文章编号 1004-6410(2004)03-0094-05

基于Java自动答疑系统的设计与实现

徐奕奕[1],陈小花[2]

(1.广西工学院计算机工程系,广西 柳州 545006;2.柳州市建设委员会,广西 柳州,545001)

摘 要:自动答疑系统是当前信息化教育发展的热点。该文对自动答疑系统的关键技术:初始领域知识库的构建、汉语词条切分技术、web页面的全文搜索技术进行了论述。该系统基于B/S模式架构,用Java技术进行开发,同时在安全性、可移植性和准确性方面做了充分考虑,测试结果表明能满足实用要求。

关 键 词:词条切分;搜索;自动答疑系统;B/S

中图分类号:TP393 **文献标识码**:A

0 引 言

信息化教育是在信息化环境下,以学生为主体,强调学生个性化学习和自主学习的新型教与学模式,是一种发展迅猛的新型教育形态。一方面丰富的网上资源和多元互动的教学环境,让学生从知识获取方式的单一性变成网络化,从局部刺激变成多元刺激,极大激发了学生的学习主动性、积极性;另一方面由于信息化教育尚处在发展阶段,在硬件提供和教育支持服务体系上不能完全满足需要,传输信息单一,传输速度慢,实时性差等问题,使得效果受到很大影响。在进行自动答疑系统的设计时,针对实际情况,对最常见的Internet使用方式的支持加以考虑,基于传统的B/S模式,Tomcat为web服务器,MySQL为后台数据库,采用Java语言进行开发,突出系统的实用性、安全性、稳定性和可移植性,为学生提供了自适应性学习机制。

1 自动答疑系统应解决的问题

自动答疑系统主要解决的问题有:(1)提供基于Internet的自动答疑功能;(2)提供"疑问-解答"库生成、检索、管理功能;(3)提供与网络课程软件的标准接口;(4)提供利用"疑问-解答"库生成"样板题"与一问一答式教学课件的功能;(5)提供滞后式答疑功能,即对计算机无法自动回答的问题时,将问题反馈给相应教师,由教师回答后再返回给提问者,并更新答疑资料库[1]。系统主要的功能是自动答疑,要体现智能性,用户不仅可以输入关键词的组合来寻求问题的答案,也可以输入自然语言描述的问题。如"C语言的基本数据类型是什么?",另外用户还可以对答案的性质做限制,如答案修改最新时间,指定答题教师等。

在用户问题提出后,系统提供了在"疑问-解答"库中搜索与问题相关答案的过程,并转至教师BBS讨论区两种手段,后者是对系统现有答疑能力的补充。提问和讨论是一个统一的整体,这体现在提问没有得到相关的答案材料,系统自动将问题转贴在讨论区域;而讨论的材料也可以当作答案材料被提问搜索。如果相同问题的提出达到一定频度,将自动生成问答式课件,实际上就是将用户与教师的讨论追加到指定的ppt文件中去,可以转为教学子系统的一部分。另外,如果用户的问题不能及时得到解答,将被系统标识为"滞后回答"问题,用电子邮件形式发到相关教师电子信箱中去。

总之,系统应是:(1)开放性,可扩展。任何一个答疑系统的领域知识都不可能覆盖该领域的所有方面,更不可能包含用户潜在所有问题的答案,所以答疑系统必须是开放性的。方便系统维护人员进行领域知识的

收稿日期:2004-06-29

作者简介:徐奕奕(1980-),女,湖南邵阳人,广西工学院计算机工程系教师。

增加、删除和修改,而仍然保持答案知识的结构良好。(2)灵活、方便地接入环境。即实现5个ANY(Anywhere,Anytime,Anywhere,Anyssystem,Anyapplication)。支持任何人、任何地点、任何时间,采用任何系统都能访问,也就是对用户的硬件、软件设施不能做太多要求。

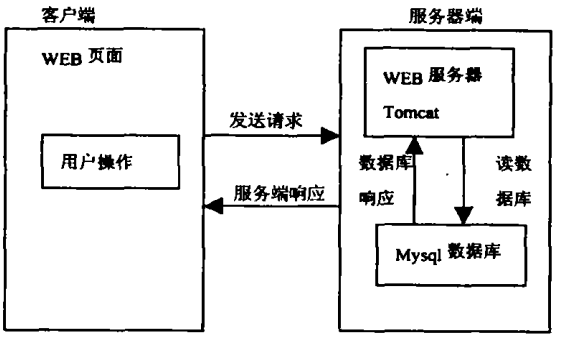

图1　总体方案

2　系统设计与实现

自动答疑系统基于传统的B/S模式开发,图1描述采用的具体方案。其中,数据库在动态网页中扮演很重要的角色,这里的"动态",就是将数据保存在数据库之中,需要时将数据由数据库中取出,再以HTML形式传送给浏览器。本系统选用MySQL数据库为后台,MySQL在数据量的支持,检索速度,管理功能,稳定性方面都能满足预定的要求。表1以"疑问-解答"库中的题库为例给出其数据字典[2]。

表1　题库的数据字典

题库数据表		questiontable						
	字段名	数据类型	大小	默认值	允许空	唯一	主键	自动增加	备注
1	q_ID	Int	50			*	*	(1,1)	问题编号
2	q_NAME	Varchar	100						课题名
3	q_TYPE	Varchar	50	0					类型:0表示专业课;1表示基础课;2表示其它
4	q_CONTENT	Text							内容
5	q_MARK	Varchar	50						解答特征
6	q_ANSWER	int							解答次数
7	q_INQUIRE	int							查询次数
8	q_IMAGE	Varchar	50						图片存放
9	q_DATE	Date Time							入库日期
10	q_DIR	Varchar	10						存放地址
索引	字段名	Date Time					排序		
	q_ID	DB_qstiontable					升序		

系统涉及的典型用户有三类,一是使用系统来解答自己疑问的普通用户(学生),二是答疑系统所涉及领域的教师用户,三是系统维护管理人员。普通用户是系统服务的对象,他们使用系统解答自己的疑问,并对整个系统进行评价,参与到系统开发需求分析与测试的全过程。教师用户的作用非常关键,他们负责系统初始领域知识库的构建,即"疑问-解答"库的初建,该库应按国家教委提出的远程教学规范建设,包括题库,课件库,案例库及各种相关素材,要求库的覆盖面较为全面,且要负责解答系统暂时不能解答的问题。系统维护管理人员由专门人员来担任,负责系统安全,可靠运转。从用户角度出发,自动答疑系统主要包括下面几个模块,如图2所示。

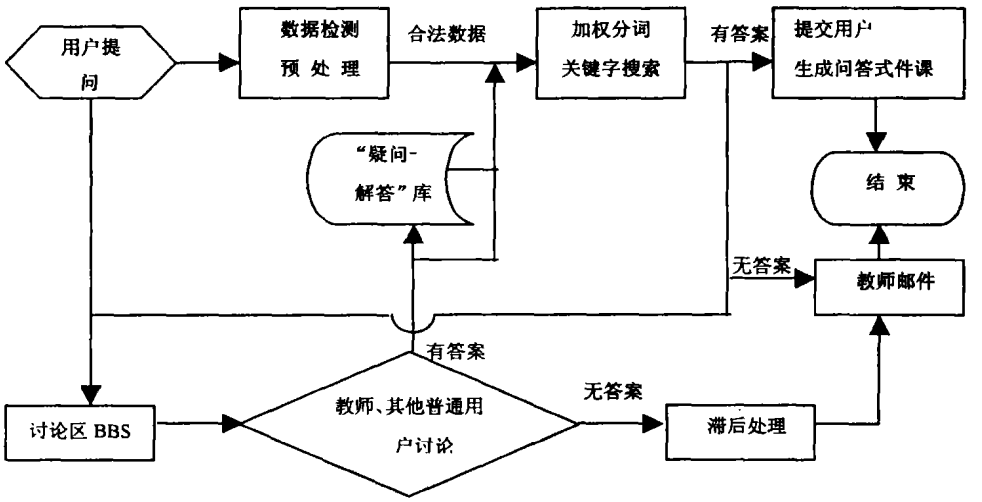

图 2 系统模块图

2.1 提问模块

提问模块实质是提供给普通用户的一个web 页面。当用户提出问题后，提问模块要进行预处理，加权分词和搜索等功能。预处理子模块的作用是对于问题文本进行第一次分解处理，主要是将自然语言的问题文本，根据标点符号、西文字符等分解成子串[3]。加权分词模块的作用主要是在预处理子模块的基础上，将问题文本的系列子串进一步分解成与系统有关的加权关键词的组合。加权关键词的组合与前面提出问题的限制条件相结合，就形成了搜索答案的条件。关键字全文搜索子模块则根据生成的搜索条件，在系统的领域知识库，以及讨论形成的材料中搜索与问题相关的材料，并按照匹配程度返回结果。

2.2 讨论区模块

讨论区模块是普通用户使用自动答疑系统的另外一种基本手段。用户可以参加Web 方式的同步讨论(如BBS 和实时聊天等)，参与者可以是教师用户也可以是其它普通用户。讨论要具备自动更新功能，使BBS 能对最新没有回答的问题置顶，而聊天室可以让客户端定时刷新。另外，用户提问没有得到系统满意解答的问题，系统自动把问题转发给教师用户的指定邮箱。

2.3 问答式课件生成模块

问答式课件生成模块的核心算法是统计算法，统计哪个知识点的问题最集中，例如用户提出同类型题目超过一定数量，如十次，则会自动将用户与教师的讨论追加到指定的ppt 文件中去(即"疑问－解答"库中的课件库)，实际设计中把该html 页面转换为txt 文本，可做为教师教学时的原始资料。

3 关键技术

自动答疑系统首先需要解决的问题就是汉语的词条切分，并从自然语言文本中抽取出能够代表问题的关键词。而关键字全文搜索目的是查找与问题相关的答案。为了使系统能最大效率地工作，必须选择切实可行且匹配精度较高的算法。

3.1 汉语词条切分

词条切分简称分词，是自动文本分析的前提，对汉语进行文本分析时需要进行词干抽取、词法分析后再进行分词。本系统分词所采取的方案是：先使用一部分基本的分词词典(常用词词典)进行串匹配分词，同时使用将串频统计的方法来加权，确定关键字，即包括两个过程－－分词和加权。自动答疑系统的材料是以HTML 文件的格式存储的，所以要先把文档中的文本抽取出来，包括HTML 文件的BODY 中的正文文本、TITLE 标记的标题文本和在HTML 文件头部中以META 标记指定的关键词序列。对于后面两者的文本还要做特殊标记，作为加权时的一个权值确定依据，接着对得到的文本序列进行初始的子串切分。因为在实际应用中用户输入的不可能都是中文，因此要对英文、汉字和数字、标点符号进行区分。最后对没有实际意义但使用频率太高的词进行停用，建立一个停用表，如"了，的，是"之类，不进行检索，以提高效率。

常见的两种基于字符串匹配算法是直接匹配和链接匹配算法。在直接匹配算法中,采用汉字字符序列的典型字典序列,即按字符出现的频度或其ASCII值大小来顺序排列。比如,"星期一/星期二/星期三/星期四",而链接匹配算法如图3所示。

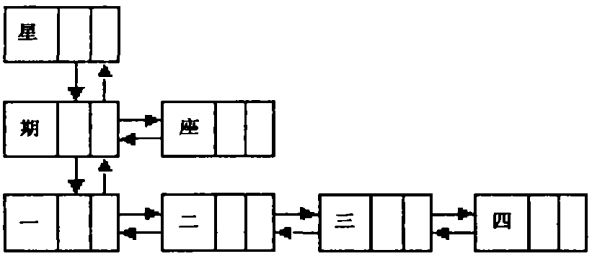

图3　链接匹配法示意图

显然,直接匹配法较为简单,本系统在其基础上采用了典型算法——最大匹配法(Maximun Matching Method),即设词表中最长的词由m个字组成,每次切分时总是从待切分的句子中截取一个长度为m的字段W,查找分词词典,若存在则匹配成功,W被切分出来,否则从W去掉最后一个字,进行新的匹配,循环到匹配成功为止。如果匹配不成功,可以跳过该过程,把其新关键字追加到关键字索引表中。

加权的目的就是确定各个关键词对于答案材料的匹配程度,将匹配程度(相似度)比较高的结果排在前面。比如答案"C语言的基本数据类型有字符型、整型和实型。"其中"数据类型"的匹配程度最高,但是,根据语义和语境来抽取可近似表示答案材料语义的关键词,是一个需要高度智能的问题,这对现有的技术来说难以完全实现。本系统采用的是:对材料中相邻共现的各个字组合的频度进行统计的方法来计算匹配程度,然后再根据各个词的权值,计算分词结果中各个词在同一篇文档中的权值和,权值和超过某个规定值的文档将被按照权值和的大小依次返回。因此一个词相关的记录包括所属文档ID(Identifier)、权值、最佳匹配值m·

3.2　关键字全文搜索

全文搜索的任务是根据用户的需求,提供一组来源于索引库中的相关信息。对此需求的表达通常是汉字查询,在索引库中的每个文档中搜索每个(或所有)关键字。处理查询时有一个简单的方法可打开并扫描每个文档,寻找每个查询词。如直接对多个关键词进行模糊匹配:like"%keyword1%" and like "%keyword2%",这样在处理查询时,需要打开每个文档并搜索关键字词,比较浪费时间[4]。

一个简单的解决方案是:将原始文档中所有基本元素的位置信息记录在索引库中,建立一个索引表,那么处理查询时就不用扫描每个文档了。唯一的要求是用反向索引相互比较文档,并选择"疑问-解答"库中与查询最有关联的文档。在汉语中可选择的基本元素可以是字,也可以是词,从而形成了两种索引库结构,即基于字表的索引库和基于词表的索引库。字表法是将原始文档中每个字的位置信息记录在索引库中;而词表法则是以词为单位,将其位置信息记录在索引库中,需要使用切分词典,因而适用于特定领域中内容相对固定文档的全文搜索,优点是索引库比较小,检索速度快,缺点则是不能适应跨领域的文档处理要求;字表法采用对每个字的出现位置进行统计,不需要任何词典,适用范围强。当然,采用字表法的检索精度没有采用词表法那么高。对于本系统来说,其答疑内容基本上是针对各学科或者专有领域的。本系统采用词表法,建立以"课程名"为第一关键字的索引库来组织全文搜索。

4　问题与展望

该系统经过测试,系统界面友好,支持多用户使用,答案各部分信息完整、丰富,对用户提问的响应平均时间是3秒,最长是10秒,基本达到预期目标,但是该系统仍存在以下问题:

(1) 智能化、自动化程度不够。下一步将建立决策支持系统,主要是引进数据仓库相关技术进行存储、分析、归纳,使系统有一定决策能力。

(2) 对数据库的全文搜索只能针对网页,通用性不够,且答案时有歧义,不能精确定位,分词算法有待改进。

(3) 图文声像一体化尚处在初步阶段,比如语音提问尚未实现。下一步的目标是从学生心理出发,融入生动的多媒体信息,使自动答疑系统更为人性化、更为高效。

[参 考 文 献]

[1]余胜泉,何克抗 .基于 Internet 的数学系统[M].四川:西南师大出版社,2000.
[2]张海藩 .软件工程[M].北京:人民邮电出版社,2002.
[3]郭庆琳,樊孝忠 .自动化应答系统中自然语言理解技术的研究[J].计算机应用研究,2004,(6):24-25.
[4]车 东 .在应用中加入全文检索功能——基于 Java 的全文索引引擎 Lucene 简介[EB/OL].http://www.chedong.com/tech/lucene.html,2004.9.

The design and implementation of the auto-answering system based on Java

XU Yi-yi[1], CHEN Xiao-hua[2]

(1. Dept. of Computer Engineering, Guangxi University of Techology, LiuZhou 545006, China;
2. Liuzhou Committee of Construction, Liuzhou 545001, China)

Abstract: An auto-answering system is a very promising research field of the development of information education. This paper discusses the key technology of the auto-answering system such as the construction of the initial knowledge base, word entry segmentation of Chinese ideograph and the full-tex search technology of web page. By adopting Browser/Server structure and Java technology, fully the security, graft and veracity. The testing results shows that this system can satisfy the practical requirements.

Key words: word entry segmentation of Chinese ; full-text search technology; auto-answering system; Browser/Server

(责任编辑 李 捷)