$是对应于本体中的关键词列表，

c_{ij} $(0<i<n,0<j<m)$是指文本中第 i 个关键词在第 j 个领域本体中的最相似概念。该矩阵中每一列对应一个领域本体，结合前面的关键词向量模型，取一个加权值α $(0<\alpha<1)$，得出文本与领域本体的总体相似度为

$$S_j = \sum_{0<i<n,0<j<m} (w_i + \alpha k_i d_j) \quad (0<\alpha<1)$$

(3) S_j为文本与第 j 个领域本体的总体相似度，取前三个最大的值，得出与文本最为接近的三个领域本体。根据矩阵列矩阵c_{ij} $(0<i<n)$，即文本关键词列表在各个领域本体中对应的最相似概念矩阵，通过一定的推理，找出在该领域内部其它隐含的概念，这样可以变相对文本进行挖掘，找出文本中隐含的意思，从而提高索引的准确性和全面性。

通过上述的 3 个步骤，建立了网页到本体的映射，更确切的说是到本体意群的映射，这里的意群是指本体中的一个语义范围，它有大有小，根据领域概念大致的分布来进行定义，这里有两种方法来对这些意群进行定义。

（1）两个意群间的共有概念达80%以上可认为是同一意群。

（2）以一个意群为中心，以一定语义距离为半径，在这个圆内的意群为相似意群，圆环内的意群为相似意群，相似度随着半径的增加而减少。

可以设想为每个意群设定一个ID，随着意群的增加，意群的限定会越来越困难，这时需要对意群进行合并和重构，或建立意群间的关系，这实际上是对本体内的一种语义细分，只是我的一种想法，但能在多大程度上改进本体的功能还不能确定，希望能给这方面的研究提供一点贡献。有的意群可能横跨几个小意群，这时就可以进行整合。

通过对意群的整合，可以控制意群的数量，减少不必要的意群划分，提高处理的效率。还有个好处就是便于领域本体间意群的映射，原来是概念或属性间的映射，代表的语义信息较少，而建立意群间的映射，可以更大程度上反映本体间语义上的联系，但是这样建立映射的复杂度比较高，需要充分考虑内部的语义关系，进行精确的定义。

4.3 博客页面排序算法研究

现在知名的页面排序技术来自Google的PageRank算法，它是Google排名运算法则（排名公式）的一部分，是Google用于用来标识网页的等级/重要性的一种方法，是Google用来衡量一个网站的好坏的唯一标准。在揉合了诸如Title标识和Keywords标识等所有其它因素之后，Google通过PageRank来调整结果，使那些更具“等级/重要性”的网页在搜索结果中排名获得提升，从而提高搜索结果的相关性和质量。Google的PageRank根据网站的外部链接和内部链接的数量和质量俩衡量网站的价值。PageRank背后的概念是，每个到页面的链接都是对该页面的一次投票，被链接的越多，就意味着被其他网站投票越多。这个就是所谓的“链接流行度”——衡量多少人愿意将他们的网站和你的网站挂钩。PageRank这个概念引自学术中一篇论文的被引述的频度——即被别人引述的次数越多，一般判断这篇论文的权威性就越高，如果一篇高质量的论文引用了你的文章也说明你的论文具有很高的影响力。

它很重要的一个核心思想是将互联网上的页面都看成一个整体，而不像其他的搜索引擎将一个网页单独处理，从而忽视了页面之间的联系。它的算法是没有人工干预的，全靠算法来控制，大多数人认为一个网页重要的网页就会有很高的PR值，这体现了一种民主的思想，这样可以防止那些利用关键词进行

恶意拼凑来获得排名的行为。根据上面的集成语义的综合博客主模型，可以设定一定的标准来对它进行定量化，便于分析一个博客的整体价值，在这里，每个模块的权重是不一样的。

在模型中，有些共同行为，可以用计分的方式来对此进行量化，比如：

访问	留言	有留言	分享	转载	被转载	发表日志
1	2	1	2	2	8	4

表 4-1　博客页面重要属性权重表

这些积分只是作为一种参考，具体在实施过程中需要考虑很多细节问题，设定一些规范。对于不同的因数，应该区别对待。而且每个模型内部也有区别，比如好友间有访问很频繁的，也有访问很少的，这就需要界定一个标准，设定等级，经常访问的好友在评价中占的权重更大，这个原理也可以应用于其他模块。所在博客圈的人气和你在里面的活跃程度共同决定了该博客圈对你的影响程度。

设好友圈、博客圈和好友价值模型的得分数分别为Q_f，Q_c，Q_o，设q_f，q_c，q_o分别为仅根据表 5-1 算出的一个博客在好友圈、博客圈和自身价值模型中的得分，没有加入模型内部权重的设定，下面对每个模型内部进行权重分析。

好友圈：最近一个月的互访次数，一般取互访问次数的均值，如果访问主要集中在己方，这样的得分是很少的，而对方的得分会很多。设该博客对 n 个好友的访问次数分别为：$VA_i(1<i<n)$，n 个好友对他的访问次数分别为：$V_i(1<i<n)$。具体步骤如下：

1）首先根据上面的思想，设计它的每个好友的权重公式，$W_{fi}=2^{\frac{V_i-VA_i}{V_i}}/n(1<i<n)$，当$V_i>VA_i$时，$V_i$越大，权重越大，反之$V_i<VA_i$时，$VA_i$越大，权重越小，这里除以 n 是考虑每个博客的权重有限，好友越多，平均每个好友获得的权重就会受到影响，这点有待更好的算法来解决。

2）得到权重后，W_{fi}乘以在好友中的得分q_{fi}，得到在每个好友中的最终得分：$Q_{fi}=W_{fi}\times q_{fi}(1<i<n)$

3）最后得出在好友圈中总得分为：$Q_f=\sum_{1<i<n}Q_{fi}$

博客圈：博客圈排名和自己在圈内的活跃程度。很多博客圈以发表文章数量、留言评价、访问次数等来对里面的每个注册用户进行积分，但这里不能采用这个积分，原因有两个，一是里面的积分标准与我设定的标准不一样，而是博客圈间的标准也不一样，由于上述两个原因，我决定采用表 4-1 定义的标准。为了简化处理，只是抽取一些关键的元数据，比如在发表文章数、文章评论次数、文章显示在首页的次数等。还有一个问题，就是一个用户可能同时在几个博客圈中，每个博客圈的权重也要进行划分，这个参考该博客圈在整个博客圈中的排名、平均一周更新数量、精华文章的数量等因数来衡量。

大致流程如下：

1） 首先按照模型得到该用户在每个博客圈中的得分 $q_{ci}(1<i<n)$。

2） 根据博客圈权重计量方法，综合排名、更新数量和精华文章来计算博客圈的权重 W_{ci}。

3） 根据上述两个数据，可以得出该用户在第 i 个博客圈中的得分为 $Q_{ci}=W_{ci}\times q_{ci}(1<i<n)$，进而得出该用户在博客圈中的得分 $Q_c=\sum_{1<i<n}Q_{ci}$。

自身价值模型：主要分析该用户最近的活动量，比如发表日志及留言情况、上传图片、转载、链接、分享等，可以按照表 5-1 来进行计算。这里就不做详细介绍了。

通过这些模型以及设定的计分方法，可以得出每个模型中的分数，还需要设定模块整体权重，好友圈为 α，为 0.4，博客圈为 β，为 0.1，自身价值模型为 γ，为 0.5。这是一个整体的权重，没有科学的依据，只是为了这里的分析的需要，考虑到有些用户没有加入博客圈，加上博客圈中的评分涉及很多不确定因数，比如博客圈间标准的不一样等，所以将其权重设为 0.1 来平衡差别，对于没有博客圈的用户，好友圈自动变为 0.5，与自身模型一样。

这样可以得出一个博客的总体得分数：$Q=\alpha Q_f+\beta Q_c+\gamma Q_o$，以此作为博客排序的重要指标之一。

链接应单独做一个因数进行分析，防止那些以链接来获得积分的博客。我借鉴了 Google 中的思想，每个博客都被分配了一个链接权重，当它链接的博客越多时，平均分给每个博客的分数就越少，当它出链的数量越少，每个博客获

得的分数越多。然后将所有这些链接分数加起来就组成了链接这一块的总分数。对于双向友情链接过多或单向链出数过多的页面进行权重降低的限制。现在很多网站摸清了搜索引擎的运行机制，采取恶意的行为来提高网站排名，比如拼凑关键词、黑链接、论坛签名、购买友情链接等，但是忽略了最重要的，就是网站的内容，原创的内容，有价值的内容，这才是关键。Google 已做出一些调整，在某种程度上衡量一个网站的整体效用来进行排名更加合理化。

看了 Google 创始人 LarryPage 的 PageRank 算法思想，我觉得这个算法在博客搜索中可以借鉴。可以将所有的博客用户页面看成一个整体，用系统的方法来对页面信息进行抽取，按照统一的标准进行处理，博客页面间信息是相互关联的，主要有链接和好友两种关系。得出这些信息后，只需要确定哪些是该博客的好友或有到这个博客的链接，就可以很容易地将这些数据调出来加以应用。

这里的排序只是作为一种重要的参考指标，最重要的还是内容的相关度，在内容相关度上接近的博客页面就采用这种排序方式进行修正，以达到挖掘潜在草根博客的价值。

第 5 章 本体在博客营销中的应用

5.1 博客营销

博客营销是将博客作为平台和渠道进行企业广告宣传和促销营销模式。作为企业自媒介的博客特点在于“博客值得信任，企业博客与传统的公关以及广告行销手段相比，具有营销成本低、时效性强、参与性与互动性突出、读者信任度高以及营销效果易见等特点和优势。目前，企业博客已经被应用于沟通消费者、处理媒介关系以及产品事件行销、树立行业先锋形象和企业文化宣传等诸多领域。”

随着博客的不断发展，聚集了大量的人气，博客正在实现了网络与现实的接轨，企业从这里看到了巨大的商机，纷纷涉足了博客营销这一领域，以求来分一杯羹。博客营销颠覆了许多旧的营销理念，传统的“4P”理论已经远远跟不上时代的发展。科特勒先生对他本人提出的、可以说是现代市场营销理论基石的“4P”模型进行了修正，提出了最新的市场营销模型：CCDVTP。所谓CCDVTP是指：创新（Create）、沟通（Communicate）、价值传递（Deliver Value）、目标市场（Target）和获利（Profit），CCDVTP模型就是：针对目标市场，通过创新、沟通和价值传递，实现赢利。博客营销完全符合这一理念，通过创新一种全新的沟通模式来培育目标市场，更重要的是一种价值传递。它是一种非官方的沟通方式，是一种自底向上、互动、平等、自由、开放的方式，这更容易让人接受，而不是一种“推”的方式。

价值传递就是一个该如何与客户沟通的问题。品牌的价值主张，也不是简单地追求说服客户，而是要引起客户与你的心灵共鸣，认可你的价值观，这样才会成为你的忠诚用户。

博客营销的精髓在于给予和分享，将自己的心得体会与知识通过一种巧妙的方式传递给大众，让大众在轻松之余产生深刻的思考，被你的这种无私分享的精神所打动，这样才达到了博客营销的目的。博客营销主要有个人博客营销、企业博客营销和博客圈营销，这里主要讨论企业博客营销，通过博客来提升公司的名气，介绍公司的产品，丰富和用户的互动方式。博客营销的优势有如下几点：

1）细分程度高，定向准确。企业可以利用博客进行一些在线调查，得到一些用户的评论信息，这些信息对于企业非常宝贵，在交流的过程中可以发现问题并仔细考虑用户的需求，根据这些信息进行产品的细分。

2）互动传播性强，信任程度高，口碑效应好。通过不断的沟通，可以增进双方的互信程度，开发性不断得到提高，给用户一个更加全面的认识。

3）影响力大，引导网络舆论潮流。通过不断地“博客”来聚集人气，提升行业知名度，引领网络舆论的前沿，扩大影响力，传递一种价值理念。

4）与搜索引擎营销无缝对接，整合效果好。现在出现了许多博客搜索引擎，更加看重博客的整理质量，通过博客搜索引擎对博客资源进行整合，扩大博客营销的传播渠道。

5）有利于长远利益和培育忠实用户。博客营销是一个思想交流的过程，是一个潜移默化的过程，应该从长远的角度来认识它的价值。一旦一个用户认可了你的价值，就会长期信任你，经常关注你的动态，博客营销用利于公司的长远利益和培育忠实用户。

5.2 基于本体的博客营销模型的设计

营销的方式有很多，从体验式营销、一对一营销、关系营销、连锁、品牌营销、深度营销，到网络营销、整合营销、直销、数据库营销、文化营销等，不仅在众多营销案例中能见到它们的身影，而且其阵营还在扩张。其中的网络营销，从最初的品牌网络广告、搜索引擎营销、窄告、插件类网络推广产品、通过第三方平台发布商业信息，发展到视频网络广告、社区营销、博客营销、电子邮件营销、即时通讯营销等多种形式，家族规模庞大，而且实力不凡，品牌网络广告成了新浪、搜狐等门户的主要来源，搜索引擎营销助百度成功登陆纳斯达克。博客营销作为最近几年才兴起的一种新型营销模式，已经吸引了广泛的关注，相信在未来几年，它将改变传统营销模式，引发营销理念的变革。

在博客中已运用得很成熟的 RSS 技术正发挥着越来越重要的作用，现在很多大型的网站都支持 RSS2.0 技术，很大原因是它简单实用，使得这项技术的影响力与日剧增。但是它有个缺点，不能处理结构化的数据，没有一套用于扩展的标准。RSS1.0 是采用 RDF 的格式来设计的，扩展性和兼容性要些，被 Web 专家和技术标准组织所认可。

RSS 是一种共享资源模式的延伸，而本体是一个领域的知识表示集合，这二者的结合会产生什么的效果呢？RSS 在语义的表达上很欠缺，现在网上的 RSS 资源很多，它只负责共享资源，里面定义的是内容的相关信息，但对资源内容的意义和价值没有一套评价机制，这给它在企业博客营销方面带来了不便。而本体是领域的知识集合，是概念模型的规范化的说明，可以利用本体的语义特性来对企业 RSS 资源进行整合，通过搜索引擎对 RSS 资源进行分析，找出相关的 RSS 资源，建立企业 RSS 资源到本体的映射，这样就能在博客搜索的同时，推出企业的相关服务。具体的步骤如下：

1） 企业将自己的 RSS 资源提交给搜索引擎，搜索引擎对里面的内容进行分析，建立 RSS 资源到本体的索引，企业的更新都会被搜索引擎进行跟踪分析。

2） 企业有自己的服务功能模块，这些模块也可以进行营销，用户搜索的内容与该公司 RSS 资源有相关性时，当用户访问排名靠前的博客文章时，该功能模块就会自动显示在博客日志的下面，比如，用户搜索电影相关的信息时，搜索引擎会将一个功能模块显示在日志下面，这个模块有如下功能：它可以显示最近影片的预告片，用户可在线观看，还有该影片在各大影院的放映信息，放映时间、影片介绍、打折信息、离用户最近的影院等，这就是一套完整的服务，搜索引擎来整合这些资源。

由于前面已经设计一个基于语义网的博客搜索模型，这里就简化了博客信息的详细处理。

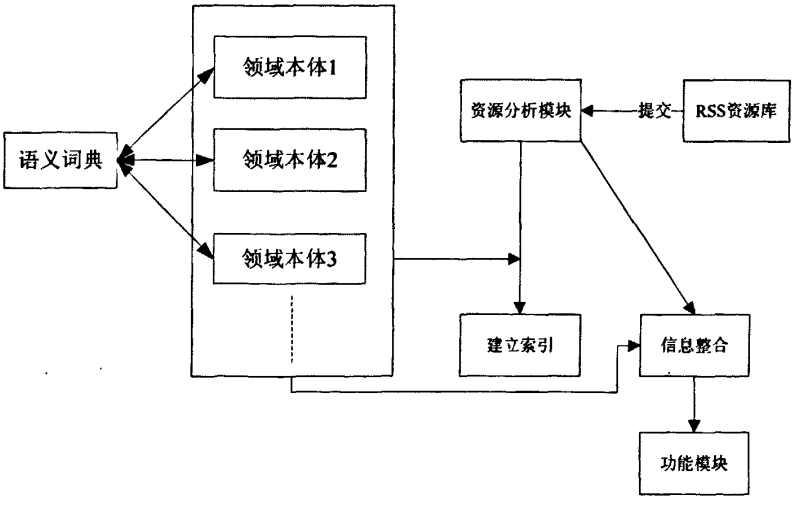


图 5-1 基于本体的博客营销模型

基于本体的博客营销有三种模式可以借鉴：

（1）企业博客内部做优化，提高企业博客的内容质量和整体功能性，来提高其在搜索引擎中的排序。

（2）通过三方工具集成相关企业的资源，以一种功能模块的形式来推广企业的相关产品或服务，在相关博客搜索中排名靠前或访问量大的博客中进行推广。

（3）在点击率高的博客中投放企业产品广告。

第一种形式是企业真正参与到博客的建设过程中来，通过幽默、轻松、专业、故事化的博客文章来吸引用户，与用户进行互动，进行详细的市场调查和分析来确定公司定位，并在潜移默化中传递公司的核心价值观，推广公司的营销理念。这是一个长期的过程，需要公司做一个长期的投入，但这个投入比起建网站的成本要低得多。博客是一个个性化交流平台，集成了多种功能，更加便于沟通，这里涉及一个与用户真诚沟通的问题，没有这种对用户的长期的培育，是不能产生长远效果的。这种形式是以提高企业博客在搜索引擎中的排名为出发点的，前面设计的基于语义网的博客搜索模型，对这种企业博客也是适用的，里面的排序模块会不一样，对于企业博客，它有友情链接，行业企业博客群，自身价值模块，这在排序算法上会有区别。但在整体设计上区别不大，可以参考前面的模型进行分析。

这里主要讨论第二种形式的营销方式，它是本体、RSS 技术和搜索技术的结合，营销模式的选择也很重要。这种模式需要第三方设计一个插件模块，还需要这个行业企业的支持才行，运用 RSS 技术收集行业信息进行整合，运用本体来规范行业内的语义表达歧义，最终以一种功能模块的形式显示在博客用户的日志中。选择在哪些博客用户中显示也是要考虑的问题，传统方法是显示在点击量高的博客用户中，运用集成综合博客主模型来对博客页面进行排序，取相关度排名靠前的博客进行投放是一个不错的选择，当然这里的算法还需要优化。

5.3 博客营销模型在旅游行业中的应用分析

旅游行业是一个发展比较快的行业，各种资源需要得到整合，一站式的服务是一个大趋势。比如携程网，提供一条龙的旅游服务，但用户只有登陆携程网才能了解到相关信息，它可以将一部分功能定制成模块，对外开发，在相关

旅游知名博客中进行推广，用户在浏览博客的同时能了解到旅游线路的相关信息，从而扩宽了携程网的推广渠道。

博客营销在旅游行业的潜力很大，因为很多人亲身去体验过，对当地风景的体会更深，他们生动的描述对希望去旅游的人有很大吸引力。还有一些驴友，经常在外面行走，到过很多好玩的地方，其中有些地方是旅游线路上没有的，这对户外爱好者也有很有吸引力。

旅游企业可以选择在相关博客中进行推广服务，比如一个以描写杭州的人文景观、风土人情为主的用户，在杭州风景相关博客搜索中的排名很高，可以考虑在里面进行旅游服务的推广。通过一个第三方集成工具，集成各大旅游公司的相关最新旅游线路、航空机票打折情况、目的地酒店的预定情况等。它能根据用户所在方位，推荐出最合适的旅游线路和附近旅行社的相关信息，从而给浏览博客的用户很大的选择性。这里用到了本体的技术，它能定制一个流程标准，解决各大旅行社间的信息表达不一致的问题，为它们之间更大的信息集成提供支撑。各大旅行社的最近旅游线路信息也能及时地显示在功能模块中，这里用到了 RSS 技术，它能自动检测到最新的数据更新，给用户提供最新、最及时的信息。

这在未来会是一个很大的趋势，这些服务都是可以定制的。目前 RSS 只是实现了内容的实时传输，用户可以在自己的网站看到其他网站的最新信息，而在自己的网站里面能集成其他网站的服务则是一个以前难以想象的事情，但确实有人在这方面进行研究，也取得了一些成果，但还没有进行市场化的运作。中科院的一个项目产品 OncePortal 是一个非常不错的产品，它能让你在自己的网站里用网易邮箱登陆邮件，登录搜狐论坛，订机票等，在自己的空间里集成其他网站的功能，你可以采用拖拽的方式将其他网站的功能放进你的空间，它实现了网站间的无缝对接，确实很强大，当然这里还需要对方网站对这块没有设限，有的网站屏蔽了这方面的扩展功能。

当然博客营销要想发挥它最大的优势，也不能脱离其它营销方式，传统的营销方式也有它的优点。二者进行很好的结合，能产生更好的营销效果。我对博客营销的前景很看好，相信博客营销借助本体的优势将发挥越来越重要的作用。

第 6 章 总结与展望

6.1 全文总结

本文对语义网技术和本体相关知识进行了介绍，通过分析现有博客搜索引擎存在的问题，将语义网技术应用于博客搜索，构建了基于语义网的博客搜索系统模型，并对关键子模块进行了详细分析，对模型中涉及的相关算法进行了研究，为基于语义网的搜索引擎的实现做了大量理论研究工作。所做的工作有以下几点：

（1）介绍语义网和搜索引擎相关知识，对搜索引擎的相关原理和模块进行了分析，对语义网中的关键构成元素本体相关知识进行了介绍，它们为基于语义网的博客搜索提供了技术和理论基础。

（2）对博客及博客搜索现状和相关研究进行了分析，总结了它们优缺点，提出了基于语义网的博客搜索模型。

（3）对基于语义网的博客搜索各个子模型进行了分析和构建，重点在索引模块和排序模块，并对模块中的相关工作流程进行了研究。

（4）对关键技术与算法进行了分析，包括领域本体的构建、语义词典的结构分析、基于本体意群的索引算法和基于集成语义的综合博客主模型的排序算法。

（5）对博客模型在博客营销中的应用进行了分析，并对博客营销的前景进行了预测。

本文的创新点有如下几点

1）提出了一个集成语义的综合博客主模型，并对模型中的各个子模块进行了研究，设计了一套基于该模型的排序算法。针对博客的个性化特征和丰富的内涵，还有博客之间的紧密联系，该模型能在整体上衡量一个博客的价值，能更深层次的挖掘博客的潜在价值。

2）提出了一种基于本体意群的索引模型，并对索引算法进行了研究。传统的倒排序索引建立方式没有涉及语义层次，只是一些技术性的改变，虽然在一定程度上提高了索引的质量，但对于处理呈爆炸性增长的网络信息流量已越来越力不从心，迫切需要能在语义层次上对资源进行组织，提高资源的处理效率。

6.2 研究展望

本文对博客搜索系统模块进了分析和设计，对相关算法进行了研究，文中的不足有以下几点：

（1）由于时间和精力有限，没有完成原型系统的开发工作，不能很好地检验这种模型的效果，这是一个遗憾。

（2）排序算法是凭自己的想法进行构思的，只是大致上进行了量化说明，但没有经过系统科学的检验，缺乏一些客观性。

（3）没有考虑到搜索的现实情况，一般大型搜索都是基于分布式的搜索，海量数据的存储，这也是系统设计上一个缺陷。

本文有许多理论还存在很多漏洞，没有更深入地考虑其中的很多搜索细节，实现分布式和语义网技术在博客搜索中的运用将是自己下一步的重点工作，还需要不断提高自己的实际动手能力和思维多样性能力。

我相信博客搜索会成为了今后十年网络发展的一个很大的趋势，它加快了虚拟网络到现实的转化，提升了大众的话语权，这种独特的沟通方式已经吸引了众多的网民，下一步的发展趋势是搜索引擎转化为博客的搜索，朝着个性化的方向发展，全文搜索引擎将被专业化更高的垂直搜索引擎所取代。

参考文献

[1] Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead, Matthew Stiff: Definition of the CIDOC Conceptual Reference Model[R]. ICOM/CIDOC CRM Special Interest Group, Version 3.4.9, 30th November 2003.

[2] Technorati. http://www.technorati.com

[3] BlogStreet. http://www.blogstreet.com

[4] DayPop. http://www.daypop.com

[5] BlogPulse. http://www.blogpulse.com

[6] Sphere. http://www.sphere.com

[7] 黄波，主题搜索引擎的研究与应用[D]，成都理工大学，2007.11

[8] 肖亮.垂直搜索引擎的研究与实现[D]，北京交通大学，2008.3

[9] Sugiura,Atsushi;Etzioni,Oren:Query routing for Web search engines: architecture and experiments[J],Computer Networks, Volume:33, Issue:1-6,June,pp:417-429

[10] Query routing for Web search engines: architecture and experiments.Atsushi Sugiura Oren Etzioni. Computer Networks 33 (2000) 417-429.

[11] XDSearch: an efficient search engine for XML document schemata , Eric Jui-Lin Lu.Yu-Ming Jung Department of Information Management,Chaoyang University of Teclmology, Expert Systems with Applications 24 (2003) 213-224.

[12] Evolutionary document management andretrieval for specialized domains on the web.Mihye Kim Paul Comptom. Human-Computer Studies 60 (2004) 201-241.

[13] Rccipon, Hcrvc; Makalowski, Wojcicch: The biologist and the World Wide Web: an overview of the search engines technology, current status and future perspectives[J].Current Opinion in Biotechnology, Volumn:8 Tssuev:1,1997,pp 115-118

[14] Gruber TR.A translation approach to portable ontology specifications. Knowledge SystemLaboratory, 1993.Technical Report, KSL 92-97.

[15] Deng ZH, Tang SW, Zhang M, Yang DQ, Chen J. Overview of ontology.Acta Scientiarum Naturalium Universitatis Pekinensis, 2002,38 (5):730-738.

[16] W. N. Borst and J. M. Akkermans. Engineering Ontologies. International Journal of Human-Computer Studies, 2002,46(2):365-406

[17] John F. Sowa. Top-level ontological categories. International Journal of Human and Computer Studies, 1995, 43(5):669-685

[18] S. Decker, F. van Hannelen, J. Broekstra et al. The Semantic Web--on the roles of XML and RDF. IEEE Internet Computing, 2000, 45(7):985-999

[19] M. Uschold and M. Gruninger:Ontologies, Principles, methods, and applications. Knowledge Review, 1996, 11(2):93-155

[20] T. R.Gruber, A translation approach to portable Ontology specifications. Knowledge Acquisition, 2002,5(2):199-221

[21] B. J.Wielinga, J.Sandberg, and G.Schreiber. Methods and techniques for knowledge management:What has Knowledge Engineering to Offer. Expert Systems with Applications,2003,13(1):73-84

[22]W3C, 'Web Ontology Language (OWL): Overview W3C Working Draft 10 February 2003'. http://www.w3.org/TR/owl-features/

[23] Uschold M. Ontologies:Principles, Methods and Applications. The Knowledge Engineering Review , 93-115,1996 11(2)

[24]Jorg-Uwe Kietz, Raphael Volz, Alexander Maedche,Extracting a Domain-Specific Ontology from a Corporate Intranet , Proceedings of the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop , Lisbon , 2000

[25] Gruninger, M . and Fox , M.S. Methodology for the Design and Evaluation of Ontologies ,Workshop on Basic Ontological Issues in Knowledge Sharing , I JCAI-95, Montreal ,1995.

[26] IDEF 网址. http://www.idef.com

[27] 陈禹主编. 建模分析与设计方法. 北京清华大学出版社，1999.

[28] Chang C K, Garcia-Molina H . Conjunctive Constraint Mapping for Data Translation. In : Third ACM Conference on Digital Libraries,Pittsburgh, Pa., June 1998.49 -58

[29] Chawathe S, Garcia-Molina H, Hammer J, Ireland K, Papakonstantinou Y,Ullman J, Widom J.The TSIMMIS Project: Integration of Heterogeneous Information Sources. In :16th Meeting of the Information Processing Society of Japan(I PSP94), Tokyo, Japan, 1994.7-18.

[30] Weinstein P C, Birmingham P. Creating Ontological Metadata for Digital Library Content and Services. International Journal on Digital Libraries,1998,2 (1):19 -36.

[31] Mena E, Illarramendi A, Kashyap V, Sheth A P. OBSERVER: An Approach for Query Processing in Global Information Systems Based on Interoperation Across Preexisting Ontologies. In: Proceedings of the First IFCIS International Conference on Cooperative Information Systems(CoopIS'96). Brussels, Belgium, 1996.19 -21

[32] Mitra P., Wiederhold G., Kersten M. A Graph-Oriented Model for articulation of Ontology Interdependencies. In:Conference on Extending Database Technology, Konstanze Germany, Mar. 2000.

[33]A. Rodriguez, M. Egenhofer. Determining Semantic Similarity Among Entity Classes from Different Ontologies. IEEE Transactions on Knowledge and Data Engineering, 2003,15(2): 442~456.

[34]Alexander Maedche, Steffen Staab. Measuring Similarity between Ontologies. In: Proceedings of the European Conference on Knowledge Acquisition and Management 2002. Madrid, Spain, October 1-4, 2002. LNSC/LNAI 2473, Springer, 2002: 251~263.

[35]Satoshi Sekine, Kiyoshi Sudo, Takano Ogino. Statistical Matching of Two Ontologies. In Proceedings of the SIGLEX99: Standardizing Lexical Resources, Maryland, USA, 1999: 69-73.

[36]各类人士对博客的不同理解和定义.http://www.blogger.com

[37]宋双峰. 什么是"博客". 《中国记者》 2004 年第 10 期第 81 页

[38]徐晓波. 论博客现象及其技术基础和传播特性. 来自中国学术论坛网 http://www.frchina.net/dataldctail.php?id=13315

[39] Makoto Nakatsuji, Yu Miyoshi, and Yoshihiro Otsuka. Innovation Detection Based on User-Interest Ontology of Blog Community.2006.9.

[40] Knud M¨oller, Uldis Boj¯ars, and John G. Breslin. Using Semantics to Enhance the Blogging Experience. Digital Enterprise Research Institute, National University of Ireland.2006.4.

[41] Yoonjae Jeong, Dongman Lee. A Rapid Egocentric Search Scheme Using Authority Estimation in Blog Space. Digital Media Laboratory, Information and Communications University,2008.12.

[42] Yun Chen, Flora S. Tsai, Kap Luk. Machine learning techniques for business blog search and mining. School of Electrical and Electronic Engineering, Nanyang Technological University .Expert Systems with Applications 35 ,(2008):581–590.

[43] Sachit Rajbhandari, Frederic Andres, Motomu Naito. Semantic-Augmented Support in Spatial-Temporal Multimedia Blog Management. Lecture Notes in Computer Science. 2007.9: 4438 (215-226)

[44] RSS 2.0 Specification.http://blogs.law.harvard.edultech/rss(Accessed 2007-4-10)

[45]The Great RSS vs. Atom News Feed Debate. http://www.lawtechguru.com/archives/2004/02/13_the_great_rss_vs_atom_news_feed_debate.html(Accessed 2007-4-10)

攻读学位期间发表的学术论文及参与的科研项目

[1] 聂规划, 章志龙, 王锐．基于语义词典的电子商务推荐系统模型研究.情报杂志[J], 已收录，将于 2009 年 12 月份发表

参与项目

[1] 国家科技支撑计划“电子商务与现代物流共性集成技术研究开发”。项目来源：国家科技支撑计划，编号：2006BAH02A08。

[2] 湖北省科技攻关计划“电子商务与物流语义集成技术研究”。项目来源：湖北省科技厅，编号：2007AA402A48。

[3] 省自科“基于语义网的网络消费心理知识挖掘及其本体构建研究”。项目来源：湖北省科技厅，编号：2007ABA190。

[4] 省自然科学基金“网络环境下基于本体的电子商务推荐系统研究”。项目来源：湖北省自然科学基金，编号：2006ABA303。

致 谢

随着论文工作的结束，我也即将结束我的硕士研究生学习生活。回首往事，感触颇多，这一路走来，师长、朋友、家人的帮助让我获益匪浅。没有他们的帮助，就没有我今天的成绩。

首先，我由衷地感谢我的导师聂规划老师。在攻读硕士学位期间，聂老师渊博的学识、和蔼可亲的态度、严密的思维逻辑以及追求卓越的工作精神一直是我心目中的学习典范，他不断给我指明前进的方向，让我在今后的人生道路上少走弯路。两年半的时间里，无论在学习还是生活上，聂老师都给了我很多关心和帮助。在此，谨向聂老师致以崇高的敬意和真诚的感谢！

另外感谢研究所的陈冬林，刘平峰，王惠敏，申学武，杨爱民，傅魁，刘勇军，曹洪江等老师，你们对我的帮助和指导，让我获得了实际的锻炼，也加深了我对自身的认识，要不断去克服自我，向更高的目标奋斗。还要感谢我的师兄师姐们，付志超，龚璇，丁峰，杨敏，左秀然，感谢他们在研究生生涯中给予我的帮助和关心，他们给了我许多宝贵的意见，对我的论文写作提供了很大的帮助。还要感谢王锐，付梦，夏欢，李丽颖，梁越岭，蒋祥杰，李晓菲，徐尚英等，感谢他们在工作上对我的帮助及合作。

在研究生学习期间，我还要感谢跟我一起学习、一起生活的所有经济研2007级同学，和他们一起度过了一段美好的时光，将成为我人生中弥足珍贵的回忆，使我在前进的道路上不是一个人在奋斗。

最后，感谢我的父母，在成长的道路上他们给了最无私的爱，用他们的勤劳的汗水来教会我许多做人的道理，感谢他们多年来的养育之恩，到了儿子回报你们的时候了。

章 志 龙

2009年11月5日