doi:10.13582/j.cnki.1674-5884.2018.01.013

基于 Lucene 的在线答疑系统的构建

赵肆江[a,b],李锋[b]

(湖南科技大学 a. 知识处理与网络化制造湖南省普通高校重点实验室;b. 计算机科学与工程学院,湖南 湘潭 411201)

摘　要:通过分析在线答疑系统的交互模式,提出自助答疑和交互式答疑无缝衔接的答疑流程和框架;针对自助答疑过程中信息模糊检索的匹配问题,提出基于 Lucene 的全文检索方法,阐述了中文分词、主题词提取、索引建立和问题检索等关键技术。实验表明,该方法能有效地检索相关答疑信息,提升答疑效率。

关键词:在线答疑;Lucene;全文检索

中图分类号:G434　　**文献标志码**:A　　**文章编号**:1674-5884(2018)01-0066-04

网络教学和辅导作为一种新的学习模式(如 MOOC[1])在现代教育技术中占据重要地位,成为课堂面对面教学等传统教学方式的重要补充。其中基于网络的在线答疑系统拓展了学习的空间,丰富了学习内容和种类,为教师和学生提供课内和课外知识分享、经验交流和解决疑难的平台,成为传统教学和网络教学的重要组成部分和补充[2-3]。在线答疑系统中一个重要的功能是对提出的问题进行检索,常见数据库中的检索方式为基于关键词的精准查询和模糊匹配查询,这两种方式最大的缺点是难以检索和关键词不完全匹配的内容,难以根据语句与段落内容进行检索,且难以根据检索结果与被检索内容的相关度进行排序。因此,在线答疑系统中需要解决如何有效地对检索的问题进行全文检索并排序等问题。

Apache 软件基金会(Apache Software Foundation,ASF)旗下的 Lucene 是一个开源的全文搜索引擎工具包[4]。Lucene 具有良好的系统架构,并具有跨平台、性能优异和易用性等优点,非常适用于海量文本数据的全文检索。国内外很多学者研究将 Lucene 应用于文本分类、全文检索、图像检索等领域[5];相对于英文来说,中文分词更具复杂性,多位学者针对中文检索领域研究和改进 Lucene[6]。因此,本文采用 Lucene 技术实现在线答疑学习平台,实现问题的全文检索功能,并按照检索问题的相关度对检索内容进行排序,从而让提问者容易从已有题库中找到所需答案,达到减少重复提问、提高答疑效率的目的。

1　网络答疑流程分析

根据陈丽提出的网络教学交互模型,交互主要包括 3 个层面:学生与媒体界面、学生与教学要素和学生的概念与新概念[7]。信息交互的实体主要在于学生与学生、学生与教学资源和学生与老师之间,如图 1 所示。

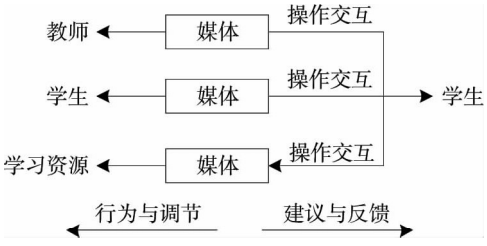

图 1　教学交互模型

收稿日期:20170816

基金项目:湖南省自然科学基金面上项目(2017JJ2081);湖南省教育厅科学研究项目(17C0646);湖南科技大学 2016 年教学研究与改革项目(G31607);2016 年湖南省普通高等学校教学改革研究项目(406);2017 年湖南省普通高等学校教学改革研究项目(252)

作者简介:赵肆江(1980-),男,湖南邵阳人,讲师,博士,主要从事众源地理信息、地理信息系统建模等研究。

在线答疑模式按照信息交互的方式可分为自助式答疑和交互式答疑。自助式答疑指学生自主地在系统问题库中搜索所需信息而获取知识;交互式答疑指通过学生与老师、学生与学生之间相互讨论获取知识的过程,交互的方式可分为线上和线下两种。

下面分析网络答疑流程,讨论如何将自助式答疑和交互式答疑 2 种方式无缝衔接。网络答疑流程中的提问过程如图 2 所示,提问一般由学生发起,学生针对自身难以理解和掌握的内容提出相关问题。为了提升答题的效率,首先采取自助式答疑,根据学生提出的问题搜索库中包含的类似问题,并根据问题的相似度进行排序。学生依次查看相似问题,如果找到满足要求的答案,则可以进行评价,并结束提问过程;否则,转而进行交互式答疑,交互式答疑首先要求学生提交问题、设定提问对象(可以为所有的教师和同学,也可针对具体的某位教师)和回答方式(在线回答或电子邮件等),然后提醒相关教师对问题进行解答;最后教师解答问题并可与学生进行在线交流,并将相关内容纳入问题库,以备用于后续的提问检索。

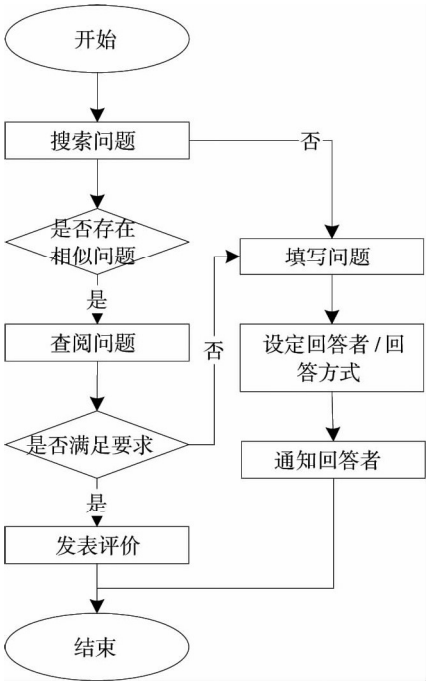

图 2　自助模式和交互模式相结合的提问过程

上述过程中,自助式答疑模式效率高,快速直接地解答学生的提问,其缺陷是题库的完整性建设以及如何快速有效地检索相似问题。交互式答疑模式的优点是回答问题的覆盖面广,交互性强,可以彻底消除学生学习过程中的疑问,丰富课间与课外知识;但相对自助式答疑而言效率较低,问题难以得到马上解答。网络在线答疑应该将二者有机结合起来,尽量满足学生的提问需求,以达到课外教学辅导的目的。

2　基于 Lucene 的答疑系统的关键技术

自助式答疑根据学生所提问题智能检索相关问题,是在线答疑系统中的重点和难点,因此本文主要介绍自助式答疑部分。根据前一节中的分析可知,题库建设和相似问题的检索是自助式答疑中需要解决的两个基本问题。相似问题检索不仅在学生提问排序中起到重要作用,同时在题库建设过程中也有非常重要的作用,尤其是在解决题库中问题冗余的问题上,一方面,在学生提问的时候可以直接查看相似问题,避免了重复提交相似问题,造成冗余;另一方面,在题库周期性质量检测过程中,可以检测出相似题目,去除一些提问和回答质量不高的问题。

为了解决传统检索方法难以检索与关键词不完全匹配的内容和难以根据检索结果相关度进行排序的问题,需要实现全文检索,因此,本文采用 Lucene 实现基于全文检索的问题检索。Lucene 是 ASF 旗下的开源文本搜索引擎,提供了全文的检索引擎和索引引擎,可以非常方便地嵌入到需要全文检索功能的各种应用当中。下面依次从中文分词、索引和排序等方面介绍在线答疑系统中全文检索的技术实现。

2.1　中文分词技术

分词技术是全文检索的核心技术之一,中文分词是实现问题模糊检索和排序的关键。Lucene 中提供的分词器中主要有 2 个中文分词器:CJKAnalyzer 和 ChineseAnalyzer,但二者采用简单的一元或二元分词方式,导致分词准确性不够,且容易产生大量的无用词条,性能难以保证。针对其缺陷,产生了与 Lucene 兼容的第三方分词器:IK_CAnalyzer,PanGuAnalyzer,MMAnalyzer(JE 分词)和 PaodingAnalyzer 等[5],这些算法一般采用基于词典的分词算法,提升了分词的准确性,提升了效率。

根据周敬才等人的研究[5],从分词效果和性

能等方面综合考虑，本文在 Lucene 中集成 PaodingAnalyzer 分词算法。

2.2 索引

索引是检索得以高效执行的关键，通过索引可以快速有效地查询数据库中存储在索引中的字段，因此是搜索引擎的核心。Lucene 采用倒排索引机制（Inverted index）实现索引，该机制存储全文中提取的关键词和关键词在多个文档中存储位置的映射。因此，索引表中存储词语集合（Terms）和文档集合（Documents）之间的对应关系，每个词语均对应 1 个或者多个链接，这些链接对应包含该词语的文档。当检索某个词语的时候，可以通过索引表快速地检索到相对应的文档。设词语集合 $\boldsymbol{Terms} = \{t_1, t_2, \cdots, t_n\}$，文档集合 $\boldsymbol{Documents} = \{d_1, d_2, \cdots, d_m\}$，倒排索引的典型组织方式如图 3 所示，其中 d_{ij} 与 Documents 集合中的某一个文档对应，在倒排表中表示第 i 个词语 t_i 所对应的文档，该词语对应的文档数量为 $j(j \geq 1)$，一般而言，文档信息应该包含关键词的词频和位置等信息。

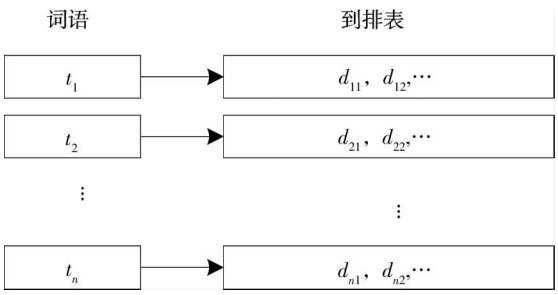

图 3 Lucene 的倒排表

2.3 排序

前述索引表能实现对单个词语（Term）进行检索，并可以根据词语出现的频率实现简单的相关度排序。但是事实上，检索的关键词往往不会局限于单个词语，而是根据几个关键词、一句话、甚至整个问题进行全文检索。此时，仅有前述的倒排索引表难以实现检索和对检索结果的排序。倒排表中还需存储答疑问题的多个主题关键词，然后根据这些关键词和检索语句的关键词进行匹配，再根据相似度进行排序，方可满足检索要求。本文采用词频逆向文件频率（Term Frequency – Inverse Document Frequency，TF – IDF）技术提取主题词，然后对检索结果进行排序。TF – IDF 是一种常用于信息检索与处理的文本处理技术，主要用于评估字词在文件中的重要程度（也称为主题词），其基本原理为字词的重要性与其在文本中出现的频率成正比，但是与其在语料库中出现的频率成反比。

具体的排序过程如图 4 所示，首先对待检索问题进行分词处理；得到问题的分词结果后进行 TF – IDF 处理，得到问题的主题词；然后根据主题词采用余弦相似度与数据库中的问题进行匹配；最后按照匹配程度进行排序。

2.4 问题索引建立和检索

在线答疑系统中，问题索引的建立和检索过程如图 5 所示，首先将数据库中已有的问题通过分词、TF – IDF、倒排索引等技术创建索引，形成倒排索引文档；然后对待检索的问题采用分词和 TF – IDF 技术提取主题词；根据主题词确定 n 篇数据库中的近似文档；最后采用余弦相似度对问题主题词和 n 篇文档进行相似度计算，根据相似度的高低对检索结果进行排序。

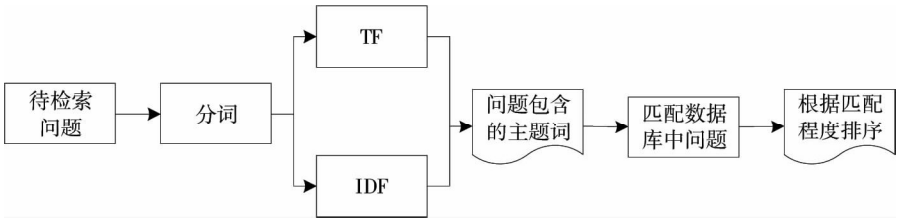

图 4 问题检索排序

3 结语

在线答疑是网络教学和辅导这种新的学习模式的重要组成部分，本文提出了自助答疑和交互式答疑无缝衔接的答疑流程和框架，在自助答疑方面实现了基于全文索引的在线答疑系统，能快

速有效地查询数据库中已经存在的问题,并能根据待查问题的相关度排序,便于提问者快速找到自己所需问题的答案,提升了答疑的效率。

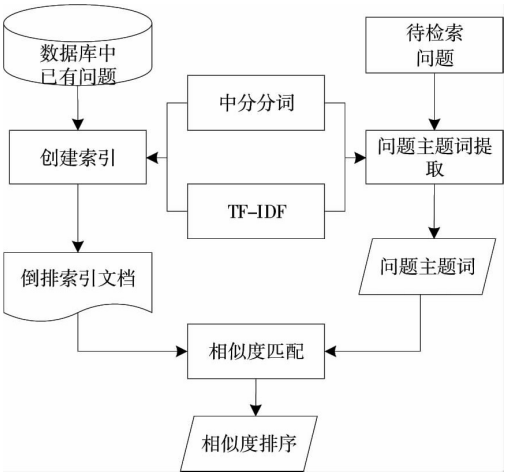

图 5　问题索引的建立和检索过程

参考文献:

[1] 杨敏, 杨洁. 慕课:班级授课制的终结者[J]. 当代教育理论与实践, 2014(5): 71 – 73.

[2] 张银. 答疑系统学习机制的分析与思考[J]. 中国远程教育, 2006(2): 36 – 38.

[3] 常耀龙. 基于 JFinal 框架的校内课程在线答疑系统的设计与实现[D]. 长沙:湖南大学, 2016.

[4] 郁红英, 高英. 基于 Lucene 的网络学习智能答疑系统的设计与实现[J]. 软件导刊, 2012(1): 80 – 81.

[5] 周敬才, 胡华平, 岳虹. 基于 Lucene 全文检索系统的设计与实现[J]. 计算机工程与科学, 2015(2): 252 – 256.

[6] 张文元, 周世宇, 谈国新. 基于 Lucene 的地名数据库快速检索系统[J]. 计算机应用研究, 2017(6): 1756 – 1761.

[7] 陈丽. 远程学习的教学交互模型和教学交互层次塔[J]. 中国远程教育, 2004(5): 24 – 28.

On Construction of Online Question Answering System Based on Lucene

ZHAO Yijiang[a,b], LI Feng[b]

(a. Key Laboratory of Knowledge Processing and Networked Manufacturing;
b. School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201, China)

Abstract: A question – answering process and framework featured by seamless connection between the self – service question answering and interactive question answering are proposed by analyzing the interactive mode of the online question answering. To solve the matching problem of fuzzy message retrieval in the process of self – service question answering, a full text retrieval method based on Lucene is put forward. Meanwhile, the key technologies, such as Chinese segmentation, thematic word extraction, index construction and question retrieval, etc. are illustrated. The experiment result shows that the proposed method can effectively retrieve relevant information about the question answering and improve the efficiency.

Key words: online question answering system; Lucene; full text retrieval

(责任校对　蒋云霞)