基于语句相似度计算的智能答疑系统机理研究

李春生,卢鹏飞,张可佳
(东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)

摘 要:在使用互联网进行在线学习的过程中,为了解决现有答疑方式答疑实时性差、准确度低、效率低的问题,提出了一种基于语句相似度计算的智能答疑方案。首先分析现有的答疑方式及其不足;其次详细阐述了智能答疑系统的工作流程、总体结构和相关数据库结构,针对原有答疑方式检索效率低的问题加入了常用问题库,并引入基于字符串匹配的分词方法完成对学习者提出的问题的拆分;最后结合基于词信息的语句相似度计算方法对语句相似度进行计算并将结果呈现给学习者,以达到提高答疑系统的准确度、效率以及实时性的目的,满足学习者的需求。实验结果表明,基于语句相似度计算的智能答疑方案相对于原有答疑方案具有较高的准确度与效率。

关键词:分词;相似度计算;智能;答疑系统
中图分类号:TP302 **文献标识码**:A **文章编号**:1673-629X(2018)04-0091-04
doi:10.3969/j.issn.1673-629X.2018.04.0019

Research on Mechanism of Intelligent Question Answering System Based on Sentence Similarity Computation

LI Chun-sheng, LU Peng-fei, ZHANG Ke-jia
(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: In order to solve the problem of poor real-time performance, low accuracy and low efficiency of the existing question answering methods in the process of online learning with the Internet, we present an intelligent question answering scheme based on the similarity calculation of sentences. First, we analyze the existing methods of question answering and their defects, then elaborate the work flow, the overall structure and the related database structure of the intelligent answering system. The common problem database is added for the problem of low retrieval efficiency of the original question answering method, and the word matching method based on the string matching is introduced to complete the separation of questions raised by learners. Finally, the similarity calculation method based on the word information is used to calculate the statement similarity of which the result is given to the learner, so as to improve the accuracy, efficiency and real-time performance of the system and meet the needs of learners. Experiments show that the scheme has higher accuracy and efficiency than original answering scheme.

Key words: word segmentation; similarity calculation; intelligent; question answering system

0 引 言

使用互联网进行在线学习的过程中,出现了多种答疑方式,主要可以分成 3 类:基于电子邮件、BBS 的答疑方式,该方式只适合规模较小的网络课程,且实时性差;基于 QQ、微信的答疑方式,该方式是对第一种答疑方式的延伸,将答疑移到了即时通讯软件上,一定程度上提高了实时性,但只限于特定时段进行答疑;基于关键字检索的答疑方式,该方式类似于搜索引擎的检索,这种答疑方式虽然实时性强,但由于系统的智能性差导致准确度低[1-3]。

针对上述问题,文中提出一种基于语句相似度计算的智能答疑方案,解决现有答疑方式实时性差、准确度低的问题。

1 智能答疑系统机理研究

1.1 答疑流程与系统结构设计

通过分析研究,智能答疑系统应具有针对性强、智能性高的特点[4]。当学习者使用自然语言输入问题

收稿日期:2017-04-10 **修回日期**:2017-08-24 **网络出版时间**:2017-12-05
基金项目:黑龙江省自然科学基金面上项目(F2015020);黑龙江省教育科研规划重点课题(GJB1215013)
作者简介:李春生(1960-),男,博士,教授,博导,研究方向为人工智能及其应用、模式识别与人工智能;卢鹏飞(1993-),男,硕士研究生,通讯作者,研究方向为人工智能与智能系统。

后,系统会根据词库中的内容对自然语言进行分词处理,然后系统根据分词结果与问题库中存在的问题进行相似度计算,并按照语句的相似度,将计算结果返回给学习者。最终显示的结果与学习者提出的问题是否一致由学习者决定,如果学习者满意则答疑结束,如不满意则将问题交由教师进行答疑,教师答疑后将问题添加到问题库中。因此随着答疑系统的使用,问题库中的问题将会更充实,系统将会变得更加实用[5-6]。

根据需求与分析设计了智能答疑系统的系统结构,如图 1 所示。

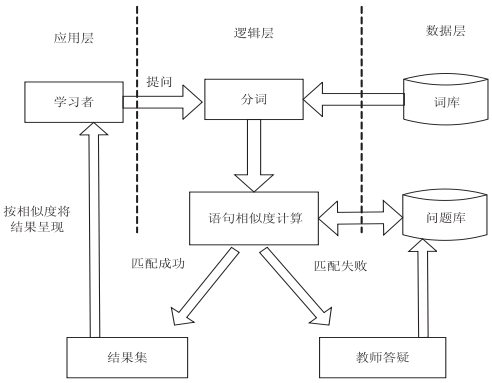

图 1 智能答疑系统结构

该系统分为三层,分别是应用层、逻辑层、数据层。应用层为学习者提供了一个友好的答疑界面,满足用户的答疑需求;逻辑层主要负责语句分词和相似度计算,为答疑的精确度提供了保证;数据层主要存放各种数据,为整个答疑系统的运行提供数据支撑[7-8]。

1.2 数据库结构设计

(1)问题库结构设计。

设计问题库的目的主要是存放一门课程所涉及到的问题,它是智能答疑系统数据层的核心。为了提高系统的效率,问题库分为常用问题库和一般问题库。当学习者输入问题后,系统先从常用问题库检索,如未检索到,再从一般问题库检索,常用问题库中存放的问题由学习者的提问频率决定[9]。问题库的结构如图 2 所示。

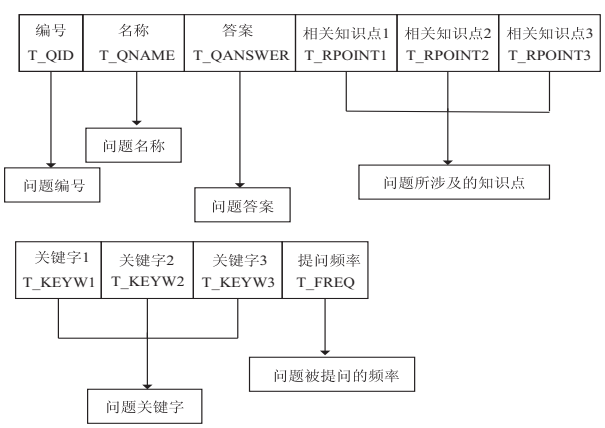

图 2 问题库结构

T_QID、T_QNAME、T_QANSWER 分别存储问题的编号、名称、答案;T_RPOINT1-T_RPOINT3 存储与该问题相关的知识点连接。如学习者检索到“数组的概念”答案后,该问题的 T_RPOINT1-T_RPOINT3 字段可能存储的就是“一维数组的定义”、“一维数组与二维数组的区别”、“数组的使用”等与该知识点相关的问题,方便学习者了解相关的知识;T_KEYW1-T_KEYW3 存储问题关键词,方便语句相似度计算时的分词。

(2)词库设计。

设计词库的主要目的是在分词中使用,词库为分词的准确提供了资源。词库分为专业词库和常用词库,两词库的表结构相同。词库的结构如图 3 所示。

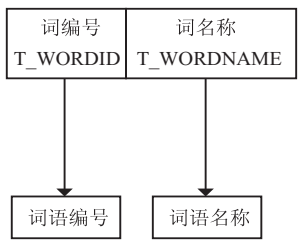

图 3 词库结构

(3)同义词库设计。

设计同义词库的目的是增加系统的准确性,使系统充分理解学习者提出的问题。同义词库的结构如图 4 所示。

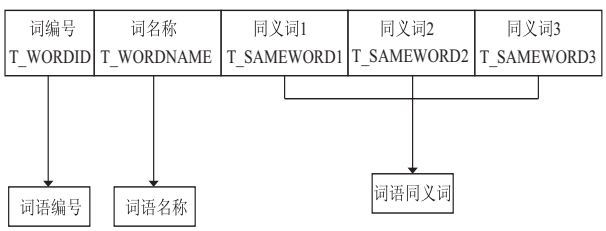

图 4 同义词库结构

在同义词库中如某个记录的 T_WORDNAME 字段存储的是“概念”,那这个记录的同义词存储的可能是“含义”、“定义”等与之相关的同义词。

(4)学习者问句库设计。

设计学习者问句库的作用是存储学习者提出而问题库中不存在的问句,教师答疑后由教师决定是否将问题加入到问题库中。学习者问句库的结构如图 5 所示。

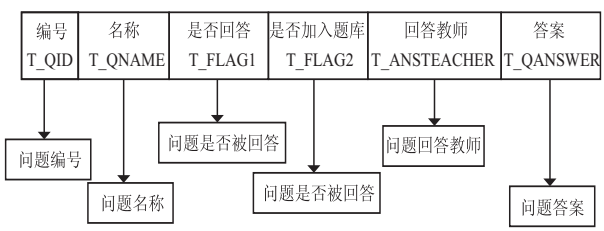

图 5 学习者问句库

2　智能答疑系统的关键技术

目前常用的分词方法主要有三种,分别是基于理解的分词方法、基于统计的分词方法和基于字符串匹配的分词方法[10-11]。通过分析与研究,发现学习者提出的问题具有专业性强、语句短小的特点[12]。出于对提问问题的特点以及效率等多方面因素的考虑,该系统采用正向最大匹配算法实现分词。正向最大匹配算法的设计思想如下:用 MAXLEN 表示词库中最大词长,从词库中查找长度为 MAXLEN 的词,按照从左至右的顺序,与句长为 MAXLEN 的子串进行匹配,若匹配成功则切分出字符串,指针后移 MAXLEN 位继续匹配,若失败 MAXLEN 长度减一位后继续匹配,直至匹配的字符串长度为 2;若还未匹配成功,则将当前汉字切割成词,然后后移一位继续匹配[13-14]。如:“C#数据类型的种类”,分词的处理步骤如下:

Step1:从专业词库中匹配出“C#”,“数据类型”。

Step2:对剩余的字串“的种类”进行分词,从常用词库中匹配出“的”,“种类”。

所以最后的分词结果为“C#”,“数据类型”,“的”,“种类”。

(1)词形相似度计算。

词形相似度计算的原理是通过计算两个句子中相同词的个数来比较相似度。它的核心思想是如果两个句子相同的词数越多,那这两个句子的相似度就越高。计算过程是将两个句子 L_1、L_2 分词后存放在数组 $ArrayL_1$ 和 $ArrayL_2$ 中,经过计算后得到的两个句子中相同词的个数用 SameWorld 表示。如果有相同的词重复出现,则取最小的数。len(L_1)表示句子 L_1 分词后词的个数,词形相似度计算的公式如下:

$$SimilarS(L_1,L_2)=\frac{SameWorld}{\max(len(L_1),len(L_2))} \quad (1)$$

其中, $SimilarS(L_1,L_2)\in[0,1]$。

例如:L_1=一维数组的概念是什么,L_2=什么是一维数组。

分词后的结果为:$ArrayL_1$={“一维数组”,“的”,“概念”,“是”,“什么”},$ArrayL_2$={“什么”,“是”,“一维数组”}。$len(L_1)=5$,$len(L_2)=3$, $SimilarS(L_1,L_2)=3/5=0.6$。

(2)句长相似度计算。

句长相似度计算是通过计算两个句子的长度来比较两个句子的相似度,它的核心思想是如果两个句子的长度越相近,那么这两个句子的相似度越高。len(L_1)表示句子 L_1 单字的个数,句长相似度计算公式如下:

$$SimilarLen(L_1,L_2)=1-\frac{abs(len(L_1)-len(L_2))}{len(L_1)+len(L_2)} \quad (2)$$

其中,$SimilarLen(L_1,L_2)\in[0,1]$。

例如:L_1=一维数组的概念是什么,L_2=什么是一维数组。

$len(L_1)=10$,$len(L_2)=7$,$SimilarLen(L_1,L_2)=1-abs(10-7)/17\approx0.82$。

(3)字形相似度计算。

字形相似度计算的原理是通过计算两个句子中相同字的个数来比较相似度。它的核心思想是如果两个句子相同的字数越多,那这两个句子的相似度就越高。它是对词型相似度的一个补充,词型相似度的准确性取决于分词的结果,而分词的准确性依赖于词库的词,如果词库的词不完整,那么词型相似度的结果就会大打折扣。因此利用字形相似度计算来弥补词型相似度计算存在的一些不足。字形相似度计算的过程类似于词形相似度。首先将两个句子 L_1、L_2 的单字拆分后存放在数组 $ArrayL_1$ 和 $ArrayL_2$ 中,计算出的两个数组中相同字的个数用 Same 表示,如果有相同的字重复出现,则取最小的数。$len(L_1)$ 表示句子 L_1 中字的个数,字形相似度计算的公式如下:

$$Similarchar(L_1,L_2)=\frac{sanme}{\max(len(L_1),len(L_2))} \quad (3)$$

其中,$Similarchar(L_1,L_2)\in[0,1]$。

例如:L_1=一维数组的概念是什么,L_2=什么是一维数组。

以上两句话中相同的字数为 7 个,$Similarchar(L_1,L_2)=7/10=0.7$。

(4)语句相似度计算。

语句相似度的计算公式为:

$$\begin{aligned}Similar(L_1,L_2)=&Q_1*SimilarS(L_1,L_2)+\\&Q_2*SimilarLen(L_1,L_2)+\\&Q_3*Similarchar(L_1,L_2)\end{aligned} \quad (4)$$

其中,Q_1、Q_2、Q_3 表示各个相似度计算的子项所占整个语句相似度计算的权重,且 $Q_1+Q_2+Q_3=1$。设 $Q_1=0.4$,$Q_2=0.2$,$Q_3=0.4$。则 L_1 与 L_2 最终相似度为:

$$\begin{aligned}Similar(L_1,L_2)=&0.4*0.6+0.2*0.82+\\&0.4*0.7\approx0.68。\end{aligned}$$

3　实验结果及分析

由于一直以来关于智能答疑系统的评测没有一个统一标准,只能通过大量的实验来检验系统的准确率。以《C#程序设计》课程为例,把《C#程序设计》课程的问题输入到系统中,从两个方面测试系统,使用基于关键字的检索方法和基于相似度计算的检索方法对比测试系统的答疑准确度[15]。使用带常用问题库的答疑系统与不带常用问题库的答疑系统对比测试系统的答

疑效率。

准确度计算公式如下：

准确度 = 答对题数/总题数 (5)

检索效率的计算公式如下：

检索平均时间 = 检索总时间/检索总题数 (6)

实验结果如表 1 和表 2 所示。

表 1 准确度对比

问题个数	基于关键字的检索方法	基于相似度计算的检索方法
30	0.49	0.65
60	0.41	0.61
90	0.52	0.68

表 2 效率对比 s

问题个数	基于关键字的检索方法	基于相似度计算的检索方法
30	0.24	0.17
60	0.29	0.21
90	0.32	0.28

实验结果表明，基于关键字的检索方法与基于相似度计算的检索方法相比，后者的准确度更高。带常用问题库的答疑系统与不带常用问题库的答疑系统相比，前者的检索效率更高。但由于受制于词库的限制，有的词不能被正确划分，直接影响了答疑效果。因此，接下来主要的目标是如何提高分词的正确率。

4 结束语

为解决现有答疑方式答疑实时性差、准确度低、效率低的问题，提出一种基于语句相似度计算的智能答疑系统的实现方案。测试结果表明，相对于原有答疑方式，该答疑系统在实时性、准确度、效率方面都有一定的提高，一定程度上满足了学习者的需求。

参考文献：

[1] 康文宁，杨志强. 相似度计算在智能答疑系统中的研究及应用[J]. 计算机技术与发展，2010，20(2)：71-74.

[2] KANAAN G，HAMMOURI A，SHALABI R A，et al. A new question answering system for the Arabic language[J]. American Journal of Applied Sciences，2009，6(4)：797-805.

[3] 付春捷，胡 萍. 基于中文分词的智能答疑系统的设计[J]. 科技视界，2014(14)：14.

[4] 郑晓洁，张 琳. 一个基于多课程本体的简单答疑系统[J]. 现代计算机，2013(16)：53-57.

[5] SEENA I T，SINI G M，BINU R. Malayalam question answering system[J]. Procedia Technology，2016，24：1388-1392.

[6] 肖坤峨，虞 泉. 基于 WEB 的智能答疑系统的研究与构建[J]. 软件，2015，36(6)：31-36.

[7] 陈品帆. 基于 Web 的远程答疑系统的关键技术探讨[J]. 信息技术与信息化，2015(2)：52-53.

[8] 李攀飞，敖永红，叶昭晖，等. 基于大规模在线学习平台的智能答疑系统研究与设计—以“教育技术”MOOC 为例[J]. 工业和信息化教育，2015(11)：33-37.

[9] LIU Y F，HE C. Design and implementation of question and answer system based on mobile phone platform[J]. Advanced Materials Research，2014，1049-1050：1977-1980.

[10] 徐 晓. 智能答疑系统的设计与研究[J]. 微型机与应用，2014，33(5)：8-10.

[11] 郭文俭. 基于课程教学网站的智能答疑系统的设计与实现[D]. 长春：吉林大学，2015.

[12] 赵 静，党丽琼. 基于自然语言理解的在线答疑系统设计与实现[J]. 计算机时代，2015(5)：10-12.

[13] ZHANG P Y. Sentence similarity metric and its application in FAQ system[J]. Advanced Materials Research，2013，718-720：2248-2251.

[14] 管 燕，仲兆满. 基于课程知识本体的 FAQ 库自动生成方法研究[J]. 中国远程教育，2010(13)：68-72.

[15] DONG W J，GENG G H. Research and implementation of intelligent question answering system in MOOC[J]. Applied Mechanics and Materials，2014，678：639-643.

(上接第 90 页)

选择方法[J]. 计算机科学，2016，43(10)：225-228.

[6] YANG Y. An evaluation of statistical approaches to text categorization[J]. Information Retrieval，1999，1(1-2)：69-90.

[7] 樊存佳，汪友生，王雨婷. 一种改进的 CHI 文本特征选择方法[J]. 计算机与现代化，2016(11)：7-11.

[8] 裴英博，刘晓霞. 文本分类中改进型 CHI 特征选择方法的研究[J]. 计算机工程与应用，2011，47(4)：128-130.

[9] 闫 屹，张燕平，耿筱媛. 基于 CHI 值特征选取和覆盖的文本分类方法[J]. 计算机技术与发展，2008，18(5)：79-81.

[10] 吕 锋，王 虹，刘皓春，等. 信息理论与编码[M]. 北京：人民邮电出版社，2004.

[11] 邱云飞，王 威，刘大有，等. 基于方差的 CHI 特征选择方法[J]. 计算机应用研究，2012，29(4)：1304-1306.

[12] JIANG X Y，JIN S. An improved mutual information-based feature selection algorithm for text classification[C]//5th international conference on intelligent human-machine systems and cybernetics. [s. l.]：IEEE，2013：126-129.

[13] DING X，TANG Y. Improved mutual information method for text feature selection[C]//8th international conference on computer science & education. [s. l.]：IEEE，2013：163-166.

[14] 熊志斌，刘 冬. 朴素贝叶斯在文本分类中的应用[J]. 软件导刊，2013，12(2)：49-51.

[15] 李荣陆. 文本分类及其相关技术研究[D]. 上海：复旦大学，2005.