智能答疑系统的设计与研究

徐 晓
（江苏联合职业技术学院无锡机电分院，江苏 无锡 214028）

摘 要：提出了一个高效的、科学的智能答疑系统。介绍了该系统开发的环境、分析了该系统组成的各个模块以及关键技术，最后实践证明了该系统提高了教学效率和教学手段，达到了较为理想的教学效果。

关键词：智能答疑系统；知识库；中文分词技术；RSS 技术

中图分类号：TP302.1 **文献标识码**：A **文章编号**：1674-7720(2014)05-0008-03

The research and design of intelligent question answering system

Xu Xiao
(Wuxi Electromechanics Campus of Jiangsu Joint Vocational and Technical College, Wuxi 214028, China)

Abstract: This paper deals with an efficient and scientific intelligent question-answering system. Besides, it introduces the system development environment and analyses the structure module of system function summarizing the key technique in the system. It turns out that this system has enhanced the teaching efficiency and improved the teaching means.

Key words: intelligent answering system; knowledge repository; Chinese word segmentation technology; RSS technology

随着 Internet 上远程教学普及，远程教学中的答疑成为人们关注的焦点之一。学习者从听众变成索求者，当遇到无法理解需要帮助时，及时的答疑和帮助成了必不可少的内容。在远程教学中建立智能答疑系统，可以使得学生在任何时间、任何地点都可以得到解答。教师也不必一直在线回答学生问题或重复回答相似问题。答疑系统自动回答学生的问题，一方面提高学生学习热情，提高解答效率，另一方面可以节约教师的时间，间接提高工作效率。

本文通过.net 建立一个简单高效的智能答疑系统，教师将疑难问题按一定组织方式，存放到知识和信息库中。学生提交问题时，通过中英文分词技术来分析并自动地匹配学生所提出的问题，自动地给予问题解答。当在知识库中没有搜索到信息可以采用电子邮件或是通过在线方式征求解答，有人解答后，系统自动将解答发给学生。

1 开发环境

系统采用 C# 编码，利用 VS2005+SQL2005 数据库平台开发智能答疑系统，使用 ADO.Net 实现对数据库的访问。

2 模块的设计

智能答疑系统是一个智能适应性的知识库系统，在教学设计阶段，教师将最常见的疑难问题按一定的组织方式，存放到知识库中，当学生在遇到疑难问题时，对学生以自然语言形式提出的问题进行预处理，主要是采用分词技术对问题语句进行切分处理，提取出匹配所需要的关键词，根据预先建立的基于关键词的索引结构，将答案快速定位，找出问题匹配度最高的答案。若在知识库中没有搜索到信息可以采用电子邮件或是通过在线方式征求解答，有人解答后，系统自动将解答发给学生。总之，智能答疑系统是一种支持同步和异步答疑以及讨论的系统。如图 1 所示，该系统包括以下几个模块：

(1)提问模块

学生可以使用提问模块来寻求问题的答案，这是使用系统的基本手段之一。

对于问题文本首先采用正规表达式取出中文和英文，然后采用分词技术，主要是按照一定的策略将要分析的字串与词典中的词条进行匹配来分解一系列的子串。然后在系统的知识库中以及讨论形成的材料中搜索与问题相关的材料，并按照相关程度返回结果。

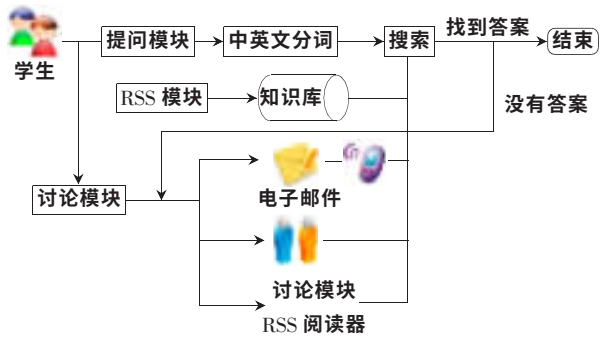

图 1 自动答疑系统的模型

(2)讨论模块

讨论模块是学生使用智能答疑系统的另外一种基本手段。用户可以参加 BBS 和聊天室等来进行讨论。

当学生在没有得到系统满意的问题解答时,系统提供了给教师发邮件模块来请求教师解答,并且提供了给教师手机发短信模块,以便提醒教师邮箱里有提示;还提供讨论模块,请求系统别的学生帮助解答;还提供了 RSS 阅读器,可以在其中查看预订的相关网上讨论社区中的资源。

(3)RSS 模块

RSS 模块可以简单地理解为一种方便的信息获取工具。RSS 获取信息的模式与加入邮件列表模式相似,无需登录到各个提供信息的站点即可自动获取。该模块主要建立一个 RSS 阅读器来订阅知识点,一旦网站上的这些知识点被更新,就会自动发送到链接源阅读器中。这样就可以不断更新和扩充知识库中的内容,如图 2 所示。

图 2 RSS 模型

3 关键技术

3.1 正规表达式

正规表达式允许快速有效地处理文本。被处理的文本小到一个电子邮件地址,大到一个多行的输入框内容。正规表达式的使用不仅允许使用一个定义模式来校验文本,而且还允许从匹配一个给定模式的文本中提取数据。

本系统使用正规表达式来取出学生问题中的中文和英文,然后执行相应的中英文分词技术,如图 3 所示。

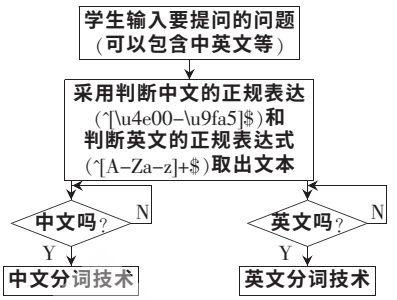

图 3 正规表达式处理流程

3.2 中文分词技术

中文分词(Chinese Word Segmentation)指的是将一个汉字序列切分成一个一个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。现有的分词算法可分为三大类:基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。基于字符串匹配的分词方法又叫做机械分词方法,它是按照一定的策略将待分析的汉字串与一个“充分大的”机器词典中的词条进行配,若在词典中找到某个字符串,则匹配成功(识别出一个词)。按照扫描方向的不同,串匹配分词方法可以分为正向匹配和逆向匹配;按照不同长度优先匹配的情况,可以分为最大(最长)匹配和最小(最短)匹配;《计算机基础》课程的智能答疑系统采用的是机械分词方法中的正向最大匹配算法。

3.2.1 分词算法

采用基于字符串匹配的分词方法,它是按照正向最大匹配法(由左到右的方向);将待分析的汉字串与一个“充分大的”机器词典中的词条进行匹配,若在词典中找到某个字符串,则匹配成功(识别出一个词)。

例如,对一个字符串 S,从前到后扫描,对扫描的每个字,从词库中寻找最长匹配。比如假设 S=“我是中华人民共和国公民”,词库中有“中华人民共和国”,“中华”,“公民”,“人民”,“共和国”……等词。当扫描到“中”字,那么从中字开始,向后分别取 1,2,3,……个字(“中”,“中华”,“中华人”,“中华人民”,“中华人民共”,“中华人民共和”,“中华人民共和国”,“中华人民共和国公”),词库中的最长匹配字符串是“中华人民共和国”,那么就此切分开,扫描器推进到“公”字。

3.2.2 数据结构

哈希表是一种高效的数据结构。哈希表最大的优点,就是把数据的存储和查找消耗的时间大大降低,几乎可以看成是常数时间;而代价仅仅是消耗比较多的内存。然而在当前可利用内存越来越多的情况下,用空间换时间的做法是值得的。另外,编码比较容易也是它的特点之一。

本系统采用哈希表(Hashtable)记录词库。首先将词典中的词进行处理,对每一个词语,如果该词语有 N 个字,则将该词语的 1,1~2,1~3,…,1~N 个字作为键,插

入相应词长度的哈希表中，而同一个键如果重复插入，则后面的值递增，如图 4 所示。

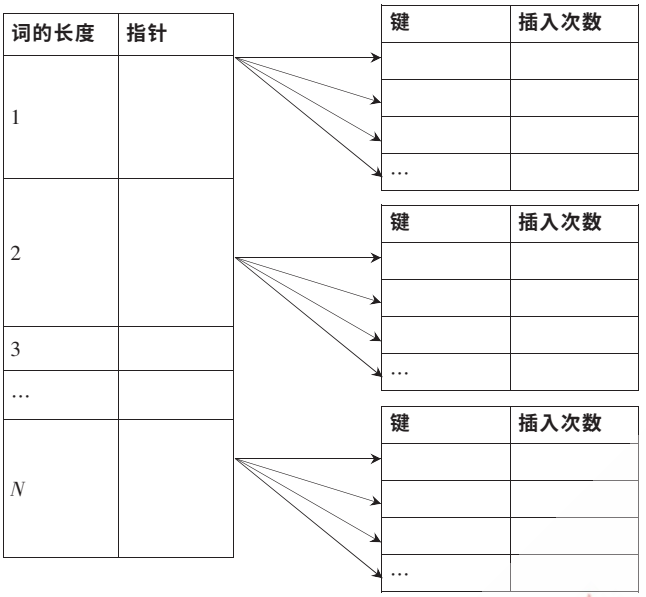

图 4 哈希表

该智能答疑系统能在一定程度上减少答案域的搜索范围，并能获得准确的答案。实践证明该智能答疑系统是有一定的智能性、主动性和方便性等特点，提高了教学效率和教学手段。

参考文献

[1] 赵成龙.一个基于 Web 的智能答疑系统的设计与实现[D].南京：东南大学，2004.

[2] 陈银凤.RSS 技术的应用和发展趋势探讨[J].内蒙古财经学院学报，2008，6(1)：98-102.

[3] 余战秋.中文分词技术及其应用初探[J].电脑知识与技术，2004，(32)：81-83.

[4] 陈挺.中文字段匹配算法[J].计算机工程，2003，29(13)：118-120.

[5] 柳泉波，黄荣怀，何克抗.智能答疑系统的设计与实现[J].中国远程教育，2000，121(8)：43-45.

(收稿日期：2013-11-13)

作者简介：

徐晓，女，1978 年生，工程硕士，主要研究方向：计算机信息管理。