远程教育中自动问答系统相关技术探究

杨亚菲
（国家开放大学 信息化部，北京 100039）

摘 要：远程教育已经成为我国终身教育体系的一个重要组成部分。答疑解惑是远程教学过程中决定教学质量的重要因素，而自动问答系统可以在少量人力的参与下及时且准确地定位问题并做出解答。文章介绍了国内自动问答系统的发展现状和研究意义，分析了自动问答系统的问题分析、信息检索和答案抽取以及各部分使用的关键技术，最后阐述了现代远程教育自动问答系统的研究方向。

关键词：远程教育；自动问答；问题分析；信息检索；答案抽取
中图分类号：G434；TP391 文献标志码：A 文章编号：1673-8454(2018)20-0094-03

随着计算机和网络等相关技术的成熟发展，远程教育已在教育界起到越来越重要的作用，而问答系统在远程教育平台中可以起到及时解决学生的疑问以方便其学习的作用。但目前随着远程学习学生规模的增长，现有的问答方式无法及时解答学生问题。基于此现状，本文介绍了自动问答系统并探讨了涉及到的相关技术。

一、自动问答系统国内现状

1.自动问答系统

自动问答系统（Automatic Question and Answering System，简称“问答系统 QA”），是对于用户使用自然语言描述的问题，基于大量非同构数据自动搜索出简洁且准确答案的信息检索系统。问答系统在及时解决学生在学习过程中所遇问题的同时，还促进了学生学习的积极性，此外可以使教师不用将大量精力花费在重复回答相似问题上，而是集中于教学的改革和研究。

2.国内远程教学中的问答系统

目前我国在远程教育领域的问答系统主要分为以下三种：

（1）没有独立的问答部分，教学中的交互只能通过使用电子邮件、留言板或聊天室等简单方式进行。这种远程教育系统可以视为电视大学的网络版，没有展现网络教学的优点。

（2）具有简单问答方式的问答系统，这种系统类似于 BBS 形式，为师生提供交互环境，或是在教师的主导下以线上聊天的方式进行交互，这种方式比较粗糙地实现了师生之间的互动，但不一定是及时的。

（3）采用比较复杂的技术在某种程度上实现自动的问答系统。这类系统减少了教师参与，缩短了问答互动延时。根据采用技术不同大致分为三种类型：①基于 FAQ 库的智能答疑系统。基于常见问题库（Frequently Asked Question，FAQ）的 QA 是指将常见问题与对应答案存储到常见问题库，系统使用自然语言处理技术分析问题并抽取出关键词，然后在 FAQ 库中匹配和提取最优项反馈给用户。②基于全文检索的问答系统。这种系统搜索答案的范围是相关文档。系统利用自然语言处理技术分析用户问题，然后在文档库中使用信息检索技术搜索文档并按照查询相似度排序文档，最后提取出与问题相似度较高的文档返回给提问者。③面向知识自动化的问答系统。这种系统利用知识自动化的方法对虚拟空间的大数据进行深度开发和智力挖掘，以有效解决不确定、多样且复杂的问题。

二、自动问答系统研究意义

现代远程教育是我国终身教育体系中至关重要的一个组成部分。根据现代远程教育的特点，它不再是使用以往教育中教师灌输知识、学生被动学习的方式进行教学，而是在教学过程中更侧重以学生为核心的自主学习，因此，及时回答疑问是决定远程教育教学质量的必要环节。然而在远程教学过程中，教师和学生在时间和空间上的分离导致难以实现师生一对一实时互动，所以需要问答系统解决学生在自主学习过程中遇到的疑难问题，这使得问答系统对远程教育质量高低起着决定性作用，因此，研究问答系统对我国现代远程教育的发展具有深远的意义。

三、自动问答系统实现关键技术

无论采用何种方式进行分类，QA 系统架构通常包括三个主要过程：问题分析、信息检索和答案抽取。具体流程

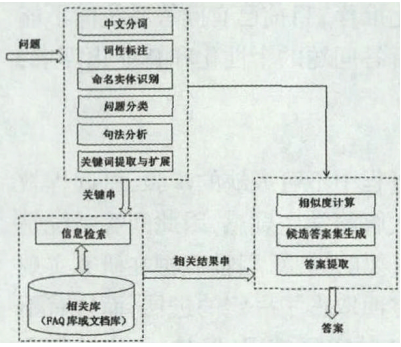

图 1 问答系统基本体系结构

为首先对用户提出的问题进行自然语言分析处理；然后在相关库中对关键词进行检索，得到问题答案候选集；最后使用问题答案候选集对问题进行相似度计算以提取出最优答案。其基本体系结构如图 1 所示。

1.问题分析

问题分析是指对用户提出的问题进行预处理和文本分类，其中用到的关键技术有中文分词、命名实体识别、词性标注、问题分类、句法分析以及关键词提取与扩展等。这些涉及到的自然语言处理各个细分部分的性能都直接或间接影响着整个问答系统的性能。下面分别介绍以上涉及的关键技术。

中文分词是将汉字序列切分成单个独立的词。例如对句子“如何理解会计系统设计内部控制的方向？”进行中文分词的结果为 “如何 理解 会计 系统 设计 内部控制 的 方向”。近乎所有涉及到中文文本处理的任务都在中文分词的基础上完成，因为在中文信息处理中，一组单词通常被用作最小处理单位。常见的分词技术有基于字符串匹配法、基于统计原则和基于理解的方法。

命名实体识别即专名识别，用于识别文本中具有特定含义的对象，主要是人名、地名、组织名、专有名词等。在具有问句的文本中命名实体基本具有可以区别其它文本信息特殊的含义，因此识别命名实体不仅利于后续信息检索性能的提高，而且在抽取答案时给相似度计算提供较好的特征。

词性标注又称词类标注或标注，用于标注分词结果中每个单词的正确词性，以确定每个单词的词性是名词、动词、形容词或其它词性。词性标注是自然语言处理领域中许多任务必不可少的步骤，例如句法分析、文本分类、信息抽取以及语音识别等。常见的词性标注方法包括基于统计模型的注释方法、基于规则的注释方法、统计方法与规则方法相结合的注释方法。

问题分类是指通过确定问题的目标答案的类型来为随后的答案抽取提供语义限制。问题分类可以缩小候选答案搜索空间，以提高定位答案的准确性。因此，问题分类对提高问答系统的性能方面起到关键性作用。常用分类方法有贝叶斯分类方法、支持向量机、最大熵等。

句法分析是分析句子的词语语法。在对句子中的单词串进行句法分析之后，会构造出一个解释句子语法结构的句法分析树。对文本的预处理仅限于句子中词及其相关属性级别，则不能分析出句子中词与词之间以及句子与词之间的关系。而句法分析可以准确揭示认识对象的结构特征并迅速把握语义，所以本文预处理过程中句法分析必不可少。

关键词是指在某种程度上可以代表句子主要含义的词或词组。在对句子进行文本处理的任务中，通过提取关键词可方便理解句子的主要语义，为后续信息检索和答案提取操作降低难度。常用的关键词提取方法可分为有监督学习算法和无监督学习算法。此外关键词需要扩展，因为在不同问题和答案中具有相同含义的关键词或许具有不同的词型，如果不扩展关键词的词型，可能会降低定位答案的准确性。

2.信息检索

信息检索是通过相应的检索技术将提取出的关键词在相应的知识储备数据库中进行信息检索，得到候选问题集；信息检索的本质是将描述用户所需信息的问题特征与存储在信息中的检索标识符进行比较，并找到与问题特征一致或基本一致的信息。当前的问答系统无论是基于知识库检索问题答案，还是基于全文相关文档检索信息，都需要对相关信息数据创建索引，然后搜索索引以获取信息检索的结果。关于信息检索目前有多种检索模型，比如布尔模型、向量空间模型以及概率检索模型。对于布尔模型，文档与用户查询由其包含的单词集合来表示，这种模型简单直观但无法结合数据的相关性，而且其搜索结果也不细致。向量模型把文档看成一个向量，将查询也表示成同一高维空间的向量，计算给定的查询向量和每个文档向量的 tf-idf 作为权重值进行检索。对于概率模型，排序文档与用户查询相关的概率作为最优处理方式进行检索。以上三种模型中，布尔模型计算速度比其它两个模型快，但是性能低于后两个模型。

3.答案提取

答案提取处理过程中相似度计算是关键技术。答案提取的主要任务是对问题与候选集中的问题进行文本分析以及相似度计算提取最终答案，相似度计算可以从关键词词类型相似度、句子长度相似度和关键词词序相似度等不同角度计算用户问句和候选问题集中问题之间的相似度，若计算的相似度值大于预先设定的阈值，则检索的结果满足预期，便得到与用户输入问题最为相似的问题，抽取其对应的答案返回给用户。其中关键词词类型相似度使用句子中关键词的相同类型个数来度量，并且相同的个数越多，相似度越高。句子长度相似度

是使用句子的长度来衡量，句子之间长度相差越小，相似度越高。关键词次序相似度是使用关键词在句子中的位置来衡量，关键词的位置越相近，相似度越高。

四、现代远程教育中自动问答系统研究难点分析

1.中文领域问答系统研究

目前远程教育方面比较先进的问答系统是在英文环境下研究与开发的，而且英文环境下提供了大量先进技术和资源使用，而较少有研究涉及到其它语言包括中文领域。而且中文的语言结构比较多样，相同问题因句子语境的不同可能表达的含义不一致，使得问答系统处理的数据源具有一定复杂性。另外，问答系统的整个处理过程需要很多步骤，每个步骤都需要改进算法提高性能，提高最终结果的准确性。因此，中文领域问答系统的研究仍面临诸多挑战，我们可以借鉴国外的技术和成果，但将此领域国外相关技术应用于中文领域仍需进一步努力以达到最佳适用度，这就需要我们加强对比与分析相关工具与方法，进而找到更好应用于远程教育方面的中文领域问答系统。

2.与其它系统关联的问答系统研究

现代远程教学信息日益增多，其中除了包括结构化数据之外，还包括大量的半结构化数据以及非结构化数据，数据结构的复杂性使得问答系统在远程教育信息资源整合方面存在困难。而且就目前国内远程教育方面自动问答系统的研究现状看，问答系统很少与其它相关系统如教务系统、考试系统、学习系统等进行关联，导致无法利用相关系统中有价值的数据更好地进行有针对性的回答、扩展数据来源以及增加数据内容丰富性。所以在信息整合和推理方面的方法和技术并不成熟，对问答系统与其它系统的关联研究还有很大的发展空间。我们可以在将问答系统与远程教学中相关系统建立关联的方向多做尝试，使关联系统的相关数据为问答系统服务，以提高其定位答案的准确率。

3.满足现代远程教育个性化与智能化需求的问答系统研究

现代远程教育主要是针对相关领域相关专业相关课程的知识进行的学习，根据现代远程教育的特点，适用的问答系统应该具有实时性、准确性以及正确性，可以达到能快速且准确解答学生疑问的目的。但随着教育行业相关政策的推行，问答系统已经不能满足当前教育形式的发展需求。除了以上基本要求外，问答系统尤其需要具备个性化推荐功能，通过对学生的课程学习情况、科目考试情况等学习行为信息分析，并从中挖掘出有价值的数据构建针对学生个性特点的学习模型，进而实现相关问题的个性化推荐。目前已有问答系统尚不能满足现代远程教育对解答问题的个性化和智能化需求，仍需进一步研究。

五、结束语

答疑解惑是学习过程中不可或缺的步骤，而远程教学中问答系统可以高效解答学生疑惑，因此问答系统是远程教学体系中极其重要的模块。目前我国在研究实现自动问答系统的技术方面取得了一定的进展，而且有不少领域已经实现了系统的实际应用，但是，将比较契合现代远程教育的问答系统投入使用需要更多努力。本文对远程教育领域问答系统的相关技术进行了阐述，并就目前的研究难点进行了分析，希望对相关研究者有一定的启示和帮助。

参考文献：

[1]刘里，曾庆田.自动问答系统研究综述[J].山东科技大学学报(自然科学版),2007(4):73-76.

[2]李爽，陈丽.国内外网上智能答疑系统比较研究[J].中国电化教育,2003(1):80-83.

[3]江耿豪.基于 FAQ 的自动答疑系统的设计与实现[J].计算机时代,2009(12):39-41.

[4]曾帅，王帅，袁勇等.面向知识自动化的自动问答研究进展[J].自动化学报,2017,43(9):1491-1508.

[5]邓实福，刘挺，秦兵等.问答系统综述[J].中文信息学报,2002,6(16):46-52

[6]张黎，徐蔚然.中文分词研究[J].软件,2012,33(12):103-108.

[7]孙镇，王惠临.命名实体识别研究进展综述[J].现代图书情报技术,2010(6):42-47.

[8]江会星.汉语命名实体识别研究[D].北京:北京邮电大学,2012.

[9]Ma J,Xiao T,Zhu J,et al.Easy-First Chinese POS Tagging and Dependency Parsing[C].COLING,2012:1731-1746.

[10]Ma J, Zhu J, Xiao T, et al. Easy-First POS Tagging and Dependency Parsing with Beam Search[C].Meeting of the Association for Computational Linguistics. 2013:110-114.

[11]孙宏林，俞士汶.浅层句法分析方法概述[J].当代语言学,2000(2):74-83+124.

[12]郑丁山.基于 moodle 平台答疑系统的设计与实现[J].计算机光盘软件与应用,2013,16(9):101-103.

[13]康毅.面向客服的自动问答系统关键技术研究[D].沈阳:东北大学,2014.

[14]王正华，韩永国.自动问答系统设计与实现[J].软件导刊,2014,13(9):111-113.

（编辑：鲁利瑞）