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。据我所知，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，也不包含为获得河北农业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：赵全来　　签字日期：2011年6月7日

关于论文使用授权的说明

本学位论文作者完全了解河北农业大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权河北农业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，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：赵全来　　导师签名：

签字日期：2011年6月7日　　签字日期：2011年6月7日

学位论文作者毕业后去向：

工作单位：　　电话：

通讯地址：　　邮编：

摘 要

随着 Internet 的发展，传统搜索引擎查准率低下的缺点不断的暴露出来，个性化服务便应运而生。用户偏好作为个性化系统的核心，逐渐受到了重视。本文对用户偏好的建模以及更新方法等进行了深入的学习研究，构建了基于用户偏好的农业智能问答系统的原型系统。

该原型系统将用户的访问页面和 Web 日志作为挖掘分析对象，并不需要客户端用户过多的参与，可以由系统自动的分析推断出用户的兴趣来构建偏好模型并更新之。另外系统仿照人类遗忘规律对不同种类的偏好进行不同程度的遗忘。通过不断的更新用户模型，使之越来越贴近用户的真正意图。

本文着重从以下几个方面进行了学习和研究：

（1）全面研究了信息检索系统的关键技术，重点研究了构建基于本体的用户偏好模型的关键技术和核心问题。

（2）对用户偏好模型进行了重点研究，探讨了用户偏好模型的表示方法，建模方法，更新方法等关键技术，并提出了一种新的用户模型更新算法。

（3）在充分学习和研究了传统的农业相关网站的基础上，构建了农业智能问答系统的原型系统。

（4）认真设计了测试数据，对文中的算法和用户偏好模型进行试验和分析，验证了其有效性。

关键词：用户偏好；智能问答；农业本体；Ontology

Research on the Agricultural Intelligent Question Answering System Based on Ontology

Author: ZhaoQuandong
Major: Computer Applied Technology
Supervisor: Wang Fang

Abstract

With the development of internet,the shortcomings of lower precision of traditional search engines have been revealed.Individuation service emerges as the times require.The user profile,as the core of personalized system, is growing recognition. This paper study and research on the modeling and updating method of user profile,and bulids the Agricultural Intelligent Question Answering System.

The system excavates and analyzes the the accessed pages and Web logs,doesn't ask the client users to participate in it too much. It can automatically analyzes and deduces user profile , subsequently builds and update the model of profile .Besides this system imitate human oblivious rules to forgets the different kinds of interests in varying degrees.Through regular updating the model, making it more and more to get close the user'sreal intentions.

This article mainly studies and researches on from the following aspects:

(1) The key technology of information retrieval system is comprehensively studied in this paper, the core problem and the key techniques how to construct Ontology-Based Preference Model is mainly studied

(2) The user profile model is mainly studied in this paper,and the representation ,modeling method,updating method and so on is discussed,finally a new algorithm for updating the model is put forward.

(3) Based on studying and research on the traditional agricultural websites, the Agricultural Intelligent Question Answering System is build。

(4) Test data is designed in earnest the algorithm and the model is ,experimented and analysised, its effectiveness is verified.

Keywords: User profile;Intelligent question answering;Gricultural ontology;Ontology

目　录

1 引言

1.1 研究的背景和意义

互联网的发展为人们带来了巨大的惊喜与变革，也为人类带来了许多难题和挑战。随着 Internet 网络在全球的迅速蔓延，计算机软、硬件技术的飞速发展以及人们利用信息技术传播资讯的能力大幅度提高，万维网自从 1991 年出现以来已经发展成为一个巨大的全球化的信息空间。“新摩尔定律”指出：Internet 上的信息正以每六个月翻一番的速度爆炸性地增长着。信息变得极大丰富在为互联网用户带来极大的便利和乐趣的同时，也使他们陷入了“信息过载”的困境当中。在现实生活中，人们发现要找到自己所需的信息正变的越来越困难。功能强大的搜索引擎的广泛使用可以缓解这样的困境。但搜索引擎由用户手动触发，用户通常一次只输入很少的关键字进行查询，搜索的结果往往数据量巨大而命中率却很低，需要用户再花费大量的时间和精力进行二次查找，无法满足用户准确、快速的需求。可见，目前的信息检索技术还具有很大的局限性，具体表现在[1,2]：

（1） 查询返回集繁杂，导致查准率低下

互联网上的信息巨大，增加速度迅猛，而传统的检索只是根据用户输入的简单的关键词返回相关的结果，这样繁杂的结果集导致查准率低下，用户很难迅速准确的得到所需信息。

（2） 不能提供个性化的服务

由于用户的年龄，学历，职业，性别，地域的不同，对搜索的风格和内容也要求不同。传统的搜索只针对检索词，没有考虑用户的真正需求，导致了较低的查准率。

（3） 互联网信息的随机变化性

网络信息的的发布具有随意性和自由性，任何人和机构都既是信息的发布者又可能是信息的检索者。每天都有海量的信息发布，同时也会有大量的信息失效或者更改，这都使网络信息表现出随机性。

（4） 互联网信息的无序性

不同的网络通过 TCP/IP 协议连接在一起，但对信息资源并没有统一的组织规范。虽然对于每个网站或者网页来说，信息具有一定的组织规则，但通观整个信息网络，资源却处于无序和分散状态。

面临以上种种困惑，对于一个网站来说，如何从中脱颖而出得到用户的认可，只有提高自己的服务质量，提供更符合用户偏好，充分体现用户个性化要求的信息服务。为用户提供个性化服务，一般有三种方法：基于内容的方法，基于规则的方法，协作过滤的方法。

基于内容的方法：该方法主要是根据网页的内容与得到的用户档案的相似度确定个性化服务的内容。

基于规则的方法：该方法是指网络管理者以统计学的方法建立规则来确定个性化服务的内容。

协作过滤的方法：该方法是指在直接或者间接的取得信息的基础上将客户分为不同的类型，然后通过同类客户群的评级确定发送给客户那些服务内容。

这些方法使得网络服务提供商能够对用户提供个性化的服务形式和服务内容。然后，基于规则的方法在规则数量较多时经常很难管理，而协作过滤的方法的评价数据又有很大的片面性。为此越来越多的研究焦点聚集在基于内容的方法，也就是我们所说的依据用户偏好，为用户定制个性化服务。基于内容的方法，如果直接获得用户档案，其实就是一种静态方法，它不能及时体现用户偏好的变化。间接的用户档案管理通常是指通过自动分类技术和聚类技术得到的，即我们所说的根据用户访问的数据内容挖掘用户偏好的过程，它是一种动态方法，可以及时的跟踪用户的兴趣变化。用户访问过的数据，集中反映了用户的个性偏好，为此我们就可以通过对 web 日志文件和相关数据进行分析挖掘，进而得到用户的个性偏好，据此为用户提供个性化服务，例如针对性的信息推送等等。

近年来，随着网络技术的发展，农业服务和信息的查询也越来越多的依靠网络服务。目前农民的文化知识增长很快，农民的文化素质也在不断的提高，很多农村已经接入了宽带网络，这对于农民对农业相关知识的获取提供了良好的平台。

随着信息技术的发展，虽然国内外也涌现出了很多农业相关网站，如农搜网和神州蔬菜网等百余家。但这些传统的搜索引擎依然继承了传统网络信息服务的弊病，极低的查准率让本来不怎么习惯网络的农民更为烦恼，为此为用户提供个性化服务在农业网站上显得更为迫切和重要。

鉴于以上种种原因，本文结合用户偏好进行了农业智能问答系统的研究，这具有一定的实用意义。

1.2 国内外研究现状

在个性化服务研究的早期，建模技术并没有得到应有的重视。大量的研究集中在实现个性化服务的技术上，如推荐技术，信息检索技术，用户聚类技术等，用户建模技术只是这些研究中几笔带过的陪衬。然而随着个性化服务的发展和研究的深入，研究者逐渐意识到，个性化服务的质量不仅仅取决于具体的推荐技术，检索技术等，还取决于用户兴趣和偏好等特点的可计算描述，而后者尤其重要。所以，近年来，有关用户建模技术的研究开始从具体的个性化服务形式中脱离出来，作为个性化服务的基础技术来研究。

Syskill&Webert 是加州大学的 pazzani 等人开发的一个辅助用户浏览 Web 的导航工具，是针对单用户的系统。在用户浏览 Web 的过程中，Syskill&Webert 要求用户对每一个浏览过的页面标注“感兴趣”、“不感兴趣”或者“一般”，而后系统通过计算页面中单字与类别的互信息（Mutual Information）找出反映用户兴趣的关键词，构成用户

模型。比如，如果用户保存某个页面，则推测用户对该页面感兴趣；如果用户经常访问某页面，则可推测用户对该页面感兴趣；如果用户点击页面中某个超链接而后又快速返回，则可推测用户对该超链接的链宿页面不感兴趣；假设用户浏览习惯是从左至右、从上至下，如果用户跳过某个超链接，则可推测用户对该超链接的链宿页面不感兴趣。这种建立用户偏好模型的方法太过笼统，也极大的干预了用户的使用，导致系统的效率较低，用户使用起来较为不方便。

Letizia 是由 MIT 的 Henry Lieberman 开发的 Web Agent。该系统通过收集用户的操作和浏览行为，运用启发式规则集，对用户的浏览行为建模，从而产生用户的偏好模型。系统并不要求用户给出显式的评价，主要通过分析用户在客户端的浏览行为推断用户的个性偏好。这样偏好模型的建立比 Syskill&Webert 更加人性化。

Personal WebWatcher 是由卡内基.梅隆大学推出的个性化系统。Personal WebWatcher 的个性化服务是在服务器端提供的。它主要由两部分组成:代理服务器(proxy server)和学习器(learner),代理服务器是用户 Web 浏览器与 Web 之间的桥梁，它保存了所有访问过的 URL 地址,学习器主要是为系统提供用户模型,整个系统用 Perl 语言和 C++语言编写。

代理服务器主要由 3 部分组成:代理(proxy)、建议器(adviser)和分类器(classifer).当代理接到一个请求时，先下载请求的文档,如果该文档是 HTML 格式,将会加上一些建议并将结果发给用户，增加建议的过程是这样的:代理将下载的文档发给建议器，建议器先从文档抽取超链,接着将结果发给分类器,分类器利用学习器产生的用户模型推荐满足某一阈值的超链返回给用户。

学习器有两种版本:从头开始创建一个新模型的学习器和更新一个已存在模型的学习器.它们之间的差别是：前一个不得不定义好领域信息,同时从一个空模型开始学习;后一个可以利用已定义好的领域信息,同时修改已存在的模型。系统假定被用户访问过的文档是揭示用户兴趣的,也就是正面的例子,所有其他被忽略的文档就是负面的例子,这个假定可以省去用户参与系统学习的过程.用户的兴趣模型很简单，系统是根据指向文档的超链来预测文档的兴趣度，而不是真正文档的内容,因为检索文档是一个非常耗时的过程,在系统空闲的时候(比如晚间),也可以根据文档的内容来预测文档的兴趣度,这样会更准确一些.此外,也可以利用超链来预测文档内容，然后利用这些文档内容来预测真正文档的兴趣度。仅仅通过超链接来推断用户的兴趣，可能导致系统不能很好的理解用户的意图，当然也就不能准确的推断出用户的兴趣所在，建立起较为精确的用户偏好模型。

2007 年西安工业大学计算机科学与工程学院的李宝敏、韩岳松在本体环境下用户偏好库的查询算法扩展一文中，依据本体论，采用用户兴趣剖像算法，扩展查询算法，使得在基于本体的智能搜索中更多更准确地体现出用户兴趣爱好，进而大大提高了检索的速度和查准率,但其实验系统的关键词只是果品方面的词汇，并且在单机的条件下进行了测试，其结果具有一定的局限性；2009 年曲阜师范大学的孔繁超在个性化信息服务中用户偏好的动态挖掘一文中采用聚类、关联规则等技术，对用户偏好进行了动态的挖掘，通过追踪用户需求序列，最终产生 Top-N 产品推荐，提高了推荐

系统的推荐质量，该推荐系统目前被广泛应用于电子商务、数字图书馆等领域,但随着进一步的应用,原有算法暴露出许多缺点；2005 年兰州大学的蒋萍、崔志明在基于用户兴趣挖掘的个性化模型研究与设计一文中，采用遗忘因子的概念对用户的兴趣进行了有选择的平等的遗忘，将用户模型应用于 PMSE，提高的查询效率，但系统的遗忘算法存在一定的弊端，对不同的兴趣类型采取了一致的遗忘算法，这样影响了用户主要兴趣的体现；2009 年盐城工学院的赵雪梅、朱恩亮在网站用户偏好度的数据挖掘模型一文中，基于统计学的观点讨论了网站用户偏好度的数据挖掘模型，设计了一个网站用户信息浏览偏好度的数据挖掘模型并应用于通用商品销售系统，也在一定程度上提高了产品的查准率。

除此之外，很多学者还对用户模型的表示和更新进行了大量的研究和探索。然而，总的来说，用户偏好建模技术的研究还处于起步阶段，依然没有形成完整的技术体系，很多关键技术善待解决和探索。

与现存的偏好模型相比，本文所设计的用户偏好模型具有以下特点：

（1）模型中不但保存了用户的兴趣主题信息以及其兴趣度，还保存了该兴趣主题所对应的特征词信息与其权重，将个人偏好进行了细化。

(2)将用户偏好分为静态偏好，短期偏好，长期偏好，对不同的类型的偏好采取不同的遗忘算法和不同的结构模式。

（3）提出了一种新的遗忘算法，加入了遗忘因子修正值，使用户偏好更有效的服务于信息检索。

1.3 研究内容和设计目标

论文采用面向对象的思想方法，结合本体技术，对农业智能问答系统中的用户偏好问题经行详细的分析，并采用 JAVA 技术，结合目前比较流行的 SSH 框架对农业智能问答原型系统进行了实现和相关分析。

论文完成了以下几个方面的工作：

（1）全面研究了信息检索系统的关键技术，重点研究了构建基于本体的用户偏好模型的关键技术和核心问题。

（2）对用户偏好模型进行了重点研究，探讨了用户偏好模型的表示方法，建模方法，更新方法等关键技术，并提出了一种新的用户模型更新算法。

（3）在充分学习和研究了传统的用户偏好模型和构建技术的基础上，搭建了农业智能问答系统的原型系统。

（4）认真设计了测试数据，对文中的算法和用户偏好模型进行了试验和分析，验证了其有效性。

1.4 论文的结构

第一章：引言。主要介绍了论文研究课题的背景、来源、意义及作者的主要工作。

第二章：检索技术研究以及相关技术介绍。本章主要介绍了传统搜索引擎的原理、分类和不足，以及个性化服务的特点和关键技术。

第三章：对农业智能问答系统中的用户偏好进行了详细分析和设计。本章探讨了偏好挖掘技术，偏好建模技术，模型的分类和构建方法，以及模型的更新方法等，并对用户偏好的存储进行了详细设计。

第四章：智能问答系统的详细设计和实现。本章结合给出的原型系统框架图对其进行了详细的功能模块设计实现，并对关键模块进行了流程和效果截图的说明。

第五章：系统测试。测试了短期兴趣和长期兴趣的遗忘机制和短期兴趣向长期兴趣的转换，并且在采用模型和不采用模型的情况下对原型系统进行了测试，根据实验数据和结果证明了该系统在使用偏好模型后，在保证查全率的前提下，提高了其查准率，体现了用户的个性化，需要符合选题要求。

2 个性化搜索引擎简介

2.1 搜索引擎概述

为了更好的理解个性的问答系统，在此我们有必要先对搜索引擎做必要的介绍。那么什么是搜索引擎呢？搜索引擎(search engine)是指根据一定的策略、运用特定的计算机程序从互联网上搜集信息，在对信息进行组织和处理后，为用户提供检索服务，将用户检索相关的信息展示给用户的系统。

2.1.1 搜索引擎工作原理

抓取网页：每个独立的搜索引擎都有自己的网页抓取程序（spider）。Spider 顺着网页中的超链接，连续地抓取网页。被抓取的网页被称之为网页快照。由于互联网中超链接的应用很普遍，理论上，从一定范围的网页出发，就能搜集到绝大多数的网页。

处理网页：搜索引擎抓到网页后，还要做大量的预处理工作，才能提供检索服务。其中，最重要的就是提取关键词，建立索引文件。其他还包括去除重复网页、分词（中文）、判断网页类型、分析超链接、计算网页的重要度/丰富度等。

提供检索服务：用户输入关键词进行检索，搜索引擎从索引数据库中找到匹配该关键词的网页；为了用户便于判断，除了网页标题和 URL 外，还会提供一段来自网页的摘要以及其他信息。

2.1.2 搜索引擎的分类

图片搜索图片：图片搜索引擎是全新的搜索引擎，目前国内有安图搜。基于图像形式特征的抽取：由图像分析软件自动抽取图像的颜色、形状、纹理等特征，建立特征索引库，用户只需将要查找的图像的大致特征描述出来，就可以找出与之具有相近特征的图像。这是一种基于图像特征层次的机械匹配，特别适用于检索目标明确的查询要求（例如对商标的检索）。产生的结果也是最接近用户要求的。但目前这种较成熟的检索技术主要应用于图像数据库的检索，在网上图像搜索引擎中应用这种检索技术还具有一定的困难。

全文索引：全文索引引擎是名副其实的搜索引擎，国外代表有 Google，国内知名的百度搜索。它们从互联网提取各个网站的信息（以网页文字为主），建立起数据库，并能检索与用户查询条件相匹配的记录，按一定的排列顺序返回结果。根据搜索结果来源的不同，全文搜索引擎可分为两类：一类拥有自己的网页抓取、索引、检索系统（Indexer），有独立的“蜘蛛”（Spider）程序、或爬虫（Crawler）、或“机器人”（Robot）程序（这三种称法意义相同），能自建网页数据库，搜索结果直接从自身的数据库中调用，上面提到的 Google 和百度就属于此类；另一类则是租用其他搜索引

擎的数据库，并按自定的格式排列搜索结果，如 Lycos 搜索引擎。

目录索引：目录索引虽然有搜索功能，但严格意义上不能称为真正的搜索引擎，只是按目录分类的网站链接列表而已。用户完全可以按照分类目录找到所需要的信息，不依靠关键词（Keywords）进行查询。目录索引中最具代表性的莫过于大名鼎鼎的 Yahoo、新浪分类目录搜索。

元搜索引擎：元搜索引擎（META Search Engine）接受用户查询请求后，同时在多个搜索引擎上搜索，并将结果返回给用户。著名的元搜索引擎有 InfoSpace、Dogpile、Vivisimo 等，中文元搜索引擎中具代表性的是搜星搜索引擎。在搜索结果排列方面，有的直接按来源排列搜索结果，如 Dogpile；有的则按自定的规则将结果重新排列组合，如 Vivisimo。

垂直搜索引擎：垂直搜索引擎为 2006 年后逐步兴起的一类搜索引擎。不同于通用的网页搜索引擎，垂直搜索专注于特定的搜索领域和搜索需求（例如：机票搜索、旅游搜索、生活搜索、小说搜索、视频搜索等等），在其特定的搜索领域有更好的用户体验。相比通用搜索动辄数千台检索服务器，垂直搜索需要的硬件成本低、用户需求特定、查询的方式多样。

2.1.3 传统搜索引擎的不足

搜索引擎在信息检索方面存在的许多不足主要表现在：①查全率比较低。每个引擎的覆盖面都相当有限,据统计，没有一个搜索引擎的索引量超过整个网络网页总数的 1/6。由此可见，每个搜索引擎虽然相对查全率比较高，而实际查全率则比较低。②查准率比较低。目前通过搜索引擎检索的网络信息资源相关性非常差，浪费了用户大量的相关判断时间。③每一个搜索引擎都有自己的检索规则，用户利用不同的搜索引擎需要进行不同的适应过程，增加了用户的负担。④多数搜索引擎采用关键词检索，并提供高级检索功能，但用户很难通过组配关键词来准确表达自己的信息需求，导致检索效率低下。⑤更新速度比较慢。搜索引擎机器人只能在由系统管理员确定的一定时间间隔内跟踪特定信息，不能保证信息的及时更新，导致产生错链和死链。随着网络信息数量的指数增长，引擎数据库急剧膨胀，检索速度也将会变慢。为了解决单个搜索引擎信息覆盖面小，信息收集量有限，用户需要对不同搜索引擎进行适应的缺点，人们提出了个性化搜索引擎的概念。

2.2 个性化搜索引擎

2.2.1 个性化搜索引擎

通过上述介绍，我们了解了什么是搜索引擎，由于传统搜索引擎的不足，我们必须来构建个性的搜索引擎来弥补它的缺点。个性化搜索引擎最终要为用户提供个性化服务。根据目前个性化搜索技术实现的现状，搜索引擎要实现个性化搜索服务，可以

从以下几方面考虑[8,9]：

(1)根据用户个性特征以及用户搜索的历史信息，建立用户兴趣模型。只有构建了用户兴趣模型，搜索引擎才能根据用户的兴趣特征来提供个性化服务。用户兴趣模型是个性化搜索引擎实现的基础。

(2)还要有一个合理的机制来完善和更新用户兴趣。用户跟踪的方法可以分为显式跟踪和隐式跟踪。显式跟踪是指系统要求用户提交自己感兴趣的信息。而隐式跟踪不要求用户提供信息，所有跟踪都是由系统自动完成，隐式跟踪又可分为日志挖掘和行为跟踪。

(3)用户查询通常具有歧义性，而用户真正的查询意图大部分是直接让成员引擎去推理识别。但是不同的成员引擎有不同的推理机制，从而导致搜索引擎返回大量与用户不相关的结果。因此，为了给用户提供个性化搜索，个性化的搜索引擎需要对提供的用户查询进行查询优化，并参考用户兴趣模型最大可能地识别用户的查询意图。

(4)个性化元搜索引擎应该能够参考用户兴趣模型选择最合适用户查询的成员引擎来为用户提供搜索服务。

(5)个性化搜索引擎获得各个成员引擎返回的结果后，在合成结果时参照用户兴趣模型，过滤与用户查询不相关的结果和对结果重新进行排序。这样，即使成员引擎返回结果一样，但对于不同的用户，最终返回给用户的结果也是不一样的，充分体现用户个性化搜索。

2.2.2 个性化搜索引擎的关键技术

个性化元搜索引擎涉及的技术较多，如用户建模技术，个性化推荐技术，网站自适应技术，用户隐私保护技术等。但目前研究较多、也是最为关键的技术是用户建模技术和个性化推荐技术[8,9]。

(1)推荐技术研究目前主要的推荐技术包括基于内容过滤和协同过滤两种。由于基于内容过滤自身局限性，协同过滤推荐技术是当前研究主流。

(2)实时性研究在大型个性化推荐系统中，推荐系统的伸缩能力和实时性要求越来越难以保证。如何有效满足推荐系统的实时性要求得到了越来越多研究者的关注。

(3)推荐质量研究在大型个性化服务系统中，用户评分数据极端稀疏。用户评分数据的极端稀疏性使得推荐系统无法产生有效的推荐，推荐系统的推荐质量难以保证。

(4)多种数据多种技术的集成当前大部分的电子商务推荐系统都只利用了一部分可用信息来产生推荐。随着研究的深入，新型个性化推荐系统应该利用尽可能多的信息，收集多种类型的数据，有效集成多种推荐技术，从而提供更加有效的推荐服务。

(5)数据挖掘技术在推荐系统中的应用随着研究的深入，各种数据挖掘技术(主要包括关联规则挖掘、序列模式挖掘、聚类分析等)在推荐系统中得到了广泛的应用。基于 Web 挖掘推荐系统得到了越来越多研究者关注。

(6)用户隐私保护研究由于推荐系统需要分析用户的兴趣爱好，涉及到用户隐私

问题，如何在提供推荐服务的同时有效保护用户隐私值得做进一步深入的研究。

(7)推荐系统可视化研究推荐系统的目的是为用户提供服务，因此必须为用户提供友好的可视化界面服务。主要包括推荐结果可视化研究和推荐结果解释研究等方面的内容。

无论何种形式的个性化服务，都需要首先建立对用户的描述，然后才能据此提供针对不同用户的个性化服务，因此，用户模型是个性化服务的基础和核心。在接下来的第三章，我们将具体介绍用户偏好的建模技术，更新技术，持久化技术和方法等。

3 智能问答系统中用户偏好的分析与设计

3.1 用户偏好挖掘技术

3.1.1 数据挖掘技术简介

数据挖掘(DataMining)[14-15]：概括的说数据挖掘就是从模糊的、不完全的、随机的、大量的、有噪声的实际应用数据中，提取潜在而有用的信息和知识的过程。

数据挖掘技术来源于人们对数据库技术不断进行探究和开发的基础之上。在信息技术不太发达的情况下，各种数据都统统的被存储在数据库或者文件当中，随着信息技术的发展，人们可对数据库进行查询和访问操作，随后发展到对各种数据库的即时遍历。数据挖掘技术将数据库技术推到了一个更高点，它不仅能对以前的数据进行正常的查询乃至遍历，还可以找出这些"过时"数据之间的潜在联系，从而促进信息的传递。数据挖掘是一门综合学科，有许多的功能，现将主要功能如下：

1.分类：通过分析对象的属性和某些主要特征，为对象分门别类来描述之。例如：网站可以根据以前的用户的访问数据进行分析，可将用户分成不同的类别，比如玉米种植组，土豆种植组，水稻种植组，甲鱼养殖组等等。

2，聚类：简言之就是将一个对象集合根据某些条件分成几个类别，每个类别中的各个对象都是相似的，但与其他类的对象之间是不相似的。

3. 序列模式和关联规则的发现：关联就是指一种事物触发时与另外的事物发生的某种关系，可通过关联的可信度和支持度来描述。与关联截然不同，序列是指事物之间一种纵向的联系。

4.预测：把握并分析事物发展的规律，对其未来的发展趋势做出预测。

随着科学技术的发展和人们需求的提升，数据挖掘技术已经越来越多的应用到了基于 Web 的挖掘上。因特网上信息量无比丰富和繁杂，人们如何从非结构化的数据信息中高效的提取出有价值的信息是数据挖掘领域中的一项挑战。Web 上的数据信息十分杂乱无序，没有固定的规范，它不像我们熟悉的数据库，数据之间有规范的结构，如关系数据库，它有统一的范式，其中的信息为完全结构化的。就处理的信息而言，传统的数据挖掘技术难以处理异质的非结构化信息，而 Web 挖掘技术克服了这方面的缺点，它处理的数据主要是大量异质的 Web 信息资源，对文档结构性差，非结构化数据具有很好的挖掘力度。由于非结构化和半结构化的信息几乎不能用数据模型来清晰的表示，因此 Web 数据挖掘技术是基于很多数据仓库挖掘技术的一门新兴科学。对 Web 上的数据进行挖掘具有极大的挑战性，简言之就是针对 Web 页面内容、用户访问信息、用户注册信息、站点拓扑结构及用户的问答等信息在内的各种数据，应用数据挖掘方法以发现有用的知识的过程。它可以帮助人们从 Web 中发现新知识，改进站点设计，提供个性化服务。Web 挖掘主要可分为结构挖掘、使用挖掘和内容挖掘。

如图 1 所示：

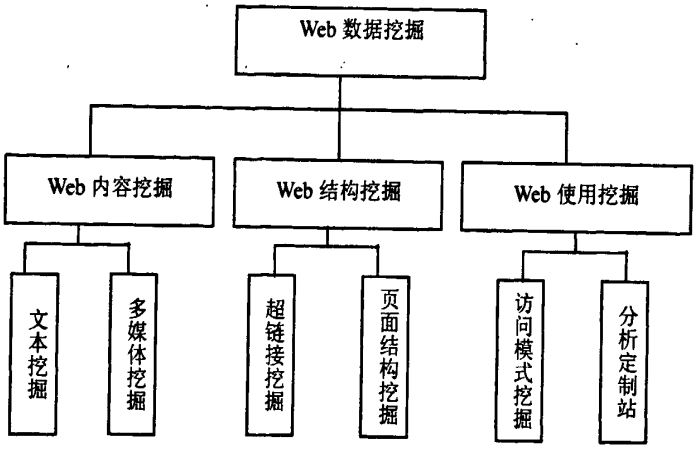

图 1 Web 数据挖掘结构图

Fig.1 the structure of Web data mining

3.1.2 Web 日志挖掘过程

服务器的日志记载了访问者的 IP，浏览器的类型，页面的大小，所访问过的页面，以及各种响应处理等等。Web 日志挖掘具体可以分为以下三个阶段：

Web 日志挖掘首先必须对日志数据进行预处理[16-22]，其任务是通过对用户访问数据的分析，组织转为数据挖掘所要求的格式并保存起来，形成会话文件等待下一步处理。预处理过程主要包括数据清洗，用户识别，会话识别和路径补存四个阶段。

表 1 WEB 日志文件记录

Tab.1 the records of Web logs

date	time	s-sitename	s-ip	cs-method	cs-uri-query	s-port
2011/3/1	09:00:01	TEST	211.68.183.152	GET	id=13809	8080
2011/3/1	09:05:09	TEST	211.68.183.152	GET	-	

c-ip	cs(User-Agent)	cs (Referer)	cs-host	sc-ststus	sc-substatus
211.68.183.151	Mozilla/4.0+(compatible; +MSIE+6.0;+Windows+NT+5.0)	...	...	200	0
211.68.183.151	Mozilla/4.0+(compatible; +MSIE+6.0;+Windows+NT+5.0)	...	...	206	0

表 2 WEB 日志格式

Tab.2 the formate of Web logs

域	date	time	s-sitename	s-ip	cs-method	cs-uri-stem	cs-uri-query	s-port
说明	请求日期	请求时间	服务器名称	服务器 IP	请求方法	请求页面	请求查询	服务器端口号

s-username	c-ip	cs（User-Agent）	cs（Referer）	cs-host	sc-status	sc-substatus	sc-win32-status
服务器端用户名	用户 IP/DNS 入口	用户代理	当前页面的 URL	主机	返回 http 的状态标识	协议子状态	Win32 状态

数据清洗：数据净化的主要任务是根据不同的应用需求来清除原始数据中不相关的数据项，例如：不代表用户兴趣的自动下载并记录在日志文件中与访问页面有关的图片音频等信息。通常可以配置一个删除列表，凡是后缀名在删除列表中的记录都需要清理掉，同时 404,301,500 等的传输错误记录也应删除，并将删除后的记录按一定的规则进行排序处理。

表 3 数据清洗后的 web 日志记录

Tab. 3 the records of Web logs after cleaning data

date	time	cs-uri-stem	cs-uri-query	c-ip	Cs-(User-Agent)	cs (Referer)
2011/3/1	10:02:01	nongye/index.jsp	id=5925	211.68.183.151	Mozilla/4.0+(compatible; +MSIE+6.0; +Windows+NT+5.0)	-
2011/3/1	10:03:06	nongye/search.jsp	id=5918	211.68.183.153	Mozilla/4.0+(compatible; +MSIE+6.0; +Windows+NT+5.0)	-
2011/3/1	10:04:25	nongye/search.jsp	id=5933	211.68.183.154	Mozilla/4.0+(compatible; +MSIE+6.0; +Windows+NT+5.0)	-
2011/3/1	10:06:22	nongye/advsearch.jsp	id=5925	211.68.183.155	Mozilla/4.0+(compatible; +MSIE+6.0;+ Windows+NT+5.0)	-
2011/3/1	10:08:58	nongye/advsearch.jsp	id=5981	211.68.183.156	Mozilla/4.0+(compatible; +MSIE+6.0;+ Windows+NT+5.0)	-

用户识别：每个用户都是都是一个独立的个体，他通过浏览器来访问站点。但是

由于本地高速缓存，代理服务器，防火墙等的存在，使得识别用户较为困难。目前的用户识别大多采用以下三条启发式原则：（1）如果用户的 IP 地址不同则认为是不同的用户。（2）如果 IP 地址相同，但浏览器软件或操作系统不同（使用代理上网的局域网用户），则认为是不同的用户。（3）如果 IP 相同，而且所使用的上网环境相同，那么可根据网站的拓扑结构对用户进行识别，如果用户所访问的页面不能通过已访问页面的任何超链接到达，则认为是一个新的用户。

表 4 用户识别结果

Tab. 4 the result of users recognition

用户 ID	用户 IP	客户端浏览器和操作系统
1	211.68.183.151	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.0)
2	211.68.183.153	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.0)
3	211.68.183.154	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.0)

会话识别：会话识别就是将用户的访问记录划分成单个的会话，不同用户访问的页面属于不同的会话，日志文件中不同的页面也属于不同的会话。由于在较长的时间里，用户可能多次访问了该站点，也很难知道用户是否为分开几次登录。所以一般利用最大的超时来判断用户是否已离开了该网站，若两次请求时间之间超过了一定的时间界限，就会被认为用户的一个会话已结束，开始了一个新的会话。

表 5 用户会话识别表

Tab. 5 the session recognition of user

会话 ID	用户 ID	用户 IP 地址	URL 访问页面序列	访问时间（分钟）
1	1	211.68.183.151	p0,p30,...	0.78，2.3，...
2	2	211.68.183.153	p0,p50,...	2.8,0.5，...
3	3	211.68.183.154	P1,p30,...	4.1,2.2，...

3.2 用户偏好模型研究

用户兴趣模型毫无怀疑是每个个性化系统的基础和核心，用户兴趣模型的建模也就是指系统从与用户相关的各种信息中构建出相应的用户模型。

3.2.1 用户兴趣模型的含义

那么什么是用户兴趣模型？所谓用户兴趣模型通常是指用来表示用户对信息相对稳定的兴趣需求的模型，它可以直接反映出当前用户在某一段时间内对信息需求的主要倾向，随着对用户操作行为的跟踪和对用户反馈信息的整理收集等，同时系统将按照特定学习方法对用户兴趣模型进行及时更新，以使其更好地体现出用户的兴趣，符合用户的真正需求。

用户偏好模型的建立是一个循序渐进的过程[23],终极目的是准确反映出用户的真正兴趣和要求。初始阶段只是一般性的兴趣模型，具有其基本特征。经过用户行为跟踪处理和用户反馈信息处理后，系统对兴趣模型不断更新和修正,以使其更准确地代表用户在某时间段内的兴趣倾向和需求。兴趣模型的演进如图 2 所示：

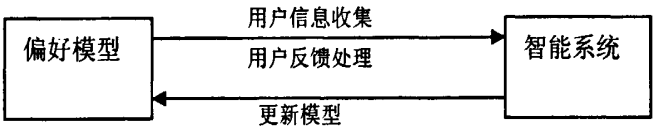

图 2 兴趣模型演进图

Fig. 2 the evolution of interests model

3.2.2 偏好模型的形式

用户偏好模型是指被系统所提取的用户信息需求、兴趣主题等的结合。偏好模型的表现形式在很大程度上决定了其模型反映用户信息需求真实性，同时也在一定程度上制约了个性化推荐的准确性。个性化偏好模型的表现形式至今仍没有统一的标准，常用的表示方法有以下几种[24]：

（1）概念表示法

概念表示法是指以用户感兴趣的主题概念来表示用户兴趣模型的方法。如果用户对苹果和梨兴趣，那么用户兴趣模型表示为{苹果，梨}。用户的兴趣概念既可以通过用户自己填写表单等资料提供，也可以由系统根据一定的规则和方法学习获得，也就是从用户所访问过的内容中根据规则提取出能够反映用户兴趣的关键词用以表达用户的偏好。

(2)关键词表示法

关键词列表法是指以用户感兴趣的关键词来表示用户兴趣模型的方法。如果用户对茶叶感兴趣，则用户兴趣模型可以表示为{花茶，绿茶，红茶}等。关键词组既可以由用户指定，也可以通过系统学习获得。系统通过学习的方法提取到能够表示用户偏好的关键词表与文本分类中的特征词的提取大同小异，二者全是通过训练样本后得到一定数量特征词集合，而不同的是后者是为了减少分类器的计算量并改善分类精度。

(3)向量空间模型表示法

向量空间模型法是指用关键词向量空间中的向量来表示用户兴趣模型的方法。vector Space Model (向量空间模型)是 60 年代末由 Gerald Salton 等人提出的，是计算机处理文档的常用方法，也是最早最出名的一个数学模型。此模型中，文档被当做是由一组正交词条向量所组成的向量空间来处理，每个文档 D 可以被表示为其中的一个特征向量{(t1，w1)，(t2，w2)，...，(tn，wn)}，其中 tn 为词条项，wn 为该项在 D 中的权值。实践证明向量空间模型法(vector Space Model)可以很好的表示用户偏好模型，但是随着文档数量的不断增加，模型会逐渐地增长，因而基于 VSM 的偏好模型表示法是以巨大的时间和空间复杂度为代价的。

(4)本体表示法

本体(Ontology)是来源于哲学领域，是对存在的系统化解释。在知识工程领域中，有关研究者将本体定义为概念化对象的明确表示。由于本体对特定领域对象的表示与描述具有规范性、可靠性等特点，所以本体被应用于信息检索领域，对文档和用户兴趣模型进行描述。本体从不同层次的形式化表示，给出概念之间相互关系的明确定义，并通过概念之间的相互关系来描述概念的语义。

3.2.3 兴趣模型的创建方法

依据模型构建过程中用户是否参与以及参与的程度，可以大致将建模技术分为用户手工定制、示例用户引导和自动分析建模三种[25]。

(1)用户手工定制

用户手工定制建模是指在建模过程中用户自己输入其感兴趣的主题词，用户既可以输入关键词，也可以选择覆盖性全面的关键词列表。在个性化服务发展的早期，该定制模式是建模的主要方法。

(2)示例用户引导

示例用户引导是指由用户主动提供相关的实例集和类别属性来建立偏好模型，如要求用户对其浏览的页面给出兴趣度。该方法的缺点是在用户浏览过程中，要求用户给出其对浏览内容的兴趣度值，严重影响了用户的正常浏览。理想的建模方法应该是无须用户参与，系统根据用户的浏览内容和操作行为隐式自动的建立起用户偏好模型。

(3)自动分析建模

自动分析建模是指在偏好模型的定制中无须用户参与，系统可以根据用户的浏览内容和浏览行为自动地为各个用户建立起其偏好模型的方法。在现有的个性化系统中，采用该方法建模的系统主要有 Personal Web WatCher,ELFI,LetiZia 等。总体来说，自动分析建模有利于提高个性化服务系统的质量和友好性,促进了个性化服务的发展。

3.2.4 兴趣模型更新方法介绍

用户偏好大致分为两种[26]，即长期兴趣和短期兴趣。在一段较长时间段内相对稳定的兴趣可以称之为长期兴趣，例如用户长时间内经常浏览苹果方面的信息，可以说明苹果是用户的长期兴趣，这种相对稳定的兴趣一般不会突然改变。相反变化较快的兴趣可以称为短期兴趣，即用户为了某种短期行为而关注的信息。例如为了购买苹果手机而关注相关的价格信息，那么该偏好可以说是用户的短期兴趣。伴随着时间和环境的改变，用户偏好也会随之逐渐发生改变。系统及时的捕捉这种变化，保持系统偏好模型与用户兴趣同步，才能不断的提高系统检索质量。

用户偏好模型的表示技术直接决定着其更新的方法，只有二者相配套才能使系统达到最好的服务质量。只有做到对偏好模型进行及时准确的更新，才能确保检索结果真正体现用户的真正需求，才能确保系统提供更优质的服务。

当前使用较为广泛的用户偏好更新技术是信息增补技术，它大致可分为两种，即直接增加信息和包含原信息调整的信息增补。

直接增加信息的信息增补技术将所获取的新的偏好信息转换成为用户偏好模型的表示形式后直接追加到用户兴趣模型，并不对原模型中的过时的偏好进行删除或者调整。该方法的缺点是对短期兴趣的"反应"较慢，因为模型中过时的信息可能对个性化系统造成严重的干扰，而新增的兴趣信息可能还需要很长的时间才能在系统中体现出来，因此就无法在服务中及时的体现出用户的短期兴趣。并且随着时间的推移，用户偏好信息不断累加，可能会出现大量繁杂无用的偏好信息和模型维护困难等问题。GrouPLens、WEBSELL、PCFinder 等系统运用了这种更新技术。

与前者不同，包含原信息调整的信息增补技术根据模型中现存的偏好信息和系统新获取的信息来分析和推测出当前用户的兴趣变化，并根据一定的算法和规则调整相应兴趣特征的权值来体现出偏好的转移。这种方法虽然对信息的信息增补技术一定程度上有所改进，但如果选取的修改算法不合适可能会导致新旧偏好间的相互干扰，直接影响到系统的效率和稳定性。

基于神经网络的更新技术是一种自适应的更新技术。当用户的偏好随时间变化时，神经网络系统也将相应的调节网络链接的权重，从而改变网络输出来体现偏好的转移和变化。

基于遗传算法的兴趣模型更新技术是一种基于自然选择和遗传机理的迭代搜索优化技术。遗传算法的基本思想是把代表问题可能潜在解集作为初始种群，在每一代进化中，根据种群中个体在问题域的适应度大小挑选个体，并借助自然遗传学的遗传算子进行交叉和变异产生新的种群。这个过程将导致种群更加适应环境，末代种群中的最优个体经过解码，可以作为问题近似最优解。基于遗传算法的更新技术可以把用户兴趣模型表示成初始种群，种群的进化就对应着用户兴趣模型的更新。

3.2.5 本体偏好模型的定义和原理

偏好模型[27,28]：Hobby={(Htl，Cl)，(Ht2，C2)，...，(Hti，Ci)，...，(Htn，Cn)}，其中 Hti=(ti，wi)(1<=i<=n)是一个二元组，表示用户偏好节点，ti 为用户感兴趣的主题， wi 表示兴趣主题 ti 的权值，n 为兴趣主题的总数，所有的 ti 构成用户兴趣主题集 T。Ci 是属于主题 ti 的兴趣特征词集，如果主题 ti 包含 m 个特征词，则 Ci 可以表示成 Ci={(eil，vil)，(ei2，vi2)...，(eij，vij)...，(eim，vim)}，其中 eij(l<=j<=m)表示兴趣主题 ti 包含的所有特征词，m 表示主题 ti 下所有特征词的个数，vim 表示对应特征词在用户兴趣 ti 中的权值。用户兴趣特征词集 C(T)是用户兴趣集中所有兴趣特征词集的并集，即 C(T)=Uci。这可用偏好模型树形象的表示。

使用 Ontology 建模的优点有很多[26]，可重用、便于查找、可靠性、有助于任务解析以及可维护性等。基于本体的用户模型用规范的结构模式描述用户的偏好，把用户偏好与语义概念层次相结合，有更强的表达能力。

第一，能够描述复杂的文本结构。每个文本都有很多的文本概念，文本的结构概

念构成文本的概念层次，文本的属性概念构成文本的属性概念层次，每个文本概念都有相应的值。

第二，能够描述偏好的语义。用户偏好中的一个词条往往包含丰富的语义。Ontology 模型能描述该词条对应的用户偏好层次概念，使偏好带有丰富的语义。

第三，可扩展性。ontology 定义了文本结构上的偏好，不同的文本涉及的概念是不固定的，Ontology 能随着文本结构的扩展而扩展。

第四，自适应性。随着现实世界知识体系的变化以及用户偏好的变化，基于本体的用户模型能够反映当前语义概念层次以及用户偏好概念层次，从而自动适应这些变化，准确表达用户的当前偏好。利用本体论的思想对领域知识库进行建模，可以使相互独立的层次有机地组成一个完整的系统，可以实现“领域知识”的共享和重用，而且“领域知识”概念模型的形式化描述便于正确性检查，可以使领域知识库的结构更加清晰，更有利于领域知识库的维护。

3.2.6 偏好模型的分类

静态用户偏好记录了当前用户的基本信息,比如用户的年龄、学历、专业、爱好等。这种偏好的特点是结构稳定,容易操作。在基于本体的用户偏好库中,每个用户都可用相对应的一个类来表示。每个类都代表唯一一种类型的用户,而类的每个实例代表一个具体的用户。该层次结构的优势主要体现在两个方面:一是在无法获取用户信息或用户信息不全的情况下可以通过类继承关系推导出其基本的特征属性；二是系统可以针对不同的用户类型建立不同风格的页面。

动态用户偏好见名思义就是指那些不稳定的，经常变化的用户兴趣。因为在大多情况下用户偏好信息属于一种隐性信息,用户很难用言语一次性的将之概括完全。还有用户的偏好与之年龄和从事的工作，专业等个人特性息息相关,即用户偏好是动态可变的。为此就要求系统不断收集用户信息,即时的调整其兴趣模型。

3.2.7 偏好模型生成算法

静态偏好生成算法：静态偏好在相对一段长的时间内是稳定的，我们将用户的专业、职业和工作区域赋予一个较高的权值，用来聚类用户并为用户推送群内消息,例如某一个用户群 50 个人，这些用户均为河北张家口阳原县的玉米种植户，其学历层次都是小学或者初中，那么系统会把该群可能感兴趣的中文问答对，推送给群内的各个用户，当然如果该群内的任何一个用户提出的新问答对也会首先推送给群内用户。其它属性，诸如用户的年龄、学历、性别、视力等作为一条普通的数据存入数据库，利用它们来控制页面字体的大小，颜色的搭配等。比如我们可以为 50 岁以上和视力很弱的人设置大字体，为色弱人群设置特定的颜色或为色盲人群进行相关颜色的文字标识等操作，这样使用户使用系统更方便，更贴心，更人性化。

动态偏好生成算法：

（1）置偏好兴趣树的根结点为用户名；

（2）置偏好兴趣树中叶节点的初始权值为 vim=5；

（3）参考偏好模型，生成偏好兴趣树，并自下而上逐层计算各父结点（各兴趣主题）Node（ti）的权值 $wi=\sum_{i=1}^{k} vi$，k 为子类的个数，vi（0<i<k+1）为其子类的权值；

（4）将兴趣树信息存入用户兴趣表 hobby。如下图 3 所示

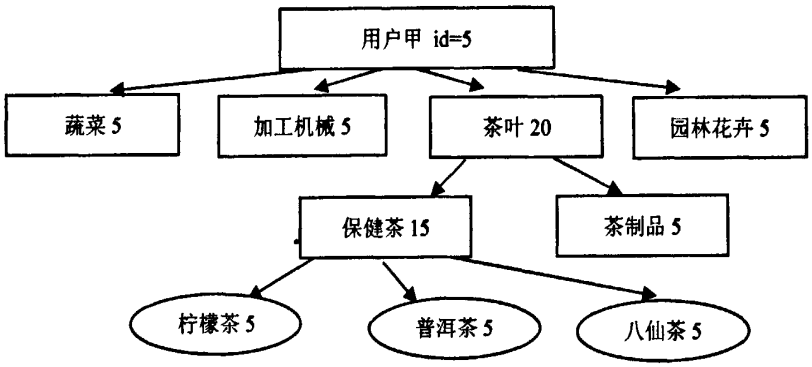

图 3 兴趣偏好模型中兴趣度的累积

Fig. 3 the accumulation of interest's degree in the model of interests and prefernces

显而易见，(1)如果我们为用户推荐信息时会推荐蔬菜类信息，加工机械类信息，茶叶类信息，园林花卉类信息，而诸如玉米种植，种猪养殖等信息则不是用户感兴趣的，应该过滤掉，其中在茶叶的推荐中，系统会着重推荐保健茶，而在保健茶中首推柠檬茶，普洱茶和八仙茶，这便体现了用户的偏好，做到系统更贴心；（2）如果用户提交查询时，例如用户提交的关键词为茶叶，那么系统首先会把保健茶中的柠檬茶，普洱茶，八仙茶检索出来排在结果的前列，次之是茶制品，如冰红茶，康师傅绿茶等，而诸如茶叶加工设备，废茶叶的处理等信息则被过滤掉，或者作为上下位词提供给用户，这样便体现了用户偏好所在，使用户更容易的找到自己想要的东西。

3.2.8 本体偏好模型的具体更新算法

用户的偏好和需求在某一段时间内是相对稳定的，但又不是一成不变的。当用户偏好和需求改变后，就要求系统对现有的偏好模型进行优化和更新。同时为了避免过时的主题词条对信息检索造成干扰，提高检索精度，同时减少存储和计算的开销，系统必须限制兴趣节点的数量，当节点数超过了制定的数量时，便要删除权重小于某一阀值的兴趣特征。

人类记忆遵循自然遗忘的规律，自然遗忘是一个渐进的过程。在此我们假设系统对偏好的遗忘也遵循这个规律，即用户的偏好也随着时间不断减弱，并且系统对偏好的遗忘速度也是从快到慢。用户近期频繁访问的词条无可置疑最能代表其最近的偏好趋向，而相对长的时间内不再访问的词条，说明用户的该偏好已经减弱或者消失，因此系统可以通过让其不断“衰老”达到去除过时词条的目的。为此，我们引入遗忘因

子对用户的现存偏好模型逐渐调整，以保证用户偏好的时效性。在对模型更新时，不仅要加入当前代表用户最新的主题词，还要对系统模型中现存的特征词进行权值的调整，即对现有词条的权值乘以遗忘因子 F(x)进行修正。遗忘因子 F(x)为 $F(x)=e^{-[\frac{\log_2(cur-fir)}{ha}\pm\delta]}$ 其中 cur 表示当前日期，fir 表示兴趣特征序词第一次在模型中出现的日期，ha 表示半衰期，ha 值应根据大量实验测试确定，也可以根据经验人为设定，δ(δ>0)为遗忘因子修正值，对于词条权值小于某一阀值 ε 的，公式中取+，否则取-，以尽快淘汰老词条。

与此同时，系统不断挖掘用户的当前兴趣，对当前用户频繁访问的特征词和主题进行权值增强，并删除一些小于某一阀值的主题词条。

3.2.9 短期兴趣向长期兴趣的转换

虽然系统对短期兴趣词条遗忘相对快，但随着时间的积累，如果用户频繁的访问该主题的话，兴趣权值就会不断的增大，系统会将兴趣权值大于某一阀值 f 的类以及对象的特征词转移到长期兴趣模型中。其具体算法描述如下：

（1）从短期偏好中筛选出符合要求的特征词和其相对应的类。

（2）依次将各词插入到长期兴趣中的对应类或新类中，并依次调整偏好树中对应节点的权值。

（3）从短期兴趣中删除该类和其对应的特征词，同时调整该类的遗忘因子和长期兴趣中对应的各个值属性。

（4）更新对应的长期兴趣类表。

3.2.10 本体相似度的计算

一个本体模型主要由一组概念的集合和一组语义关系的集合组成[29]，简单可描述如下：

语义关系：概念的语义关系主要有 Synonym 关系、Is-a 关系、Part-of 关系。其中 Synonym 关系代表两个概念之间是同义平等的关系,该关系是双向的。Is-a 关系表示一个抽象概念或泛指概念的具体化,该关系是单向的。而 Part-of 关系表示表示整体和部分的关系，这种关系是单向的。关系不同概念间的相似度也不同,比如 Synonym 关系的概念相似度要大于 Is-a 关系和 Part-of 关系。

(2)语义距离：即在本体图中连接两个节点的最短路径的长度。两个词语的语义距离越大,其相似度越低；反之,相似度越大。两者之间可以建立一种简单的对应关系。特别地当两个词语距离为 0 时,其相似度为 1;当两个词语距离为无穷大时，其相似度为 0，这可用公式 $dis(A,B)=\sum_{i=1}^{n} w_i$ 来量化。dis (A,B)表示概念 A 与 B 之间的语义距离，wi 表示连接 A,B 的最短路径上第 i 条边的权值。

(3)节点的深度：即概念与树根的最短路径中所包括的边数.因为在本体图中，每一层都是对上一层概念的细化，由此可见,在语义距离相同的前提下，两个节点的深

度和越大,概念之间的相似度越大;两个节点的深度差越小，概念之间的相似度越大可以用公式 $dep(A)=\sum_{i=1}^{n}1$ 来计算之。dep (A)表示节点 A 的深度,n 表示概念 A 与树根的最短路径中所包括的边数。

(4)节点的密度：节点的密度是指两个概念最近共同祖先的子节点的密度。本体中不同地方节点的密度是不同的，有的节点可能有上百个子结点,而有的节点可能只有几个子节点.一般来说，某个节点的子节点密度越大,说明细化的概念越具体,这些子节点间的语义相似度也就越小；反之越大，具体可由公式 $den(A,B)=n/m$ 来描述。其中 den (A,B)表示 A 与 B 最近共同祖先的子节点密度,n 表示概念 A,B 最近共同祖先的子节点个数,m 表示概念 A,B 与最近共同祖先所组成的子图的深度。

（5）调节参数：语义相似度是一个主观性相当强的概念,对于不同的应用概念的相似度也不同.调节参数正是根据系统应用的不同来设计的,这里用 a 来表示调节参数 α。

语义相似度的计算公式如下所示：

$$sim(A,B)=\left[\frac{a}{dis(A,B)+a}\right]^{a}\cdot\left[\frac{dep(A)+dep(B)}{|dep(A)-dep(B)|+1}\right]^{b}\cdot\left[\frac{1}{den(A,B)}\right]^{c}$$

其中 $\left[\frac{a}{dis(A,B)+a}\right]$ 表示语义距离对相似度的影响; $\frac{dep(A)+dep(B)}{|dep(A)-dep(B)|+1}$ 表示节点深度对相似度的影响; $\frac{1}{den(A,B)}$ 表示节点密度对相似度的影响;a,b,c 表示语义距离、节点深度、节点密度对语义相似度影响的权重且 a+b+c=1,由于语义距离在相似度计算中占主导地位，而节点密度和节点深度只是起辅助作用，所以 a 的权重相对较大，b，c 的权重相对较小。

3.3 兴趣存储方式

那么我们如何将用户偏好持久化呢？在此我们引入本体来描述用户偏好。

本体是知识的规范化的明确的说明。即本体的目标是提供对该领域知识的共同理解，确定该领域知识中公认的词汇和概念，并从不同层次的形式化模式上给出这些术语和术语之间的关系的唯一的明确的定义，基于概念和本体论的方法就是通过不同组合的概念节点和对应的权值形成 Ontology 来表示用户偏好的模型。通过调整各个节点的权值进而增加减少节点的数量来体现用户的偏好动态。我们可以形象化的表示成为一棵用户兴趣模型树。

为了区分用户的不同兴趣类别，同时考虑到用户兴趣的动态性和局部性，我们将用户的偏好表示成树形结构，简称之用户偏好树。在一般的情况下，用户偏好树只是兴趣分类参考模型的部分映射，因此，用户偏好树具有等级层次的。父节点是对子节点的共同属性的概括，而兴趣子类则是从不同的角度对父类兴趣的细化，如图 4 所示：

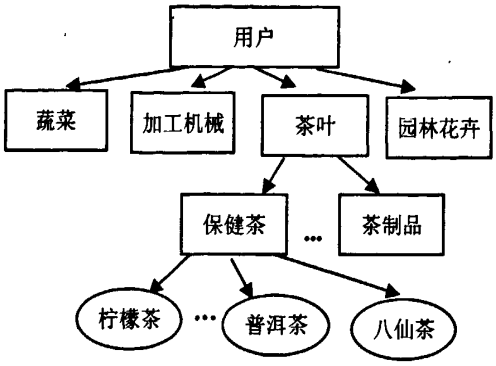

图 4 用户偏好模型树

Fig. 4 the model tree of user profile

农业智能问答原型系统采用 Myeclipse 作为 IDE，其目标是开发一个集问答和搜索于一体的智能检索系统，鉴于试验条件，我们选择 Mysql 作为数据库。

数据库中主要有三类数据表：

（1）基础数据表，包括分类词表 FLC 和用于兴趣分类的参考模型数据表 XQCK。

（2）有关页面分类信息的数据表，主要包括：历史记录表 URLS，特征词表 TZC，页面分类信息表 YMFLXX。

（3）用户偏好信息表包括：长期兴趣树节点信息表 CQXQS，短期兴趣树节点信息表 DQXQS，长期兴趣特征词表 CQXQTZC，短期兴趣特征词表 DQXXTZC，用户兴趣信息表 YHXQXX。

下面是这些表的具体表结构：

表 6 基础信息表：分类词表：

Tab. 6 the table of base information:classified word table

字段名称	数据类型	字段描述
ZTLBHId	INTEGER	主题类别号
TZCBH	VARCHAR(64)	特征词标号

表 7 基础信息表：兴趣参考模型表：

Tab.7 the table of base information:model table of reference for interests

字段名称	数据类型	字段描述
ZTLBHId	INTEGER	主题类别号
CC	INTEGER	类所在的层次描述
LJ	VARCHAR(64)	类所在路径描述
FLBId	INTEGER	父类别号
MC	VARCHAR(64)	类得名称

表 8 分类信息表：历史记录表

Tab.8 the classification information table:history records table

字段名称	数据类型	字段描述
YHId	INTEGER	用户 id
RQ	DATE	用户访问日期
LBHId	INTEGER	所属类别号
ZTCId	INTEGER	主题词编号
SFXL	BOOLEAN	是否新类别

表 9 分类信息表：特征词表

Tab.9 the classification information table:characteristic words table

字段名称	数据类型	字段描述
TZCId	INTEGER	特征词标号
TZC	VARCHAR(64)	特征词

表 10 分类信息表：页面分类信息表

Tab.10 the classification information table: information table of pages classification

字段名称	数据类型	字段描述
URL	VARCHAR(64)	站内连接地址
LBId	VARCHAR(64)	页面所属类别

表 11 用户偏好表：短期兴趣特征词表

Tab.11 User profile table:characteristic words table for short interests

字段名称	数据类型	字段描述
XQLBId	VARCHAR(64)	兴趣类别号
XQZTCId	INTEGER	兴趣特征词编号
QZ	FLOAT	兴趣权值
SJ	DATE	特征词首次出现的时间
YHId	INTEGER	用户的 id

表 12 用户偏好表：长期兴趣特征词表

Tab.12 User profile table: characteristic words table for long interests

字段名称	数据类型	字段描述
XQLBId	VARCHAR(64)	兴趣类别号
XQZTCId	INTEGER	兴趣特征词编号
QZ	FLOAT	兴趣权值
SJ	DATE	特征词首次出现的时间
YHId	INTEGER	用户的 id

表 13 用户偏好表：静态偏好

Tab.13 User profile table: static profile table

字段名称	数据类型	字段描述
DQ	VARCHAR(64)	工作地区
ZY	VARCHAR(64)	职业
ZY	VARCHAR(64)	专业
XL	VARCHAR(64)	学历
SMLX	VARCHAR(64)	色盲类型
YJDS	FLOAT	视力度数
YHId	INTEGER	用户的 id

连接数据库的核心代码如下：

```
<?xml version="1.0" encoding="UTF-8"?>
<beans xmlns="http://www.springframework.org/schema/beans"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.springframework.org/schema/beans
http://www.springframework.org/schema/beans/spring-beans-2.0.xsd">
    <!-- 定义数据源 Bean -->
    <bean id="dataSource"
    class="org.springframework.jdbc.datasource.DriverManagerDataSource">
        <property name="driverClassName">
            <value>com.mysql.jdbc.Driver</value>
        </property>
        <property name="url">
            <value>
                jdbc:mysql://localhost/nongye?characterEncoding=gbk
            </value>
        </property>
        <property name="username">
            <value>root</value>
        </property>
        <property name="password">
            <value>root</value>
        </property>
    </bean>
    <!-- 定义 SessionFactory -->
    <bean id="sessionFactory"
    class="org.springframework.orm.hibernate3.LocalSessionFactoryBean">
        <property name="dataSource">
```

```
            <ref bean="dataSource" />
        </property>
        <property name="hibernateProperties">
            <props>
                <prop key="hibernate.dialect">
                    org.hibernate.dialect.MySQLDialect
                </prop>
                <prop key="hibernate.show_sql">false</prop>
            </props>
        </property>
        <property name="mappingResources">
            <list>
                <value>net/hncu/pojo/User.hbm.xml</value>
                <value>net/hncu/pojo/Gly_User.hbm.xml</value>
                <value>net/hncu/pojo/Gly.hbm.xml</value>
                <value>net/hncu/pojo/Tiwen.hbm.xml</value>
                <value>net/hncu/pojo/Answers.hbm.xml</value>
                <value>net/hncu/pojo/Synonyms.hbm.xml</value>
                <value>net/hncu/pojo/Message.hbm.xml</value>
                <value>net/hncu/pojo/Messaging.hbm.xml</value>
                <value>net/hncu/pojo/Forbitip.hbm.xml</value>
            </list>
        </property>
    </bean>
    <!-- 定义 hibernateTemplate -->
    <bean id="hibernateTemplate"
        class="org.springframework.orm.hibernate3.HibernateTemplate">
        <property name="sessionFactory">
            <ref bean="sessionFactory" />
        </property>
    </bean>
```

4 基于本体的智能问答系统的实现

4.1 原型系统综述

基于用户偏好的检索方法与普通检索方法的最大区别在于： 对客户端提交的字符串，系统将基于用户偏好模型予以分析，并对核心词进行扩展（包括语义和中英文的扩展等），当然查询结果的匹配以及排序等也都基于其偏好进行。通过学习机制以及推理机制，一方面学习了用户在信息、需求上的偏好，另一方面，也可以对用户需求进行推理、归纳。基于用户兴趣的信息检索模型设计的关键之处有三个方面：一、根据用户的资料和行为为用户创建相对应的偏好模型；二、对用户输入的查询表达式进行查询意图分析并利用本体对关键词进行同义词的扩展；三、对查询结果进行排序并返回给客户端。

随着网络技术的发展和计算机的普及，我国新型农业的发展也开始借助了这一高科技工具，国内外涌现出了很多农业相关网站，如农搜网和神州蔬菜网等百余家。但目前的检索技术大都是基于字符串匹配的关键字检索技术，服务效率低下。为此我们提出了基于本体的农业检索系统模型,主要提供用户问答、问题分类、问题管理、信息增殖、自动应答、知识挖掘、智能检索、强大的后台管理等功能，并且提高了检索的查准率和查全率。

基于本体的农业智能检索系统旨在运用本体技术组织、管理和维护海量的信息资源，为农业相关人士提供高效优质的信息服务。通过普通用户间的问答，提供给涉农人士一个交流互助的平台；通过普通用户和专家的问答，挖掘了专家头脑中的隐性知识；通过对问答知识的分类和筛选，建立一个门类齐全的智能问答库，实现自动应答。鉴于以上各种需求，笔者在导师的帮助下，提出了基于本体的农业检索系统框架，如图 5 所示：

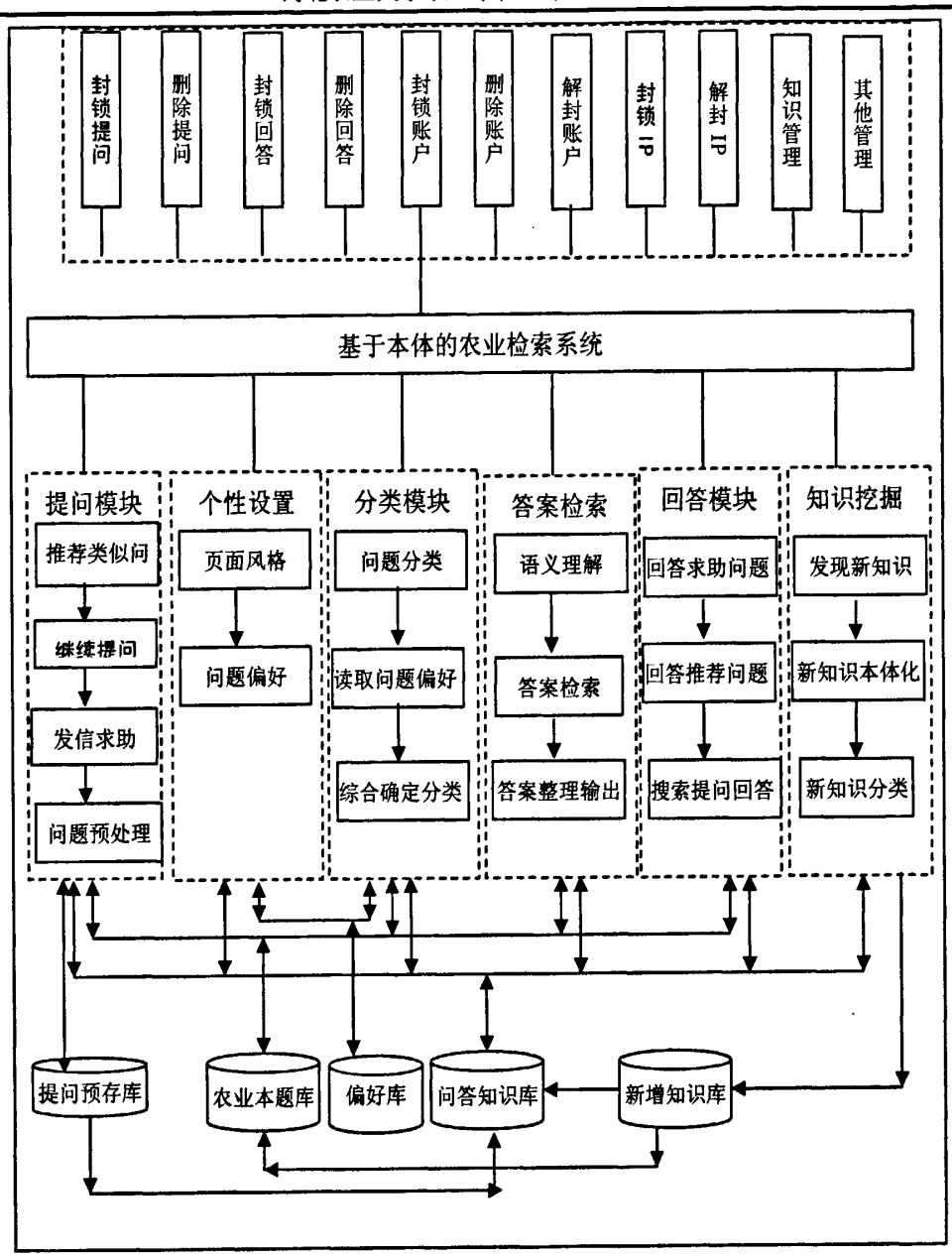

图 5 基于本体的农业检索系统框图

Fig. 5 the block diagram of the system of agricultural retrieval based ontology

提问模块。用户可以通过各种网络设备向系统提问，也可以向某一专家发求助信，提问时可以通过附件提供图像，音频等材料，以使问题更加详细易懂。在客户端，用户可以选择录制视频或者音频来提问，也可以选择以中文的自然语言来提问。对于文化较低的用户，可能文字表达能力弱，导致回答者不能准确理解问题，在此我们提供了音视频录制功能，以帮助这类人群准确的表述提问。对于用自然语言提交的问句，首先系统对问题本体化后通过计算生成提问所属的可能分类，将最相关的分类作为问题的默认分类，用户可以对分类进行调整，完成提问后进行提交，问题提交后保存在提问预存库中。接着，系统通过基于本体的搜索算法在问答知识库里检索相似的问答，

经筛选处理后呈现给用户。用户选择自己满意的答案进行确认，这时系统会自动提示用户他刚才的提问将从提问预存库中删除，视为问题已经解决，经用户确认后删除，悬赏财富将全部返还给用户，此时自动应答过程基本完成。如若用户没有找到满意答案，则问题在用户退出系统后自动存入问答知识库，以待解答。这样避免了重复意义的问答对再次存入知识库，节省了数据空间，在一定程度上提高了系统的运行和检索效率。

在此，分类技术是很关键的，只有正确的分类，才能对用户的偏好进行总结和提取，才能更高效的检索答案，才能更准确的推出问题和答案。目前，常见的基于统计的分类方法取得了良好的使用效果，但是由于机器无法理解词的语义关系,这种分类方法受词的多义、歧义现象的困扰，效率提高受限，分类准确度不高。而本体是一种计算机可理解、可共享、可重用的领域知识的表现方式,其在自动分类、语义导航、智能检索等领域的应用成为研究的热点。在该原型系统中笔者将文本自动分类技术与本体技术相结合，通过提取问题的标题和详细说明中每段首尾句中特征词的方法，对提问做分类，这样使得提问分类更加快捷有效。

4.2 提问模块的实现

提问模块。当用户登录后，我们首先根据用户的静态偏好设置系统的页面风格等。之后，用户可以通过各种网络设备向系统提问，也可以向某一专家发求助信，提问时可以通过附件提供图像，音频等材料，以使问题更加详细易懂。在客户端，用户可以选择录制视频或者音频来提问（由于实验条件所限，这项功能尚未实现），也可以选择以中文的自然语言来提问。对于文化较低的用户，可能文字表达能力弱，导致回答者不能准确理解问题，在此我们提供了音视频录制功能，以帮助这类人群准确的表述提问。对于用自然语言提交的问句，首先系统对问题本体化后结合当前用户的偏好模型计算生成提问所属的可能分类，将最相关的分类作为问题的默认分类，用户可以对分类进行调整，完成提问后进行提交，问题提交后保存在提问预存库中。接着，系统通过基于本体的搜索算法并结合具体用户的偏好模型检索类似的问答，经筛选处理后呈现给用户。用户选择自己满意的答案进行确认，这时系统会自动提示用户他刚才的提问将从提问预存库中删除，视为问题已经解决，经用户确认后删除，悬赏财富将全部按预定的规则分配给相关用户，此时自动应答过程基本完成。如若用户没有找到满意答案，则问题在用户退出系统后自动存入问答知识库，以待解答。这样避免了重复意义的问答对再次存入知识库，节省了数据空间，在一定程度上提高了系统的运行和检索效率，提问页面如图 7 所示。

在此，分类技术是很关键的，只有准确高效的分类，才能更好的更新和利用用户偏好，才能更高效的检索出用户真正需要的答案，才能更准确的推出问题和答案。目前,基于统计的分类方法效果显著，使用颇为广泛,但是由于系统本身很难理解词语之间的语义关系,为此该分类方法受到多义词的困扰,同时在处理词语的歧义方面也是困

难重重，分类效率和准确率遇到难以解决的困境。而本体是一种计算机可理解、可共享、可重用的领域知识的表现方式,其在自动分类、语义导航、智能检索等领域的应用成为研究的热点。在该原型系统中笔者将文本自动分类技术与本体技术相结合，通过提取问题的标题和详细说明中每段首尾句中特征词的方法，并结合用户的偏好模型对提问做分类，这样使得提问分类更加快捷有效。例如张三，其用户名为 zhangsan 的用户，其职业为东北地区的玉米种植散户，目前一段时间内对水果棒子的信息查询较多，这时在其短期兴趣模型中就会出现水果棒子这一小类，包括其特征词汇，现在该用户输入提问“水果棒子种植和传统棒子种植有什么区别？”，系统结合其短期兴趣模型和其静态偏好中的职业项，将预分类确定为大田作物中的种子种苗（如图 7），其分类框架如图 6 所示：

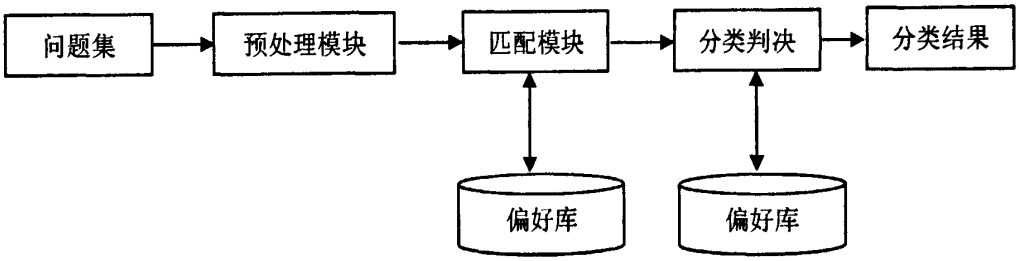

图 6 分类框架图

Fig. 6 classification block diagram

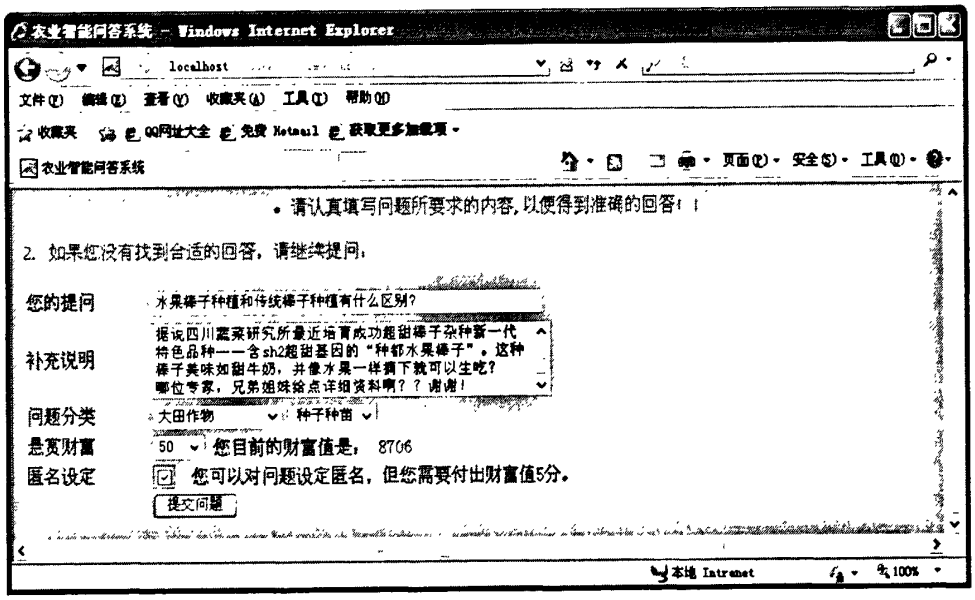

图 7 提问模块页面

Fig. 7 the page module for the questioning

4.3 偏好应用

个性设置。页面风格：用户可以根据自己的喜好，选择页面的背景颜色，字体的

大小、颜色，是否播放背景音乐，是否显示自己的登录名，语言的范围等相关设置，如用户可以选择只显示中文网页，不对关键词进行英文的扩展等，系统登录的同时，会自动加载该用户的静态偏好，如某一用户李四，用户名为 lisi 为大学本科，红绿色盲，视力低下，职业为甲鱼养殖户，则该用户登录后系统会默认为用户用文字表明网页中为红色或者绿色的地方，并为之提供大字体，设置字体加粗等相关项。同时系统会为用户提供最近最热门的点击率最高的关于甲鱼养殖和甲鱼市场等方面用户很可能关心的问题，在这会结合用户所在的地区和用户的动态偏好，通过用户对推荐问题的点击情况，按照预定的规则和算法计算推测用户偏好，并动态更新用户的动态偏好库，步步逼近用户真正需要的信息，流程如下图 8 所示：

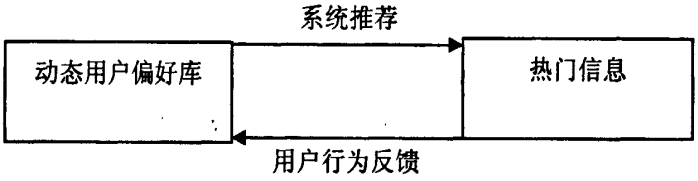

图 8 信息推送/偏好更新框图

Fig. 8 the block diagram for information push and profile renewal

问题偏好：当用户进入系统后，服务器对用户的主要行为进行跟踪和记录，包括用户搜索的关键词，用户确认的满意答案，用户经常来往的专家，用户对某一类问题的访问次数和访问时间等等。这些记录信息经过处理后都以预定格式暂存到数据库中。在用户不断应用本系统的服务时，他的动态偏好也不断的更新，当用户退出系统后，后台程序会对用户的记录进行分析，归纳，并更新到用户的动态兴趣库中。其中用户偏好与本体间结合进行的智能化检索如图 9 所示：

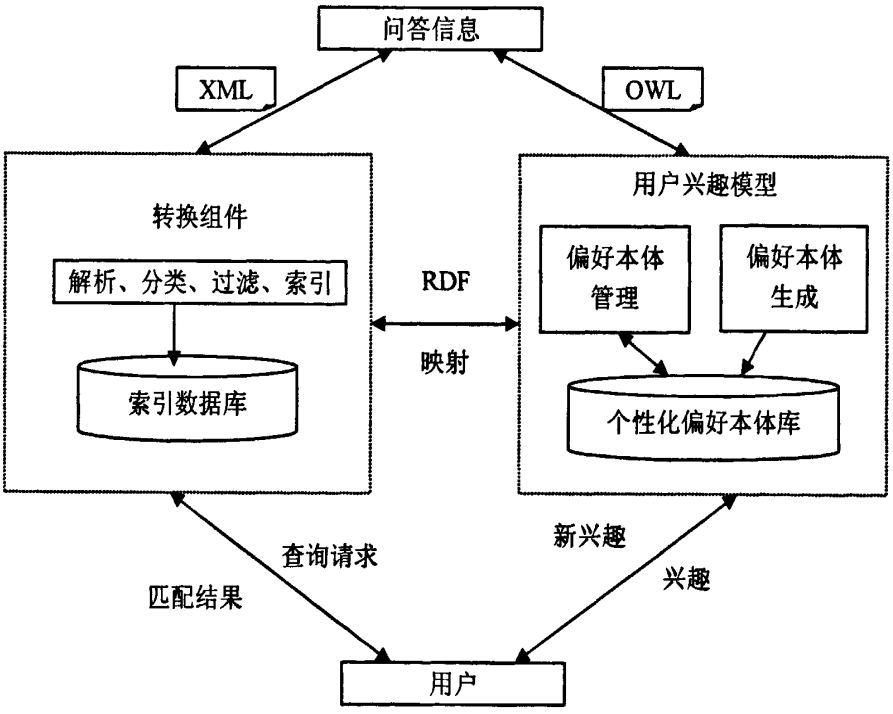

图 9 基于本体的个性化检索模型图

Fig. 9 the model of personality searches based on ontology

4.4 分类模块

分类模块：David Lee 提出[29-32]:文本分类是对给定的一组文档和一个主题集,分类器可以判断每个文档所对应的主题,即判断每个文档属于某一个主题而不是其他的主题。而且只需对文档的部分了解即可做出这样的判断。本系统将传统的文本自动分类技术与本体技术相结合，通过提取文本中每段的首尾句的特征词，并读取问答者的偏好文件，二者结合确定文本的预分类。当然如果用户对分类结果不满，还可以在提交问题时手动改变分类结果，例如用户王二小，登录名为 wangerxiao，是河北张家口阳原县的一养鸡专业大户，那么当他输入“小鸡为什么腿软？”进行提问时，系统首先会结合农业本题库为其提问进行分词。而后会自动根据他的职业结合他的偏好模型，将其问题纳入到“兽医”这一大类，如果用户的目的不是要找兽医，而是想知道小鸡腿软喂什么药，可以手动将分类调整为“兽药”这一大类下，总体流程如图 10 所示。

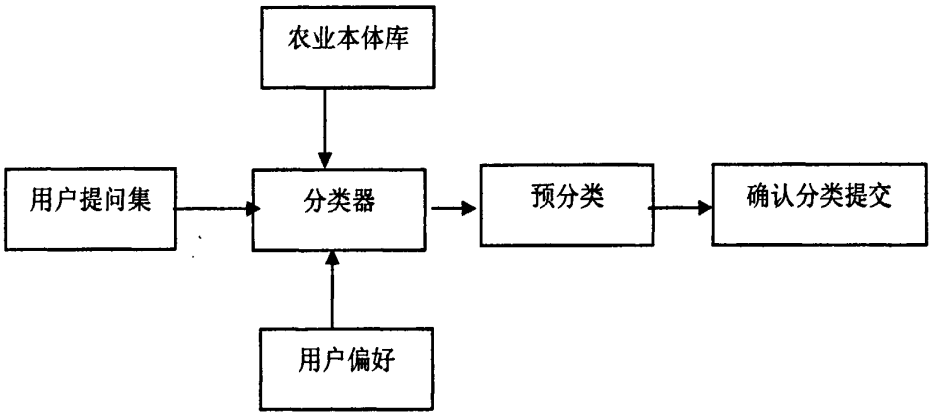

图 10 问题分类流程图

Fig. 10 the flowchart for question classification

如何让计算机高效而准确的理解自然语言？这是一个一直困扰着计算机学者的难题。本体(ontology)作为一种能在语义和知识层次上描述信息系统的概念模型,为解决这一难题提供了新思路，也奠定了坚实的基础,已受到国内外研究人员的广泛关注和议论热点。

由于叙词表和本体在表达知识结构上存在着天然的联系，所以自语义网提出之后，国内外相当一部分专家学者开始利用叙词表来建立本体的尝试，目前据统计已有十多种叙词表被成功的转换为其相应领域的本体，这样的成功都是基于叙词表和本体之间诸多的相似点[33]，比如：

(1)两者都用来描述特定的学科领域知识，可以用作特定学科信息(知识)的组织工具；

(2)两者都包含词(概念、类)及词(概念、类)之间的关系；

(3)两者都具有等级结构，并通过等级关系及词(概念、类)间关系将词(概念、类)组织起来。

通过叙词表来构建领域本体是将对应的叙词表的词汇以及词汇间的关系作为数据源，将叙词表中的叙词作为本体中的类或概念,叙词表中的叙词间的关系作为本体中类与类之间的相关联系。通过以上方法来构建领域本体中的类和类间关系，并由OWL语言来描述和表达本体,最终生成一个OWL文件来存储本体。

联合国粮农组织早在2001年就发起了一个有关农业本体论服务研究的项目,该项目组成员明确了本体在信息资源管理领域一个定义，即"包括一个领域中各类标准术语词汇,并对这些术语词汇进行准确定义,以及明确这些术语间的各种关系。"可见叙词与上述定义是吻合的,首先它是某领域内的词集,再者叙词之间有一定的结构性和相关性,可以说叙词表本身便渗透着本体论的思想和特征，,所以在某种程度上可以把叙词表看作是本体论的简单实现。目前叙词表已有的纯熟技术,对于本体论的实现具有很大程度上的借鉴意义[34]。

那么我们现在已经有农业本体的转换先例，例如；例如联合国粮农组织（FAO）成立了农业本体论服务项目小组（AOS），利用RDFS（RDF Schema）将Agrovoc叙词表转换为农业本体。在此，我们对农业本体的概念作一下简单的说明：在"农业本体论研究与应用"一书中，作者从应用的角度给农业本体下的定义是："农业学科领域中一套得到认同的、关于概念体系的明确、正式的规范说明"（钱平，2006）。

现在我们有了农业本体库，最关键的就是基于本体的分类技术，如何利用本体更好的对问答对进行高效准确的分类，是极为重要的。

如何将中文网页转化成计算机能够识别的模式，这直接影响到分类和检索的准确性，目前向量空间模型（Vector Space Model，VSM）因其较强的可操作性和可计算性被业内人士广泛运用。在该模型中，任何一个中文文档都可以被表示成这样的特征向量V（d）=（t_1，w_1；t_2，w_2；...；t_i，w_i）其中，t_i为特征项，w_i为t_i在文档中的权重。因此，要将文档与空间中的点相对应，首先需要对中文文档进行分词，由这些词作为特征项来表示文本，从而将文本信息的表示与匹配问题转化为向量空间中向量的表示与匹配问题。

在该原型系统中，中文分词工具采用中科院计算所的ICTCLA（汉语语法分词系统），该系统分词精度高，速度快，同时提供了中文分词、新词标注、命名实体标注、词性标注等功能。对于领域本体中出现的新概念可以利用该系统的用户字典功能添加到字典库中，这样便提高分词精度，保证了词库与领域本体中类和概念的同步。本文利用ICTCLA系统提供的接口函数完成对文本的分词处理。

特征选择：若直接使用上述方法得到的结果，可能得到的向量空间高达数万维；另外，很多低频词也会对分类结果产生不良影响，所以对分类文档进行特征选择是十分必要的。特征选择顾名思义就是要将低频的信息含量低的词从特征项空间中去除，从而达到降维的目的，是文本自动分类系统中的提高分类精度和速度的必要步骤。自文档分类出现以来，各种文法算法频频问世，如信息增益（Information Gain， IG）、文档频率（Document Frequency，DF）、开方拟和检验（CHI）、互信息（Mutual Information，MI）等。其中DF算法简单、实现容易、质量高，是目前业内人士首选的分类方法。每类训练集中的网页数目可能有一定的差异，所以在实验中我们采用相

对 DF 阈值的方法，将小于该阈值的那些特征词删除掉，当然 DF 阀值我们将通过实验进行合理选取。

生成特征向量：如何计算特征项在其相对应文档中的权重值 w_i 对网页分类起着至关重要的作用，目前最热门适用的是基于 TF-IDF 的传统特征权重算法。其计算公式如下：

$$w_i = \frac{tf_i * \log(N/n_k + 0.01)}{\sqrt{\sum_{i=1}^{n} tf_i^{\ 2} * [\log(N/n_k + 0.01)^2]}}$$

其中，tf_i 为 t_i 在文档 d 中出现的频率，N 为文档集中的总文档数，n_k 为出现特征项 t_k 的文档数。

预分类：预分类过程描述如下：

（1）对网页进行预处理，包括对相关文档进行中文分词并去除其中的停用词；

（2）提取文档标题内容和问答对中的首尾句，与系统中预置关键词表比较，确定单词所属类别，统计关键位置中单词在各类别中的出现频度；

（3）若属于某个类别的单词频度最大，则认为文档属于该类别；

（4）若属于两个类别的单词频度相等，则比较类别优先级，将其划分为优先级较大的类别，若优先级相同，则使用 SVM 分类器；

（5）其他情形都使用 SVM 分类器。

SVM 分类器：本系统选用了台湾大学林智仁(Chih-Jen Lin)博士等开发设计 LIBSVM 开源工具，它是一个易于使用、操作简单、快速有效的通用 SVM 软件包，可以解决分类问题（包括 C- SVC、n - SVC ）、分布估计（one-class-SVM ）以及回归问题（包括 e - SVR、n - SVR ）等问题，提供了径向基、S 形函数、线性和多项式四种常用的核函数供选择，可以有效地解决多类问题、对不平衡样本加权、交叉验证选择参数、多类问题的概率估计等。

LIBSVM 是一个免费的开源的软件包，其下载地址为 http://www.csie.ntu.edu.tw/~cjlin/。他不仅提供了 C++版的源代码，同时还提供了 MATLAB、Ruby、LabVIEW、Python、C#.net、Java、R 以及 Perl 等各种常见语言的接口，并且在 UNIX 或 Windows 平台下均可使用。另外还提供了 WINDOWS 平台下的可视化操作工具 SVM-toy。

LibSVM 在提供源代码的同时还给出了其可执行文件。如果是在 Windows 系统下使用，既可以选择直接使用软件包提供的程序，也可以进行修改后重新编译；如果是在 Unix 平台下应用的话，就必须要求使用者自己编译。 LIBSVM 在给出源代码的同时还提供了 Windows 操作系统下的可执行文件，包括：进行支持向量机训练的 svmtrain.exe；根据已获得的支持向量机模型对数据集进行预测的 svmpredict.exe；以及对训练数据与测试数据进行简单缩放操作的 svmscale.exe。它们都可以直接在 DOS 环境中使用。如果下载的包中只有 C++的源代码，则也可以自己在 VC 等软件上编译生成可执行文件。

LIBSVM 使用的一般步骤是：

1） 按照 LIBSVM 软件包所要求的格式准备数据集；

2） 对数据进行简单的缩放操作；

3） 考虑选用 RBF 核函数；

4） 采用交叉验证选择最佳参数 C 与 g ；

5） 采用最佳参数 C 与 g 对整个训练集进行训练获取支持向量机模型；

6） 利用获取的模型进行测试与预测。

4.5 答案检索

答案检索：目前基于关键词匹配的检索系统只能从字面上来匹配问答，每个提问者与信息提供者对同一概念的表示形式有所不同,故其检索到的答案中有很多冗余或者毫无用途的数据，造成误检和漏检。如果能在检索过程中系统本身可以“理解“提问，即将基于关键词层面的检索过渡到基于语义层面的知识检索,这势必大大提高系统的查准率和检索效率。基于本体的信息检索不同于传统检索方式，前者是在知识（概念）层面上进行的知识检索过程，考虑了概念之间的内在联系，可以发掘出那些不明确的或者隐含的信息和概念，不仅大大提高了信息的查全率和查准率，还使系统具备了智能化、查找定位准确等优点。笔者在学习和查阅了大量文献资料后，将本体检索技术引入了该原型系统，其检索步骤可大致概括为：1.系统接收到查询请求并对其预处理后（包括对提问进行分词处理，剔除停用词，同义词的扩展等）,利用农业本体库对其进行语义推理,并结合用户偏好本体库在问答知识库中查找合适的答案，例如用户甲是河北张家口某县的一玉米种植户，提问时输入“玉米叶黄”，那么系统将该提问分词为“玉米”，“叶黄”，同时对玉米一词进行扩展为“棒子”，“榆树棒”，当然如果该用户的学历为初中以上还会对其进行英文的扩展“corn”，而后在问答库检索玉米叶黄的相关问答。2. 对答案按照相关度和时间进行排序，返回给用户，例如系统检索到了许多关于玉米叶黄的结果，那么系统会根据答案相关度按照提问时间从现在到以前，先本地区再外地区的原则进行排序。如没有找到合适答案，则返回关键词的上下位词和相关相似词供用户选择 3。 用户找到满意答案后进行确认，否则系统将记录用户问题，直到用户得到满意回答为止。

为使答案更准确、更全面，有必要对问题进行语义理解。当提问者通过查询接口用自然语言方式提出问题时，系统首先通过语义理解模块对提出的问题进行分词处理，解析出关键词，通过一定的算法利用农业本题库对关键词进行同义词的扩展，中英文的对照，简称的对照。比如我们搜索大豆，进行语义理解后，可以同时对大豆，黄豆，青豆，soybean,soy,soya 进行检索，从而对问题做到了一定程度上的语义理解。此后，在问答知识库中按一定的相似度算法在所属的分类中检索出大于某一阀值的问答对，答案经整理和排序后呈现给用户。如果没有检索到相关的答案，可以返回主关键词的上位词、下位词、相似相关词，用户点击相关的词条后可以作为新的提问提交

给系统，再一次进行答案检索。比如若没有大豆的相关内容，系统将返回它的上位词豆类作物，下位词毛豆，相关词豆油，豆面，豆芽菜等等，其问题回答流程如图 11 所示。

```
    // 将提交的查询字符串传给 search 方法的字符串数组 cx，然后返回标记后的查询结果 result
        public String Search(String search, String wenben) throws Exception {
            if (search.length() != 0) {
                QUERY = search;
            }
            String result = null;
            // 将庖丁封装成符合 Lucene 要求的 Analyzer 规范
            Analyzer analyzer = new PaodingAnalyzer();
            // System.out.println("所检索文本为：" + wenben);
            // 读取本类目录下的 text.txt 文件
            String content = wenben;
            // 接下来是标准的 Lucene 建立索引和检索的代码
            Directory ramDir = new RAMDirectory();
            IndexWriter writer = new IndexWriter(ramDir, analyzer);
            Document doc = new Document();
            Field fd = new Field(FIELD_NAME, content, Field.Store.YES,
                    Field.Index.TOKENIZED,
Field.TermVector.WITH_POSITIONS_OFFSETS);
            doc.add(fd);
            writer.addDocument(doc);
            writer.optimize();
            writer.close();
            IndexReader reader = IndexReader.open(ramDir);
            String queryString = QUERY;
            QueryParser parser = new QueryParser(FIELD_NAME, analyzer);
            Query query = parser.parse(queryString);
            Searcher searcher = new IndexSearcher(ramDir);
            query = query.rewrite(reader);
            // System.out.println("Searching for: " + query.toString(FIELD_NAME));
            Hits hits = searcher.search(query);
            BoldFormatter formatter = new BoldFormatter();
            Highlighter highlighter = new Highlighter(formatter, new QueryScorer(
                    query));
            highlighter.setTextFragmenter(new SimpleFragmenter(50));
```

```
        for (int i = 0; i < hits.length(); i++) {
            String text = hits.doc(i).get(FIELD_NAME);
            int maxNumFragmentsRequired = 5;
            String fragmentSeparator = "...";
            TermPositionVector tpv = (TermPositionVector) reader
                    .getTermFreqVector(hits.id(i), FIELD_NAME);
            TokenStream tokenStream = TokenSources.getTokenStream(tpv);
            result = highlighter.getBestFragments(tokenStream, text,
                    maxNumFragmentsRequired, fragmentSeparator);
            // System.out.println("\n" + result);
        }
        reader.close();
        return result;
    }
    public String answers() throws Exception {
        if (search == null || search.trim().length() <= 0) {
            addActionError("请输入检索词");
            return "kong_search";
        } else {
            int flag1 = 0;
            int flag2 = 0;
            List showAll = new ArrayList();
            /** 查询用户偏好库，根据用户偏好对检索词进行处理* */
            search = MyFavorite(search);
            /** *对检索词进一步进行本体化处理** */
            search = Ontology(search);
            List all = tiwenService.queryAllSolving();
            for (int i = 0; i < all.size(); i++) {
                tiwen = (Tiwen) all.get(i);
                wenben = tiwen.getWenti();
                String biaoti = Search(search, wenben);
                if (biaoti != null) {
                    flag1 = 1;
                    if (biaoti.length() > 52) {
                        biaoti = biaoti.substring(0, 50);
                        biaoti = tiwenService.Biaoti(biaoti);
                    }
                    tiwen.setWenti(biaoti);
```

```
            }
            wenben = tiwen.getBucun();
            String bucun = Search(search, wenben);
            if (bucun != null) {
                flag2 = 1;
                if (bucun.length() > 112) {
                    bucun = bucun.substring(0, 110);
                    bucun = tiwenService.Bucun(bucun);
                }
                tiwen.setBucun(bucun);
            }
            if (flag1 == 1 && bucun == null) {
                String bucun1 = tiwen.getBucun();
                if (bucun1.length() > 112) {
                    bucun1 = bucun1.substring(0, 110) + "...";
                }
                tiwen.setBucun(bucun1);
            }
            if (flag1 == 1 || flag2 == 1) {
                if (tiwen.getNiming() == 1) {
                    tiwen.setBz2("匿名");
                } else {
                    int yhid = tiwen.getYhid();
                    User user = userService.queryUserByID(yhid);
                    if (user == null)
                        user = userService.queryUserByID(0);
                    String username = user.getUsername();
                    tiwen.setBz2(username);
                }
                showAll.add(tiwen);
                flag1 = 0;
                flag2 = 0;
            }
        }
        // System.out.println("查询执行");
        ActionContext.getContext().getSession().put("pageshowAll", showAll);
    }
    // 重新初始化一下查询几个，得到已经查好的数据
```

```
        List showAll = (List) ActionContext.getContext().getSession().get(
                "pageshowAll");
        if (showAll.size() > 0) {
            int rowsize = showAll.size();
            int i = (page - 1) * 3;
            this.pageBean = tiwenService.queryForPage(3, page, rowsize,
                    showAll, i);
            List show = tiwenService.queryForPage(3, page, rowsize, showAll, i)
                    .getList();
            // System.out.println("********page:" + page);
            ServletActionContext.getRequest().setAttribute("showAll", show);
            return "answers";
        } else {
            addActionError("抱歉！没有搜到相匹配的结果，请您修改检索词！");
            /*****没找到答案提供检索词的上下位词******/
            search=UpDownWord(search);
            ActionContext.getContext().getSession().put("search", search);
            return "noresult";

        }
    }
```

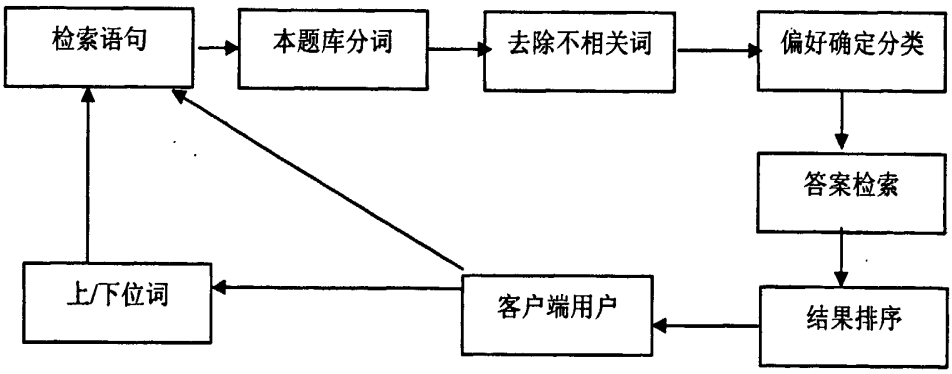

图 11 答案检索流程图

Fig. 11 the flowchart for searching answers

4.6 回答模块

回答模块：用户登录后，个人页面会为之呈现收件箱里待回答的求助问题，进入收件箱，用户点击相应的邮件后可以查看问题详情，包括问题的详细说明，悬赏财富，附件下载，提问时间等等，专家可以点击回复来回答提问，也可以转给其他的专家回答。在用户空间页面，还会为之呈现系统根据他的个人偏好，专家类型为之推荐的提

问，专家或者普通用户可以根据自己的兴趣来选择回答其中的一部分问题来挣得积分。当然用户也可以根据自己的兴趣，搜索待回答的问题进行回答，以增加自己的财富，提高自己的等级，例如用户王明，其登录名为 wangming，是河北保定市涿州的一名棉花种植专业户，系统会自动为之呈现出该阶段涿州市甚至保定市棉花种植户遇到的种植问题，而另一位用户李晓明，登录名为 lixiaoming 的用户是涿州市的一位棉花商户，则系统会为之自动呈现棉花买卖的问题，其整体回答模块流程图如图 12 所示：

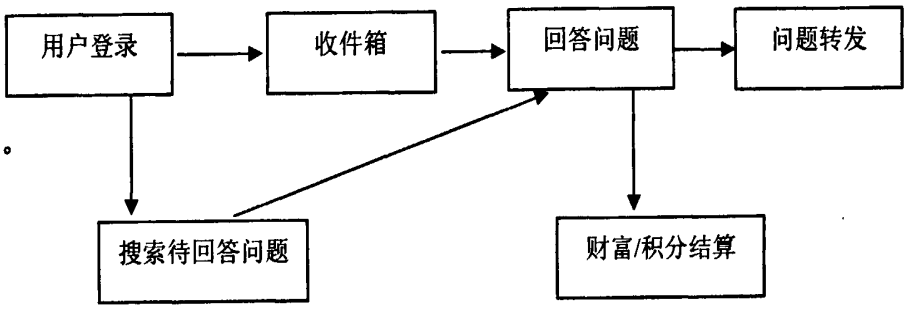

图 12 回答模块整体流程框图

Fig. 12 the holistic flowchart for answering model

用户可以通过检索来回答其他用户问题（但不能自问自答，也不能多次回答同一个用户的同一个问题）来帮助别人以挣得积分，提高自己的等级，例如一个玉米种植户 moon 为了挣得积分，简单提交“玉米“作为检索词，系统会首先读取他的用户偏好，根据其兴趣模型得知该用户是一位视力正常、初中学历、东北地区的玉米种植大户等信息，根据得到的信息设置页面风格为字体正常，背景默认，字体颜色正常显示等页面风格，根据其为玉米种植户的特点，结合其动态偏好，为其检索出玉米种植方面的信息，滤掉了玉米买卖和加工等方面的干扰信息，提高了查准率，如图 13 所示，此后检索者可以对问题进行选择性回答。做到知识的共享，如图 14 所示。

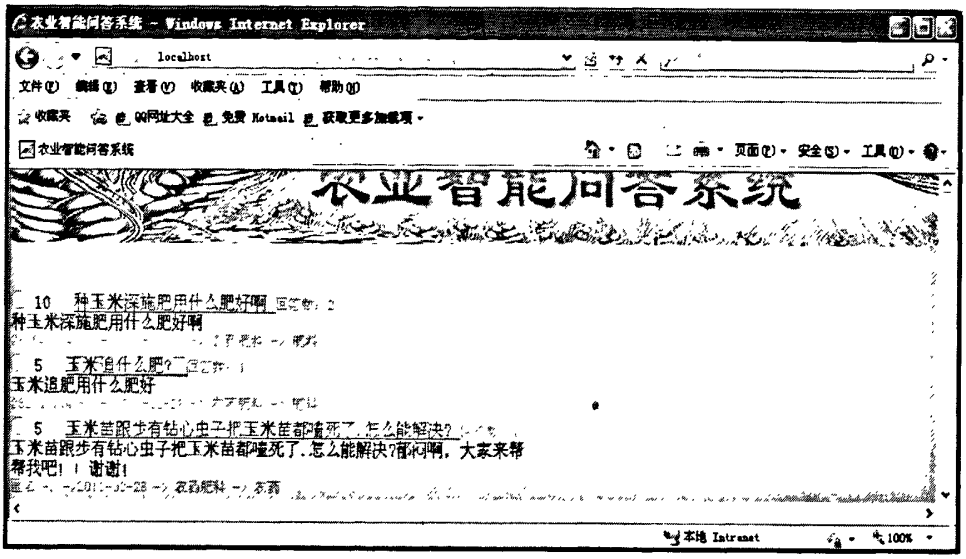

图 13 检索问题页面

Fig. 13 the page for searching questions

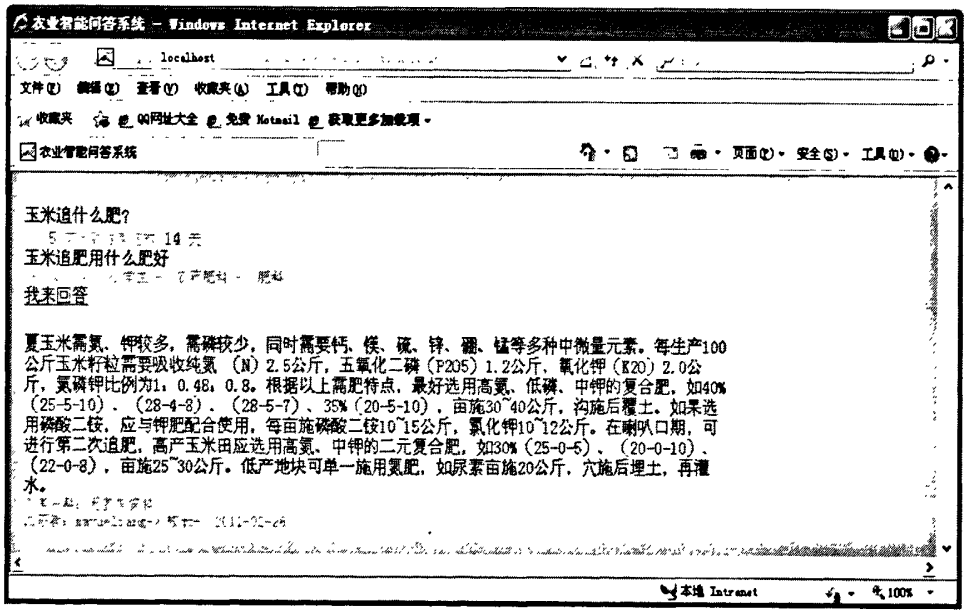

图 14 回答页面

Fig. 14 the page for answering

用户还可以回答被其他用户转发和推荐到自己邮箱的问题,其邮箱中用户可以看到其他用户转发或者直接发送过来的通知和求助信,其效果页面如图 15 所示,当然邮箱所有者可以选择性的回答一些问题,也可以不作答或全部回答,该专家可以查看问题的详细描述并下载附件查看后对问题进行作答,系统会根据用户回答的问题和拒绝回答的问题,推断出用户的当前兴趣,例如某一用户经常和一些玉米种植户联系,并且经常回答玉米价格和玉米买卖的的问题,那么系统可以将这些信息记录挖掘出来,由此来动态更新用户的偏好模型。

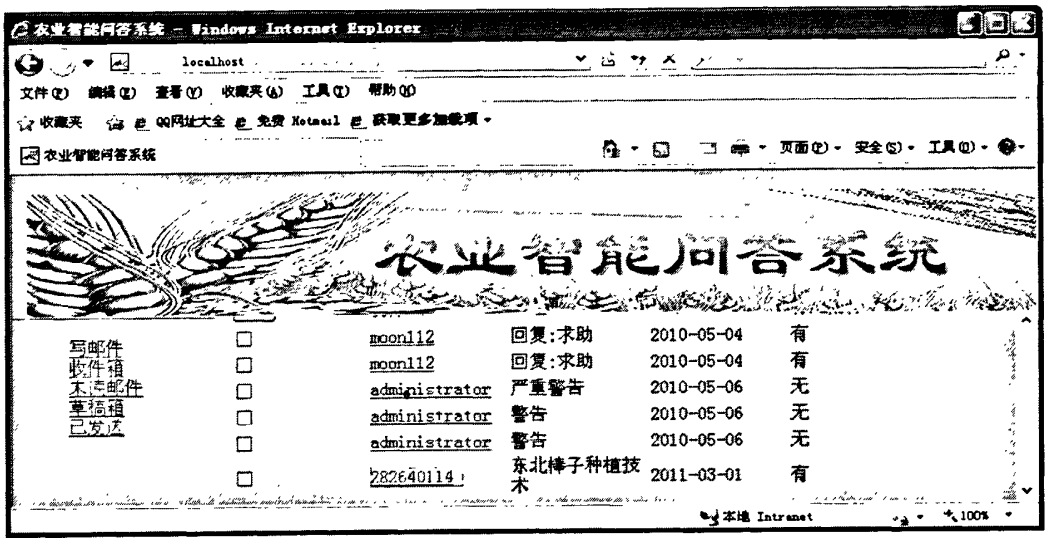

图 15 用户邮件箱

Fig. 15 the user mailbox

4.7 知识挖掘

知识挖掘：隐性知识被喻为重要的无形资产，是个人和组织保持持续发展和强大竞争力不可短少的动力源泉[35-39]。用户和专家，尤其是退休的老专家，通过回答别人的问题，把其头脑中的隐形知识经过组合形成更复杂、更系统的知识显性的表达出来，为别人解决问题。这在一定程度上解决了专家头脑中隐形知识的沉淀和共享问题。当某个问题第一次被提出时，系统会对其加以跟踪，解决后对其进行本体化处理，分类处理后加入新增知识库，由专家决定是否加入问答知识库。此外，我们还可以利用数据挖掘技术发现有用的信息和规律，供专家研究和指导涉农人士的生产，其知识挖掘的详细流程如图 16 所示。

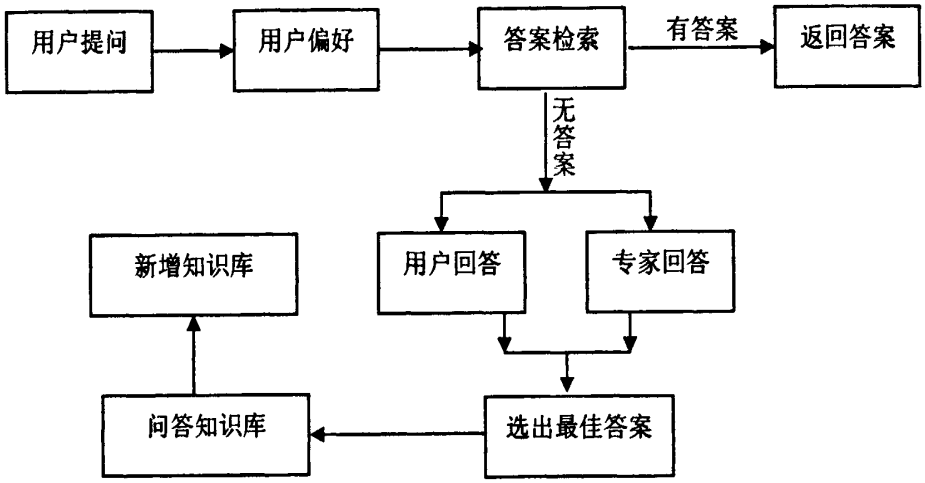

图 16 隐形知识挖掘流程框图

Fig. 16 the flowchart for mining invisible knowleage

4.8 高级检索

系统除了为用户提供一般的检索功能，还为某些特殊用户提供了高级检索功能，其目的是给用户提供更加准确和详细的检索功能，其核心源码所示：

```
public String Search(String search, String wenben) throws Exception {
    if (search.length() != 0) {
        QUERY = search;
    }
    String result = null;
    // 将庖丁封装成符合 Lucene 要求的 Analyzer 规范
    Analyzer analyzer = new PaodingAnalyzer();
    // System.out.println("所检索文本为：" + wenben);
    // 读取本类目录下的 text.txt 文件
    String content = wenben;
    // 接下来是标准的 Lucene 建立索引和检索的代码
    Directory ramDir = new RAMDirectory();
    IndexWriter writer = new IndexWriter(ramDir, analyzer);
    Document doc = new Document();
    Field fd = new Field(FIELD_NAME, content, Field.Store.YES,
            Field.Index.TOKENIZED,
Field.TermVector.WITH_POSITIONS_OFFSETS);
    doc.add(fd);
    writer.addDocument(doc);
    writer.optimize();
    writer.close();
    IndexReader reader = IndexReader.open(ramDir);
    String queryString = QUERY;
    QueryParser parser = new QueryParser(FIELD_NAME, analyzer);
    Query query = parser.parse(queryString);
    Searcher searcher = new IndexSearcher(ramDir);
    query = query.rewrite(reader);
    // System.out.println("Searching for: " + query.toString(FIELD_NAME));
    Hits hits = searcher.search(query);
    BoldFormatter formatter = new BoldFormatter();
    Highlighter highlighter = new Highlighter(formatter, new QueryScorer(
            query));
    highlighter.setTextFragmenter(new SimpleFragmenter(50));
```

```
        for (int i = 0; i < hits.length(); i++) {
            String text = hits.doc(i).get(FIELD_NAME);
            int maxNumFragmentsRequired = 5;
            String fragmentSeparator = "...";
            TermPositionVector tpv = (TermPositionVector) reader
                    .getTermFreqVector(hits.id(i), FIELD_NAME);
            TokenStream tokenStream = TokenSources.getTokenStream(tpv);
            result = highlighter.getBestFragments(tokenStream, text,
                    maxNumFragmentsRequired, fragmentSeparator);
            // System.out.println("\n" + result);
        }
        reader.close();
        return result;
    }
    public String answers() throws Exception {
        if (search == null || search.trim().length() <= 0 || key1 == null
                || key1.trim().length() <= 0) {
            addActionError("请输入检索语句和主关键词");
            return "kong_search";
        }  else {
List sy = synonymsService.queryAllSynonyms();
            // 对主关键词进行扩展
            if (key1 != null && key1.trim().length() > 0
                    && key2.trim().length() <= 0) {
                search = synonymsService.Synonums(search, key1, sy);
        } else if (key1 != null && key1.trim().length() > 0 && key2 != null
                    && key2.trim().length() > 0) {
                search = synonymsService.Synonums(search, key1, key2, sy);
            } else if (key1.trim().length() <= 0 && key2 != null
                    && key2.trim().length() > 0) {
                addActionError("请按规则出牌，您的主关键字还没有填写！ ");
                return "norule";
            }
            int flag1 = 0;
            int flag2 = 0;
            List showAll = new ArrayList();
            /** 查询用户偏好库，根据用户偏好对检索词进行处理* */
            search = MyFavorite(search);
```

```
/** *对检索词进一步进行本体化处理** */
search = Ontology(search);
List all = tiwenService.queryAllSolving();
for (int i = 0; i < all.size(); i++) {
    tiwen = (Tiwen) all.get(i);
    wenben = tiwen.getWenti();
    String biaoti = Search(search, wenben);
    if (biaoti != null) {
        flag1 = 1;
        if (biaoti.length() > 52) {
            biaoti = biaoti.substring(0, 50);
            biaoti = tiwenService.Biaoti(biaoti);
        }
        tiwen.setWenti(biaoti);
    }
    wenben = tiwen.getBucun();
    String bucun = Search(search, wenben);
    if (bucun != null) {
        flag2 = 1;
        if (bucun.length() > 112) {
            bucun = bucun.substring(0, 110);
            bucun = tiwenService.Bucun(bucun);
        }
        tiwen.setBucun(bucun);
    }
    if (flag1 == 1 && bucun == null) {
        String bucun1 = tiwen.getBucun();
        if (bucun1.length() > 112) {
            bucun1 = bucun1.substring(0, 110) + "...";
        }
        tiwen.setBucun(bucun1);
    }
    if (flag1 == 1 || flag2 == 1) {
        if (tiwen.getNiming() == 1) {
            tiwen.setBz2("匿名");
        } else {
            int yhid = tiwen.getYhid();
            User user = userService.queryUserByID(yhid);
```

```
                    if (user == null)
                        user = userService.queryUserByID(0);
                    String username = user.getUsername();
                    tiwen.setBz2(username);
                }
                showAll.add(tiwen);
                flag1 = 0;
                flag2 = 0;
            }
        }
        // System.out.println("查询执行");
        ActionContext.getContext().getSession().put("pageshowAll", showAll);
    }
    // 重新初始化一下查询几个，得到已经查好的数据
    List showAll = (List) ActionContext.getContext().getSession().get(
            "pageshowAll");
    if (showAll.size() > 0) {
        int rowsize = showAll.size();
        int i = (page - 1) * 3;
        this.pageBean = tiwenService.queryForPage(3, page, rowsize,
                showAll, i);
        List show = tiwenService.queryForPage(3, page, rowsize, showAll, i)
                .getList();
        // System.out.println("********page:" + page);
        ServletActionContext.getRequest().setAttribute("showAll", show);
        return "answers";
    } else {
        addActionError("抱歉！没有搜到相匹配的结果，请您修改检索词！");
        /*****没找到答案提供检索词的上下位词******/
        search=UpDownWord(search);
        ActionContext.getContext().getSession().put("search", search);
        return "noresult";

    }
}
```

例如用户种植户 moon123 查询玉米的相关信息，通过查询其偏好模型确定其为大学本科，视力正常的玉米种植大户，则可以为其设置预订的页面风格，并根据用户需求对玉米这个关键词进行同义词和中英文的扩展，系统根据扩展后的中心词结合用

户偏好为该用户检索到了仅与玉米种植相关的问题，过滤掉了玉米加工和玉米交易等其他用户不关心的冗余信息，这样在保证查全率的情况下，提高了系统的查准率，其具体操作界面如图 17 所示：

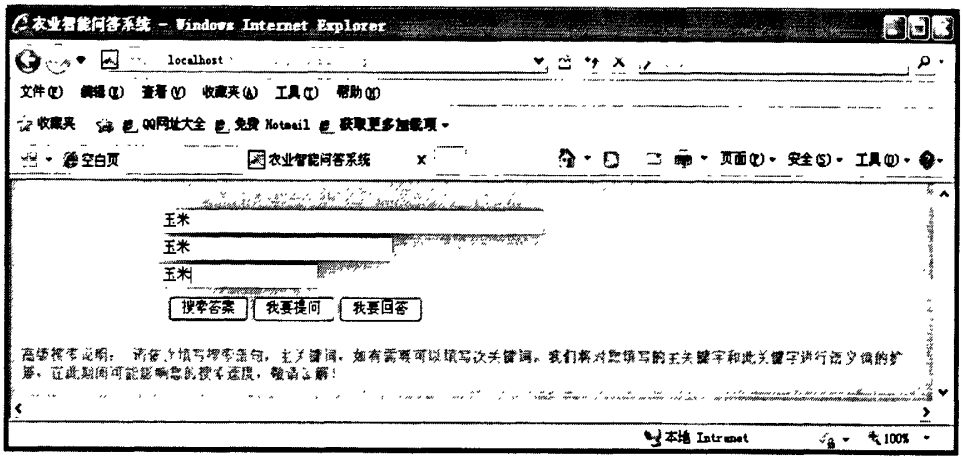

图 17 高级检索界面

Fig. 17 the page for advanced searching

检索到结果如图 18 所示，可见系统对关键词进行了同义词扩展的同时，也进行了中英文的语义拓展，并且查询结果只与玉米种植相关，过滤掉了玉米加工和玉米买卖等其他方面的冗余信息。系统提高了查准率的同时，也提高了查全率。

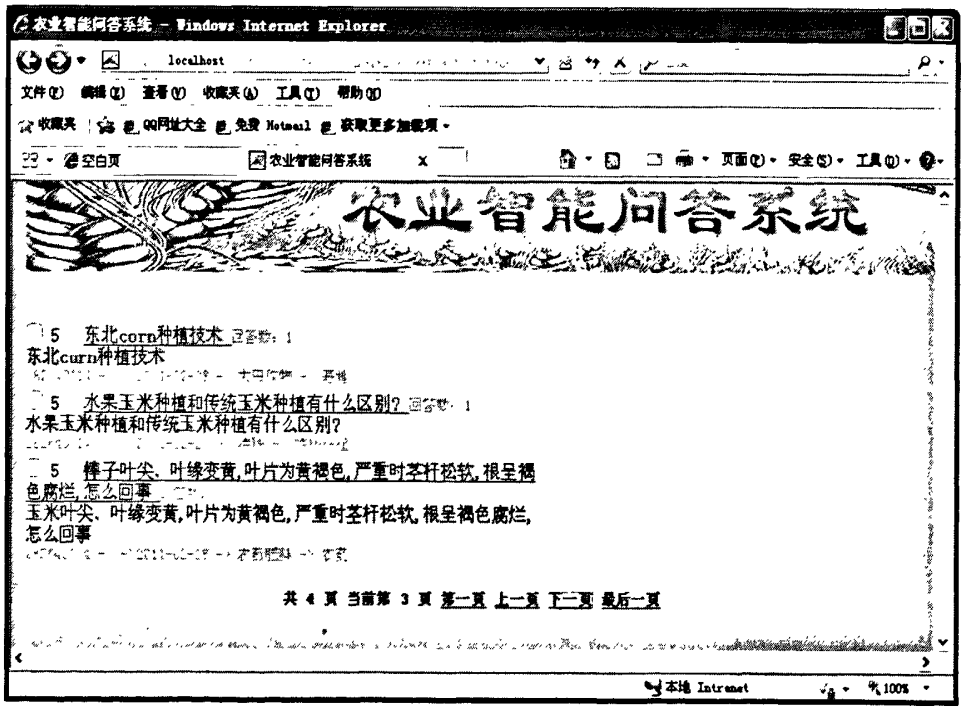

图 18 高级搜索结果集图

Fig. 18 the results of advanced searching

4.9 管理员后台管理

管理员后台：在网络和计算机飞速发展和普及的今天，不免有些素质不高的人在网络上发布一些不健康的信息。为此我们在后台提供了封锁问题，删除问题，封锁答案，删除答案，封锁用户，解封用户，拉黑 IP 等针对这些不健康信息和信息源的处理。此外，还提供了管理员对知识和其他相关问题的管理，以保证系统健康良好稳定的运行，为涉农人士提供更优质的服务。

封锁 IP 的核心实现代码如下：

```
public String execute() throws Exception {
        Date d = new Date();
        SimpleDateFormat simpledateformat= new
SimpleDateFormat("yyyy-MM-dd");
        String time = simpledateformat.format(d);
        if (reason == null || reason.trim().length() <= 0)
            reason = "您的 IP 由于多次发布非法或不健康信息，已被管理员查封！
";
        if (IP == null || IP.trim().length() <= 0) {
            addActionError("请输入要封锁的 IP！ ");
            return "kong";
        }
        Forbitip forbitip = new Forbitip();

        forbitip = forbitipService.queryByIp(IP);
        if (forbitip != null && forbitip.getState() == 1) {
            addActionError("此 IP： " + IP + "正处于封锁状态，请勿再次封锁");
            return "forbiting";
        }
        if (forbitip != null && forbitip.getState() != 1) {
            int num = forbitip.getNum() + 1;
            forbitip.setNum(num);
            forbitip.setState(1);
            forbitip.setDate(time);
            forbitipService.update(forbitip);
            addActionError("该 IP" + IP + "已经是" + num + "次被封锁！ ");
            return "forbited";
        }
        Forbitip forbitip1 = new Forbitip();
```

```
        forbitip1.setIp(IP);
        forbitip1.setReason(reason);
        forbitip1.setDate(time);
        forbitip1.setNum(1);
        forbitip1.setState(1);
        if (forbitipService.addForbitip(forbitip1)) {
            addActionError("成功封锁 IP：" + IP);
            return SUCCESS;
        } else {
            addActionError("封锁 IP" + IP + "失败，请重试！");
            return ERROR;

        }
    }
```

当系统中出现了不良提问和回答时，管理员既可以对整个问题进行查封，也可以针对问题中的某些答案进行删除，这样不但保证了系统健康稳定的运行，同时也增加了系统灵活性，其查封问题界面如图 19 所示，当问题被成功封锁后，我们可以在查封问题集中看到，也可以进一步进行删除操作。

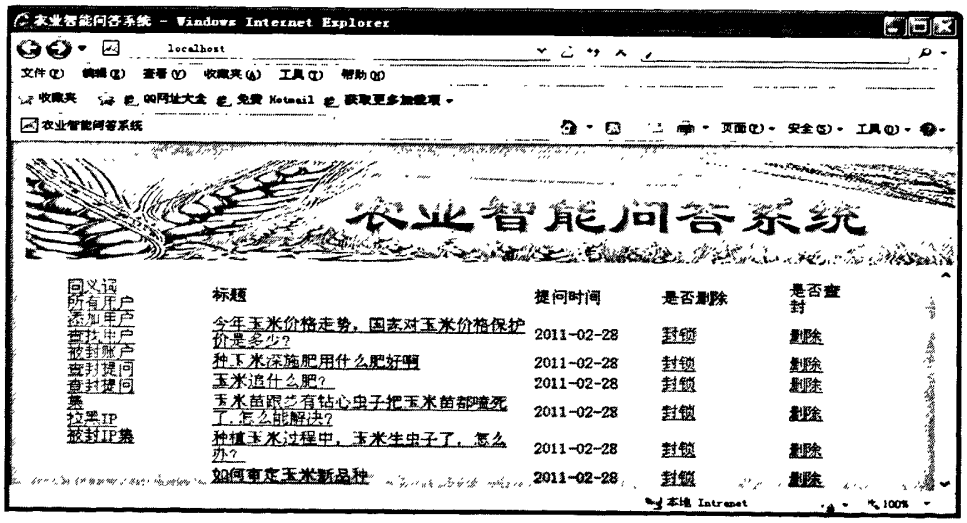

图 19 管理员查封问题页面

Fig. 19 the page of sealing up the questions by adminstrator

对于在系统中以提问为名传播不良信息，或者在别人问题中跟帖发布非法内容的用户，管理员也可以对其账号进行封锁，严重时可以删除该用户，当然也可以对用户资料进行编辑等等，其界面如图 20 所示：

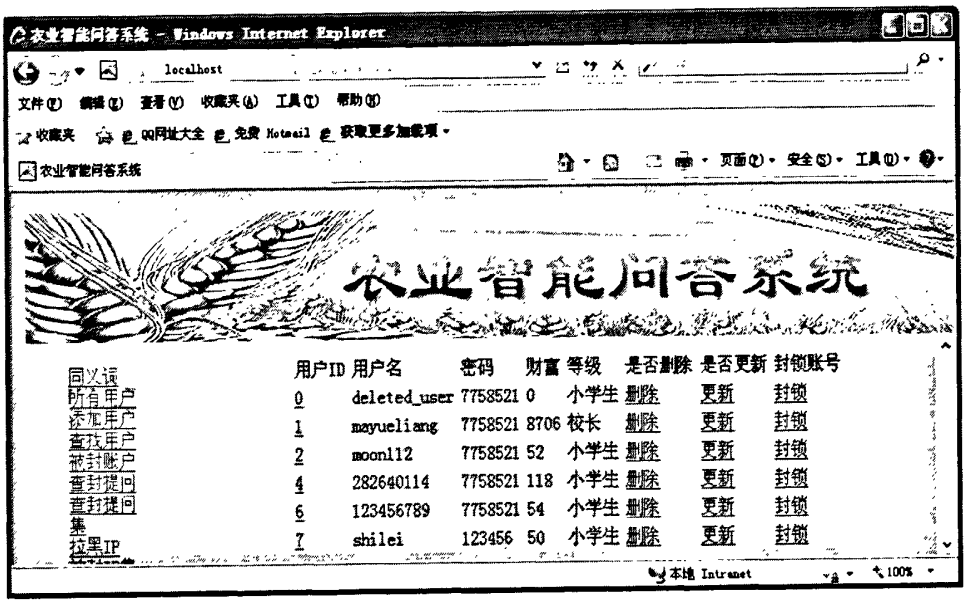

图 20 管理员对用户的操作页面

Fig. 20 the page for adminstrator operating users

5 农业智能问答系统测试

5.1 实验环境

鉴于系统特性，我们选择本实验在 Java 技术平台上进行开发，具体环境和工具为：开发语言选择 JAVA5,开发工具为 MyEclipse 6.5；使用 Struts2 框架实现 MVC 分层；采用 MYSQL 数据库和 Hibernate 等数据库技术实现数据高效存取。Java 编译和运行环境 jdk1.6.0_06, Web 服务器软件 Tomcat6.0， 本体编辑工具软件选择 Protégé3.3。具体使用的技术还包括 jsp，Spring，ext 等等，这里将不一一再做介绍。IDE 开发如图 21 示：

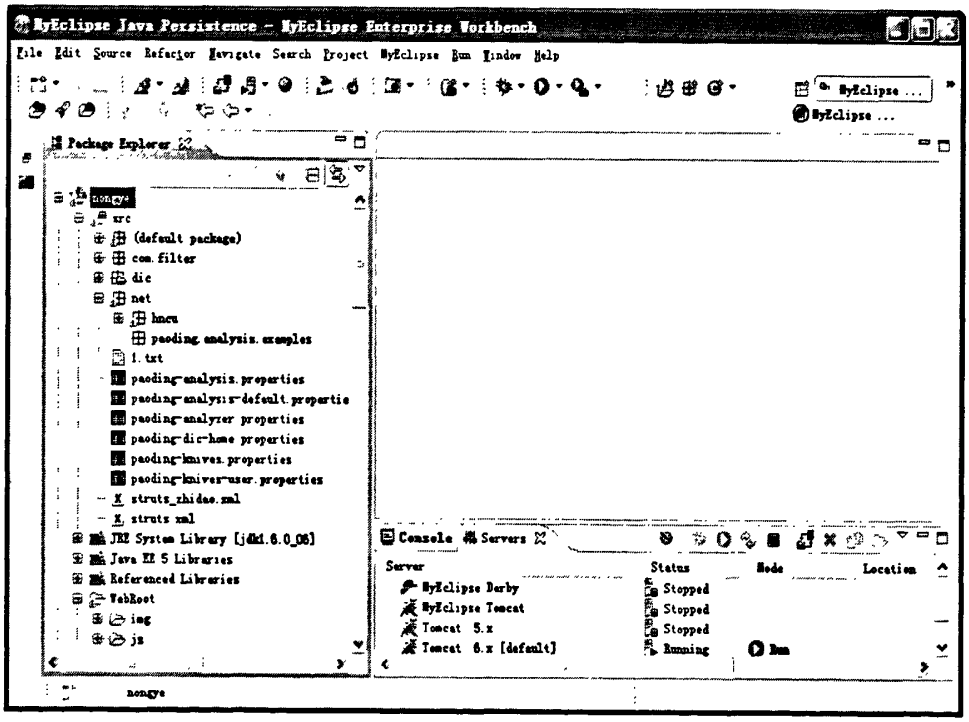

图 21 系统开发环境图

Fig. 21 the diagram of the IDE for developing

5.2 数据集和偏好建模

该实验从农博网和中国农业网分别截取网页 1580 个作为知识问答对，其中涵盖蔬菜、果业、酒业、水产、园林花卉、畜牧家禽、农用资材等多个农产品类别。其中的网页信息按类别作为问答对存入本体问答库中，供用户进行问答测试。

表 14 测试页面分布表

Tab. 14 the distribute table of testing pages

种类	玉米	大豆	土豆	棉花	茶叶	大米	种猪	甲鱼	其他
页面数	120	50	59	69	153	136	95	130	768
特征词数	1001	523	422	652	1522	889	900	956	2889

注册两个用户群，其中群组一（邀请农科的朋友来充当）：年龄 25-40，大学本科-高中学历，近视-视力正常，现拟定专业种植户和养殖散户；群组二（邀请工科的朋友来充当）：年龄 20-55，小学-初中学历，视力正常-老花，现拟为种植散户和养殖散户，两个用户群分别对其感兴趣的信息进行访问，建立起兴趣模型，结果如下：

表 15 首次访问得到的短期兴趣平均值（2011-3-10）.

Tab. 15 the short interests average for the first access (2011-3-10)

用户群一短期兴趣			用户群二短期兴趣		
类 ID	类名称	兴趣度均值	类 ID	类名称	兴趣度均值
1	玉米种植	80.5552253	1	玉米种植	65.2256981
2	大豆种植	60.3356874	2	大豆种植	50.3235482
3	农用资材	20.3698523	3	农用资材	13.1446799
4	种猪饲养	30.1465587	4	种猪饲养	38.22135114
5	茶叶加工	5.32245561	5	茶叶加工	8.22344117

（2）对短期兴趣进行遗忘，并加入第二批新的数据，得到的短期兴趣平均值（2011-3-16）

表 16 第二次访问得到的短期兴趣平均值（2011-3-16）

Tab. 16 the short interests average for the second access (2011-3-16)

用户群一短期兴趣			用户群二短期兴趣		
类 ID	类名称	兴趣度均值	类 ID	类名称	兴趣度均值
1	玉米种植	76.3522253	1	玉米种植	63.2256251
2	大豆种植	61.2236874	2	大豆种植	48.3695482
3	农用资材	18.3623523	3	农用资材	11.3631441
4	种猪饲养	12.1465556	4	种猪饲养	7.22696114
5	茶叶加工	0.32245564	5	茶叶加工	1..98521771

对比两次的短期兴趣，可以看出经过加入新数据和遗忘后，用户群的兴趣度均值发生了变化。显而易见，其中玉米种植，大豆种植，农用资材的兴趣均值上下浮动变化较小，而种猪饲养和茶叶种植的兴趣度均值变化大，有明显的大幅下降，这说明用户的短期兴趣具有随机性和偶然性，是短期不稳定的，同时也说明了遗忘因子对不同兴趣度的类遗忘进行了权值修正，使主要的兴趣度显示出来。

（3）对短期兴趣进行遗忘，并加入第三批新的数据，得到的短期兴趣和长期兴

趣（2011-3-20）

表 17 第三次访问得到的短期兴趣平均值（2011-3-20）

Tab. 17 the short interests average for the third access (2011-3-20)

用户群一短期兴趣			用户群二短期兴趣		
类 ID	类名称	兴趣度均值	类 ID	类名称	兴趣度均值
6	大米种植	10.3666253	6	大米种植	6.2253231
7	甲鱼饲养	9.22236362	7	甲鱼饲养	6.3595365
4	种猪饲养	8.1465556	4	种猪饲养	7.22696114

表 18 第三次访问得到的长期兴趣平均值（2011-3-20）

Tab. 18 the long interests average for the third access (2011-3-20)

用户群一长期兴趣			用户群二长期兴趣		
类 ID	类名称	兴趣度均值	类 ID	类名称	兴趣度均值
1	玉米种植	79.3621253	1	玉米种植	69.361358
2	大豆种植	65.2232351	2	大豆种植	50.3325642
3	农用资材	19.3623143	3	农用资材	133658941

对比用户群的偏好模型，发现短期兴趣中的类发生变化，其中大米种植主题，甲鱼种植主题新出现在短期兴趣中，这说明用户群有了新的兴趣。而玉米种植，大豆种植，种猪饲养出现在了用户群的长期兴趣中，这说明随着时间的增加，兴趣度不断增长，达到一定条件后，短期兴趣转化为了长期兴趣。而茶叶种植主题类却在用户偏好模型中消失了，这说明对其行进了遗忘。

（4）对短期兴趣和长期兴趣进行遗忘，并加入第四批数据，得到短期兴趣和长期兴趣（2011-3-28）。

表 19 第四次访问得到的短期兴趣平均值（2011-3-28）

Tab. 19 the short interests average for the forth access (2011-3-28)

用户群一短期兴趣			用户群二短期兴趣		
类 ID	类名称	兴趣度均值	类 ID	类名称	兴趣度均值
6	大米种植	29.6636116	6	大米种植	30.1233231
7	甲鱼饲养	2.2636362	7	甲鱼饲养	3.35932255
4	种猪饲养	2.14623616	4	种猪饲养	3.22352164

表 20 第四次访问得到的长期兴趣平均值（2011-3-28）

Tab. 20 the long interests average for the forth access (2011-3-28)

用户群一长期兴趣			用户群二长期兴趣		
类 ID	类名称	兴趣度均值	类 ID	类名称	兴趣度均值
1	玉米种植	80.3625313	1	玉米种植	72.3636328
2	大豆种植	68.2232151	2	大豆种植	51.332621
3	农用资材	18.3623113	3	农用资材	12.3632541

对比用户群的偏好模型，在短期兴趣中，大米种植的均值有所增长，而甲鱼饲养和种猪饲养的兴趣度均值都有所下降，这说明用户对这两类兴趣逐渐下降，而对大米种植有所增长；另外长期兴趣中，玉米种植，大豆种植和农用资材的兴趣度均值都有所变化，但幅度不大，这说明，系统对长期兴趣遗忘比较慢，而对短期兴趣遗忘较快。

5.3 评测标准

评价一种检索方法性能的两个主要指标是查全率和查准率[40-47]。本实验中所提出用户偏好方法旨在保证查全率的前提下提高信息检索的查准率，其查准率的计算公式如下：

$$查准率=\frac{检索到的相关文献数}{检索到的全部文献数}$$

5.4 实验结果与分析

在实验过程中,对于一个新的用户查询同时采用用户兴趣模型和不采用用户兴趣模型两种方法测试它们的查准率。每个用户提交了 5 个新的查询。测试结果如表 21 所示：

表 21 测试结果表

Tab. 21 the table of testing results

查询次数	用户群一		用户群二	
	使用模型（查准率）	使用模型（查准率）	使用模型（查准率）	使用模型（查准率）
1	85.23	12.12	45.31	15.85
2	70.55	20.32	50.69	32.49
3	88.36	30.49	60.78	52.65
4	40.67	25.89	50.98	38.26
5	68.46	20.65	64.38	42.67
平均值	70.65	21.89	54.42	36.38

1.从实验结果看，用户群一和用户群二的查准率均有所上升，这说明使用偏好模型的检索方式比普通方式的查准率有了一定的提高。

2.用户群一的查准率明显高于用户群二。这是由于用户群一的专业性很强，而且由于专业性和文化层次的不同，导致模型中主题特征词的类别和数量有所差异。这说明兴趣模型起了作用，但对于文化程度低和专业性不强的人群对系统的使用有待改进。

3.该实验结果数据比较少，测试用户具有了特殊的代表性，虽然用户群二由工科

专业学生充当文化层次低的农民，但毕竟其文化水平较高，导致结果数据有很大的误差，可能在实际应用中用户群二查准率仍有所下降。

4.本模型采用树形层次结构和本体的语义层次结构相对应，有利于模型的建立和将来的改进。

6 总结与展望

本文针对目前农业检索系统中查全率低下的情况，在深入研究并认真学习了搜索引擎及其各项相关技术的基础之上，设计了一个基于本体的农业智能问答系统，给出了该系统的总体设计思想、系统详细的体系结构及系统大致的工作流程，并对应用于其中的各项关键技术的实现进行了深入探讨，建立了用户兴趣模型的整体框架与工作流程，着重研究了用户偏好的收集与处理、模型的建立与更新及用户兴趣模型在改进系统中的应用。主要工作有以下几个方面：

（1）构建了用户偏好模型。用户兴趣模型的构建与更新是论文的主要研究部分，也是系统个性化的基本要求。本文结合本体技术构建了偏好模型，将用户偏好划分为静态用户偏好，短期用户偏好，长期用户偏好，并各自对其进行了不同程度的遗忘和更新。

(2)对用户提交的查询请求进行查询优化，结合用户偏好模型推测用户查询意图。

(3)对检索结果进行排序，结合用户偏好模型，通过计算检索结果与用户兴趣主题之间的相关度对查询结果进行排序，进而更好的满足用户需求。

本文构造了一个基于用户偏好模型的农业智能问答原型系统，但依然处于探索阶段，必然存在一些不足之处，需要进一步的研究与改进。下一步的研究重点主要包括以下几个方面：

（1）对于文化低的用户应该提供友好的导航式提问，以帮助农民用户更好的使用该系统，从而更好的辅助农业生产。

（2）本原型系统的实验是在实验室的小型局域网内进行测试的，对于在实际中应用效果和安全性尚未验证，这有待于接下来寻找合适的实验条件进行系统的改进。

（3）在本系统中，用户偏好的模型和数据的加入都是成批进行的，总的来说是一个“离线”的过程，接下来我们可以考虑如何实时的更新用户偏好，使系统及时的为用户提供个性化服务。

目前个性化服务是一个热点，很多学者都在尝试改进目前搜索引擎查准率低的状况，本原型系统的构建，只是本人在参考前人成果的基础上，对遗忘算法和构建方法进行了一定的改进。该原型系统在保证查全率的条件下，提高了查准率，符合实验前的预想结果。本人不才，希望在以后的工作和学习中，与广大其他业内人士互相交流借鉴，共同创建服务于社会的个性化服务。

参考文献

[1] 潘宁.基于语义技术的智能搜索引擎[D].北京:北京邮电大学,2009.1-6

[2] 滕跃.基于用户兴趣的个性化 WEB 检索[D].北京：清华大学,2004.22-30

[3] 姜华.基于本体的智能答疑系统研究[D].山东:山东科技大学,2005.1-20

[4] 滕良娟.基于本体的专家系统模型研究[D].北京:中国石油大学,2007.1-15

[5] 杨晓东.基于本体的智能决策支持系统研究[D].山东:山东科技大学,2006.15-35

[6] 曹茂诚.基于本体的语义检索技术研究[D].山东:山东轻工业学院,2008.12-23

[7] 朱珣.中文自动分词系统的研究[D].湖北:华中师范大学,2004.1-25

[8] 李彦威.基于用户兴趣的个性化元搜索引擎研究[D].河北:燕山大学,2009.5-23

[9] 吴昊.用户网页浏览兴趣模型的建模方法研究[D].北京:北京邮电大学,2009.5-40.

[10] 孙卫琴，李洪成. Tomcat 与 Java web 开发技术详解[M]. 北京:电子工业出版社.2004.

[11] 徐桔.Spring 架构在 Web 服务中的应用研究[D].南京:南京理工大学,2006.

[12] 贺胜.面向现代汉语文本处理的全文检索、自动分词通用系统[D].南京:南京师范大学,2006.1-20

[13] 郭浩军.基于本体的 Web 跨语言信息检索研究[D].北京:华北电力大学,2008.1-11

[14] 孔繁超.个性化信息服务中的用户偏好的动态挖掘[J].信息系统，2009（6）.111-113

[15] 赵雪梅，朱恩亮.网站用户偏好度的数据挖掘模型[J].盐城工学院学报，2005.30-55

[16] 李晓军.基于 Web 日志挖掘中的数据预处理[J].信息科技,2004.157-158

[17] 周则顺，水俊峰.基于 Web 日志挖掘的智能站点体系[J].武汉理工大学学报，2003（6）.72-75

[18] 左明慧.基于 Web 日志挖掘的网页推荐系统的设计[J].常州工学院学报，2007（4）.5-7

[19] 谢秋华.Web 文本挖掘的相关技术探讨[J].长春理工大学学报[D]，2010（7）.55-56

[20] 杨毅超，黄璜.基于 Web 日志挖掘的农业自适应网站设计[J].农机化研究，2009（7）.110-112

[21] 夏敏捷，张慧档.基于 Web 日志挖掘的个性化服务站点[J].微计算机应用，2006（1）.35-38

[22] 陈佳，吴军华.一种基于 Web 日志挖掘的用户偏爱度度量方法[J].南京工业大学学报，2008（2）.79-82

[23] 李宝敏，韩岳松.本体环境下用户偏好库的查询算法扩展[J].西安工业大学学报，2007（27）.480-483

[24] 梅翔，孟祥武.个性化信息服务中的用户偏好与行为分析[J].理论与探索，2008

（3）.4-6

[25] 刘超，李绍稳.农业科学叙词表向农业本体转化系统的研究与实现[J].农业网络信息，2010（1）.23-26

[26] 陈琳娜.基于本体的个性化用户模型研究[D].河北：燕山大学,2006.10-35

[27] 顾雅枫.基于用户兴趣模型的信息检索研究 [D].兰州：兰州大学,2009.25-35

[28] 蒋萍.基于用户兴趣挖掘的个性化模型研究与设计[D].兰州：兰州大学，2009（22）.75-78

[29] 陈沈焰，吴军华.基于概念语义相似度计算及应用[J].微电子学与计算机，2008（12）.96-99

[30] 张亮.基于本体的个性化元搜索引擎[D].天津大学,2005.13-14

[31] 王宏生 张琳.基于本体的文本自动分类[J].计算机与网络,2009.549-552

[32] 秦春秀，赵捧未.基于用户兴趣的个性化检索[J].情报学报，2005（4）.449-452

[33] 鲜国建.农业科学叙词表向农业本体转化系统的研究与实现[D].中国农业科学院,2008.6-20

[34] 刘超，李绍稳等. 农业科学叙词表向农业本体转化系统的研究与实现.农业网络信息[D],2010(1).23-26

[35] 顾基发.专家挖掘-实现知识的综合[J].工程研究，2010（2）.100-107

[36] 颜芳.基于隐形知识外化模型的技术信息平台研究[J].科学学研究，2008（1）.124-129

[37] 王平.基于用户偏好挖掘和主题搜索的情报推荐系统[D].浙江：浙江大学,2007.9-29

[38] 马雪.基于本体的隐性知识管理系统研究[D].西北大学,2008.3-9

[39] 黄海清，张平.用户偏好提取算法[J].无线电工程，2006（3）.16-19

[40] 吴金红.一种基于本体论的知识检索原型系统[J].情报杂志,2004,(7).45-49

[41] 马雪.基于本体的隐性知识管理系统研究[D].西北大学,2008.3-9

[42] 姜丽红，徐博艺.信息筛选中群体用户偏好聚合模型[J].上海交通大学学报，2000（6）.818-820

[43] 蒋萍，崔志明.智能搜索引擎中用户兴趣模型分析与研究[J].微电子学与计算机，2004（11）.24-26

[44] 胡昌平，邵其赶.个性化信息服务中的用户偏好与行为分析[J].理论与探索，2008（3）.4-6

[45] Mike Jasnowski. Java, XML 和 Web 服务[M].北京:电子工业出版社，2002-5: 65~67.

[46] Woongsup kim, WSCPC: An architecture using semantic web services for collaborative product commerce[J]. Computers in Industry, 2006,57: 787~796.

[47] M.M.Lehillan,L.A.Belady. ProgramEvolution: Proeesses of Software Change [J]. AeademiePress, 1985, 5: 12~16.

作者简介

赵全东，男，河北张家口人，生于1982年1月，工学硕士，毕业于河北农业大学信息科学与技术学院。

教育背景：

2008--2011：河北农业大学 计算机应用技术 硕士

2006--2008：河北农业大学 计算机科学与技术 本科

2003--2006：河北农业大学 计算机网络技术 专科

在读期间发表的学术论文

在学期间发表的学术论文：

赵全东，王芳，任力生等. 农业智能问答系统中的用户偏好研究.河北农业大学学报，2011，34(1)：127-130.

就读硕士期间参加的研究项目：

2008--2009：参与河北农业大学研究生管理与创新平台的开发。

致谢

这次论文能够得以顺利完成是所有指导过我的老师，帮助过我的同学，一直支持着我的家人对我的教诲、帮助和鼓励的结果。我要在这里对他们表示深深的谢意！

首先要特别感谢我的导师王芳教授，在我撰写论文的过程中，王老师给予细心地指导和关怀，无论是在论文的选题、构思和资料的收集方面，还是在论文的研究方法以及成文定稿方面，我都得到了王老师悉心细致的教诲和无私的帮助，特别是她广博的学识、深厚的学术素养、严谨的治学精神和一丝不苟的工作作风使我终生受益，在此表示真诚地感谢和深深的致敬。

衷心感谢学院三年来给我提供的良好科研环境，为我提供参与课题锻炼的机会，极大地提高了实践能力。

感谢我的家人，在任何时候他们都是我坚强的后盾，给我最大的鼓励和支持。

最后，衷心感谢在百忙之中评阅论文和参加答辩的各位专家、教授！

ISSN 1000-1573
CODEN HNDBEM

河北农业大学学报

HEBEI NONGYE DAXUE XUEBAO

JOURNAL OF AGRICULTURAL UNIVERSITY OF HEBEI

第34卷 第1期
2011

目 次

期刊基本参数：CN 13-1076/S * 1959 * b * A4 * 130 * zh * P * ¥10.00 * 1000 * 27 * 2011-01

文章编号：1000-1573(2011)01-0127-04

农业智能问答系统中的用户偏好研究

赵全东， 王 芳， 任力生
（河北农业大学 信息科学与技术学院，河北 保定 071001）

摘要：针对当前农业问答系统中召回率较低的现状，研究了农业智能问答系统中的用户偏好问题，探讨了用户查询记录的挖掘和学习，用户兴趣模型的构建方法等问题。试验结果表明：该方法优化了检索过程，在基本保证查全率的前提下，提高了检索系统的查准率。

关 键 词：本体；用户偏好；个性化检索；智能问答
中图分类号：TP 391 **文献标志码**：A

Research on user profile in agricultural intelligent question-answering system

ZHAO Quan-dong, WANG Fang, REN Li-sheng
(College of Information Science&Technolog, Agricultural University of Hebei, Baoding 071001, China)

Abstract: This paper deals with user profile in agricultural intelligent question-answer system. The development and study of querying records, construction method of user interest model, privacy protection of users, etc. were discussed. The results showed that the retrieval process was optimized, and precision was improved.

Key words: ontology; user profile; personalized retrieva; intelligent question answering

目前的搜索引擎，对于1个给定的查询请求，只能返回一般化的检索结果，而不能针对不同用户的特殊需要返回相应的信息，这是因为检索工具不能很好地理解用户意图，为解决这一问题，已提出多种探索方法来提高检索工具的智能性，以增强其对用户查询意图的理解能力，提高检索工具的查准率。

李宝敏、韩岳松在本体环境下用户偏好库的查询算法扩展一文中[1]，依据本体论，采用用户兴趣剖像算法，扩展查询算法，使得在基于本体的智能搜索中更多更准确地体现出用户兴趣爱好，进而大大提高了检索的速度和查准率；孔繁超在个性化信息服务中用户偏好的动态挖掘一文中[2]采用聚类、关联规则等技术，对用户偏好进行了动态的挖掘，通过追踪用户需求序列，最终产生Top-N产品推荐，提高了推荐系统的推荐质量；赵雪梅、朱恩亮在网站用户偏好度的数据挖掘模型一文中[3]，基于统计学的观点讨论了网站用户偏好度的数据挖掘模型，设计了1个网站用户信息浏览偏好度的数据挖掘模型并应用于通用商品销售系统，提高了产品的查准率。

近年来，虽然国内外也涌现出了很多农业相关网站，如农搜网和神州蔬菜网等百余家。但这些传统的搜索引擎仍仅依据用户输入的查询关键字和网页的重要性返回结果，并没有引入用户的偏好差异分析，为此本文研究并构建1种基于用户偏好的农业智能问答系统，对提高网络对农业的服务力度具有一定的实用意义。

收稿日期：2010-06-07

作者简介：赵全东(1982-)，男，河北省蔚县人，在读硕士生，从事计算机网络数据库研究.
通讯作者：王 芳(1970-)，女，河北省保定市人，博士，教授，从事计算机网络与人工智能研究.

1 用户偏好获取

1.1 用户信息的收集

显式信息收集：在用户注册时要求用户输入年龄、学历、专业、性别、职业、视力、Email、工作区域、是否色弱等相关信息，利用以上信息主要为用户设置页面风格。

隐式信息收集[4]：该方式需要收集的信息包括用户输入搜索引擎的查询关键词；用户浏览的页面；用户浏览的行为；用户下载、保存的页面和资料等；用户手工输入的其它信息；用户近期提出的问题，回答的问题，来往的专家；Web 服务器日志。

1.2 偏好信息的挖掘

数据挖掘模块被动的接受处理过的数据，通过分类、聚类等统计方法，对用户的行为进行分析。首先需要对单个用户访问过的问答对查找分类，然后将这些结果作为对单个用户偏好的描述，进而对用户再进行聚类，以完成相似用户的信息推送。

2 用户偏好的建模

2.1 偏好模型的定义

由于本体论[5-6](Ontology)对特定领域对象的表示与描述具有规范性、可重用性、可靠性等特点，为此本文将本体论应用于信息检索领域，对文档、用户的个性化模型进行描述，以优化系统服务效率。

偏好模型[7]：$Hobby=\{(Htl,Cl),(Ht2,C2),\cdots,(Hti,Ci),\cdots,(Htn,Cn)\}$，其中 $Hti=(ti,wi)(1<=i<=n)$ 是 1 个二元组，表示用户兴趣节点，ti 为用户偏好主题，wi 为主题 ti 的兴趣浓度值，n 为用户兴趣主题的总数，所有的 ti 构成用户偏好主题集 T。Ci 是属于主题 ti 的兴趣特征序词集，如果主题 ti 包含 m 个特征序词，则 Ci 可以表示成 $Ci=\{(eil,vil),(ei2,vi2)\cdots,(eij,vij)\cdots,(eim,vim)\}$，其中 $eij(l<=j<=m)$ 表示兴趣主题 ti 包含的所有特征序词，m 表示主题 ti 下所有特征序词的个数，vim 表示对应特征序词在用户兴趣 ti 中的权值。用户兴趣特征序词集为 $C(T)=Uci$。

2.3 偏好模型生成算法

静态偏好生成算法[7]：将用户的专业、职业和工作区域赋予 1 个较高的权值，用来聚类用户并为用户推送群内消息。

动态偏好生成算法[7]：(1)将用户偏好模型设为树结构，置根结点为用户信息，其他子节点为主题分类信息，叶子节点为该主题下的所有关键词。(2)设定一阀值 β，若兴趣浓度值 $wi<\beta$，则置偏好兴趣树中叶节点的初始权值为 $vim=5$，否则初始权值为 $vim=10$；(3)参考偏好模型，生成偏好兴趣树，并自下而上逐层计算各父结点(各兴趣主题)Node(ti)的权值 $W_i=\sum_{i=1}^{k}V_i$，k 为子类的个数，$v_i(0<i<k+1)$ 为其子类的权值。

2.4 偏好模型的更新

在对模型进行更新时，不仅要加入用户的最新兴趣特征序词，还要对模型中已有特征词的权值进行调整，即对现有词条的权值乘以遗忘因子 $F(x)$ 进行修正。遗忘因子 $F(x)$ 为 $F(x)=e^{-[\frac{\ln 2(cur-fir)}{ha}\pm\delta]}$ 其中 cur 表示当前日期，fir 表示兴趣特征序词第一次在模型中出现的日期，ha 表示半衰期，ha 值应根据大量实验测试确定，也可以根据经验人为设定，$\delta(\delta>0)$ 为遗忘因子修正值，对于词条权值小于某一阀值 ε 的，公式中取＋，否则取－，以尽快淘汰老此条。

2.5 个性化过滤

当用户提交查询后，将其提问进行本体化和分词处理，结合用户的偏好模型，确定搜索范围，即确定用户提问应该在哪些主题类中搜索。这样先确定分类的情况下，避免了在所有主题中搜索后再过滤，提高了搜索效率。

3 实验与结果分析

3.1 实验环境

鉴于系统特性，选择本实验在 Java 技术平台上进行开发，具体环境和工具为：开发语言选择 Java5，开发工具为 MyEclipse 6.5；使用 Stusts 2 框架实现 MVC 分层；采用 Mysql 数据库和 Hibernate 等数据库技术实现数据高效存取。Java 编译和运行环境 jdk1.6.0_06，Web 服务器软件 Tomcat6.0，本体编辑工具软件 Protégé3.3。具体使用的技术还包括 jsp，Spring，ext 等等。

3.2 数据集

该实验从农博网和中国农业网分别截取网页 620 个作为知识问答对存入本体问答库[8]中，其中涵盖蔬菜、果业、酒业、水产、园林、畜牧等多个农产品类别，供用户进行问答测试，问答对数目如表 1 所示。

表 1 问答对数目

Fig. 1 The number of Q&A

蔬菜 Vegetable	果业 Fruit industry	酒业 Vino industry	水产 Aquatic	园林 Gardens	畜牧 Farming	其他 Others
50	60	80	100	60	150	120

3.3 关键代码

系统检索部分核心代码如下所示:

```
public String Search (String search, String wenben,PianHao pianhao) throws Exception {
    if (search.length() ! = 0) {
//根据用户输入的检索词和用户偏好确定最终提交的检索词
    QUERY =PianHao( search,pianhao);
    }
    String result = null;
// 将庖丁封装成符合 Lucene 要求的 Analyzer 规范
    Analyzer analyzer = new PaodingAnalyzer();
// 读取本类目录下的 text.txt 文件
    String content = wenben;
// 接下来是标准的 Lucene 建立索引和检索的代码
    Directory ramDir = new RAMDirectory();
     IndexWriterwriter = new IndexWriter(ramDir, analyzer);
    Document doc = new Document();
    Field fd = new Field(FIELD _ NAME, content, Field.Store.YES,
  Field.Index.TOKENIZED,
Field.TermVector.WITH _ POSITIONS _ OFFSETS);
    doc.add(fd);
    writer.addDocument(doc);
    writer.optimize();
    writer.close();
     IndexReader reader = IndexReader.open(ramDir);
    String queryString = QUERY;
     QueryParser parser = new QueryParser(FIELD _ NAME, analyzer);
    Query query = parser.parse(queryString);
     Searcher searcher = new IndexSearcher(ramDir);
    query = query.rewrite(reader);
    Hits hits = searcher.search(query);
    BoldFormatter formatter = new BoldFormatter();
    Highlighter highlighter = new Highlighter(formatter, new QueryScorer(query));
    highlighter.setTextFragmenter(new SimpleFragmenter(50));
    for (int i= 0; i<hits.length(); i++) {
    String text = hits.doc(i).get(FIELD _ NAME);
    int maxNumFragmentsRequired = 5;
    String fragmentSeparator = "...";
    TermPositionVector tpv = (TermPositionVector) reader
      .getTermFreqVector(hits.id(i), FIELD _ NAME);
    TokenStream tokenStream = TokenSources.getTokenStream(tpv);
    result = highlighter.getBestFragments(tokenStream, text,
      maxNumFragmentsRequired, fragmentSeparator);
    // System.out.println("/n" + result);
    }
    reader.close();
    return result;
  }
```

3.4 评测标准

评价一种检索方法性能的 2 个主要指标是查全率和查准率[8-9]。本实验中所提出用户偏好方法旨在提高信息检索的查准率。

$$查准率=\frac{检索到相关文档数}{检索到全部文档数}\times 100\%$$

3.5 实证分析

搭建基于用户偏好的农业智能问答系统的原型系统,注册 50 位用户进行测试,对于 1 个新的用户查询同时采用用户兴趣模型和不采用用户兴趣模型 2 种方法测试它们的查准率,每个用户提交 5 个新的查询。从中随机抽取 2 位用户的结果如表 1 所示,其中 1 位年龄 25,大学本科学历,专业为园艺,视力正常,职业为花卉园艺师;另外 1 位年龄 55,小学学历,视力老花,职业为茶叶种植户,测试结果如表 2 所示。

表2 查准率分析
Fig. 2 Precision ratio analysis

查询次数 Query count	用户一 User 1		用户二 User 2	
	使用模型/% Precision ratio of used model	不用模型/% Precision ratio of unused model	使用模型% Precision ratio of used model	不使用模型/% Precision ratio of unused model
1	85.23	12.12	45.31	15.85
2	70.55	20.32	50.69	32.49
3	88.36	30.49	60.78	52.65
4	40.67	25.89	50.98	38.26
5	68.46	20.65	64.38	42.67
平均值	70.65	21.89	54.42	36.38

1. 从实验结果看，用户一和用户二的查准率均有所上升，使用偏好模型的检索方式比普通方式的查准率有了一定的提高。

2. 用户一的查准率明显高于用户二。这是由于用户二的专业性很强，而且由于文化水平的不同，导致模型中主题特征词的类别和数量有所差异。这说明兴趣模型起了作用，但对于低文化用户的使用有待改进。

3. 该实验结果数据比较少，测试用户具有了特殊的代表性，导致结果数据有一定的误差。

4. 本模型采用树形层次结构和本体的语义层次结构相对应，有利于模型的建立和将来的改进。

4 结论

本研究文采用 ontology 技术存储用户提问，挖掘用户的兴趣爱好，构建用户兴趣模型，实现用户的个性化查询。试验结果表明，使用模型的情况下用户一的平均查准率为 70.65%，用户二的平均查准率为 54.42%，均比传统查准率有了明显的提高，可见用户偏好模型在搜索问答中起到的一定得作用。

但本研究也存在很多的问题有待进一步的改进，比如对于文化低的用户应该提供友好的导航式提问等，诸如此类问题将在以后的学习中深入研究。

参考文献：

[1] 李宝敏，韩岳松. 本体环境下用户偏好库的查询算法扩展[J]. 西安工业大学学报，2007(27)：480－483.

[2] 孔繁超. 个性化信息服务中的用户偏好的动态挖掘[J]. 信息系统，2009(6)：111－113.

[3] 赵雪梅，朱恩亮. 网站用户偏好度的数据挖掘模型[J]. 盐城工学院学报，2009(22)：75－78.

[4] 梅翔，孟祥武. 个性化信息服务中的用户偏好与行为分析[J]. 理论与探索，2008(3)：4－6.

[5] 胡昌平，邵其赶. 个性化信息服务中的用户偏好与行为分析[J]. 理论与探索，2008(3)：4－6.

[6] 刘超，李绍稳. 农业科学叙词表向农业本体转化系统的研究与实现[J]. 农业网络信息，2010(1)：23－26.

[7] 顾雅枫. 基于用户兴趣模型的信息检索研究[D]. 兰州：兰州大学，2009：25－35.

[8] 卫国平. 基于概念语义的用户兴趣模型的研究[D]. 太原：太原理工大学，2008：9－13.

[9] 滕跃. 基于用户兴趣的个性化 WEB 检索[D]. 北京：清华大学，2004：22－30.

（编辑：张月清）

（上接第 122 页）

参考文献：

[1] 徐贵宝，肖延敏，王燕. 移动 IPV6 技术及其在 3G 网络中的应用[J]. 现代电信科技，2004(12)：38－39.

[2] Narten T，Nordmark E，Simpson W. RFC2461：Neighbor discovery for IP version6 (IPv6)[S]. California：IETF，1998.

[3] Cohen M I. Source attribution for network address translated forensic captures[J]. Digital Investigation，2009，5(3)：138－145.

[4] Gianluca Varenni，Loris Degioanni，Fulvio Risso，et al. WinPcap：The Windows Packet Capture Library[EB/OL]. http://www. winpcap. org/docs/default. htm，2008.

[5] Ching-Yu Lin，Ing-Yi chen，Sy-Yen kuo. Extension headers for IPv6 anycast[J]. Computer communications，2006(29)：3013－3019.

[6] habeman B. RFC 3307：Albcation guidelines for IPv6 multicast addresses[S]. California：IETF，2002.

[7] Lilian H. Detection of complex temporal patterns over data streams[J]. Information Systems，2004，29(6)：439－459.

[8] Motwani R，Widom J，Arasu A，et al. Query processing，approximation，and resource management in a data stream management system[C]//Proc Conf on Innovative Data Syst Res Pacific Grove，California：IEEE Society，2003：245－256.

（编辑：张月清）

●中文核心期刊
●教育部一、二、三届优秀科技期刊
●华北优秀科技期刊
●河北省优秀期刊
●河北省教育系统十佳期刊
●美国《化学文摘》(CA)期刊源
●俄罗斯《文摘杂志》(AJ)期刊源
●英国《动物学记录》(ZR)期刊源
●"日本科学社数据库"收录期刊
●国际农业与生物中心数据库(CABI)期刊源
●《中国学术期刊综合评价数据库》来源期刊
●《中国科学引文数据库》来源期刊
●《中国学术期刊(光盘版)》全文收录期刊
●"中国期刊网"全文收录期刊
●《农业索引》(Agrindex)数据库期刊源
●《中国农业文摘(7个分卷)》期刊源
●《中国生物学文摘》期刊源
●《中国林业文摘》期刊源
●《麦类文摘》期刊源
●《食品文摘》期刊源

责任编辑：梁　虹
英文编辑：马红军

河北农业大学学报

HEBEI NONGYE DAXUE XUEBAO

双月刊，1959年创刊

第34卷　第1期(总第155期)

2011年1月

JOURNAL OF AGRICULTURAL UNIVERSITY OF HEBEI

Bimonthly, Started in 1959

Vol. 34　No. 1(Sum. No. 155)

Jan. 2011

主管单位：河北省教育厅
主办单位：河北农业大学
编辑出版：河北农业大学期刊社
　　河北省保定市，邮编071001
电　　话：0312-7521323，7521322
主　　编：刘大群
副 主 编：马峙英　孙建恒　黄金祥
常务副主编：黄金祥
印　　刷：保定市华泰印刷厂
发行范围：国内外公开发行
国内发行：保定市邮政局
国外发行：中国国际图书贸易总公司
　　(北京399信箱)

Responsible Institution: Hebei Education Department
Sponsor: Agricultural University of Hebei
Editor and Publisher: Journal Publishing Department of Agricultural University of Hebei
Address: No. 289 Lingyusi Street, Baoding 071001, China
Telephone: +86-0312-7521323, 7521322
E-mail: xuebao@hebau.edu.cn
Editor-in-chief: LIU Da-qun
Vice Editor-in-chief: MA Zhi-ying　SUN Jian-heng　HUANG Jin-xiang
Administrative Vice Editor-in-chief: HUANG Jin-xiang
Domestic Distributor: Post Office of Baoding City, Baoding, China
Abroad Distributor: China International Book Trading Corporation (P. O. Box 399, Beijing, China)

国际标准连续出版物号：ISSN 1000-1573
国内统一连续出版物号：CN 13-1076/S

国内代号 18-43
国外代号 Q4487

国内定价：10.00元/册
　　60.00元/年

ISSN 1000-1573

9 771000 157049 01>