

单位代码: 10359
学 号: 201011120702

分 类 号: TP181
密 级:



合肥工业大学

Hefei University of Technology

硕士学位论文

MASTER DEGREE THESIS

论文题目: FAQ 问答系统中的问句相似度研究

学位类别: 学 历 硕 士

学科专业:
(工程领域) 计算机系统结构

作者姓名: 强 继 朋

导师姓名: 田卫东 副教授

完成时间: 2013 年 04 月

FAQ 问答系统中的问句相似度研究

Research on the Questions Similarity in the FAQ Answering System

作者姓名 _____ 强继朋 _____
学位类型 _____ 学 历 硕 士 _____
学 科、专 业 _____ 计算机系统结构 _____
研 究 方 向 _____ 人工智能与数据挖掘 _____
导 师 及 职 称 _____ 田卫东 副教授 _____

2013 年 4 月

合肥工业大学

本论文经答辩委员会全体委员审查，确认符合合肥工业大学
硕士学位论文质量要求。

答辩委员会签名：（工作单位、职称）

主席：陈海，中国科技大学，副教授

委员：

王浩 合肥工业大学 教授
孙丙宇 中国科技大学合肥物质科学研究院 研究员
胡弘 合肥工业大学 教授
何强风 " "

导师：

何

副教授

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得合肥工业大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：



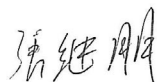
签字日期：2013年05月01日

学位论文版权使用授权书

本学位论文作者完全了解 合肥工业大学 有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权 合肥工业大学 可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

(保密的学位论文在解密后适用本授权书)

学位论文作者签名：



导师签名：



签字日期：2013年05月01日

签字日期：2013年05月01日

学位论文作者毕业后去向：

工作单位：

电话：

通讯地址：

邮编：

FAQ问答系统中的问句相似度研究

摘 要

常问问题库是问答系统的重要组成部分。问答系统通过将常见问题及其答案存储起来,形成常问问题集,来提高类似问题的答案搜索与合成效率。FAQ在使用上,存在问题集的更新和匹配新问题两个主要的问题,而解决这两个问题的关键,则在于问题(或称问句)相似度的准确计算。

本文主要针对 FAQ 自动问答系统中的问句相似度进行研究,利用中文问句和中文语言的特点以进一步提高问句相似度计算的精度,达到改善 FAQ 问答系统性能的目的。

论文的主要工作如下:

(1)现有文献中,问句相似度的计算主要借鉴普通陈述句的相似度计算方法,而普通陈述句的相似更多反映的是语句间语义上的匹配符合程度,而衡量问句间的相似性则须同时考虑问句及其答案句之间的相似程度,为此,设计了一种新的问句相似度计算方法。该方法不仅利用问句之间的语义和语法特征考察问句之间的匹配程度,还利用问句的问题类型等信息来间接刻画答案句之间的特征形象,从而获取问句的深层语义信息,以提高问句相似度计算的准确性。实验验证了该方法的有效性。

(2)由于基于《知网》的问句相似度计算方法中,词语相似度计算方法是利用相同的处理方法来处理对义词反义词与同义词近义词,从而有可能导致截然相反的两个问句拥有很高的相似度。为此,设计了一种改进的基于《知网》的问句语义相似度计算方法,该方法中不仅能够处理中文词语之间的褒贬性,也能够处理中文词语与英文单词之间的褒贬性,其中,对义或者反义的词语对问句相似度计算结果都起着副作用,从而降低了相反的反问句成为相似问句的可能。实验证明了该方法更加有效。

(3)基于上述研究,给出一个 FAQ 自动问答系统的原型系统,包含本文的一些研究方法的演示,为今后更加深入的研究提供一个平台。

关键词: FAQ 问答系统;问句相似度;问题分类;知网

Research on the Questions Similarity in the FAQ Answering System

Abstract

FAQ module is an important part of answering system. FAQ is formed through storing common questions and their answers, which can help improving the efficiency of answer seeking and synthesizing for similar questions. When using FAQ in answering system, there are two key problems, including question-updating and question-matching for new problems. And the key to these problems is the method of question similarity computation.

We focus on the questions similarity computation in FAQ system. We aim to find a new questions similarity computation method, which can use the characteristics of Chinese and Chinese questions to improve the Chinese questions similarity accuracy, and meanwhile to improve the performance of the Chinese question answering system.

Our contributions are as follows:

(1) In the existing literatures, questions similarity computation methods are mainly based on general declarative sentence similarity computation methods, where only semantic similarity between two general declarative sentences is considered. While considering questions similarity, both similarity between two questions and similarity between two answers should be considered. So a new questions similarity computation method is proposed, which takes into account not only semantic and syntactic characteristics between questions, but also information such as question type to indirectly get the characteristics of the answer, in order to extract deep semantic information of question, and finally to improve questions similarity computation accuracy. Experimental results show that the method can achieve better accuracy.

(2) For question similarity computation method based on Hownet, same word similarity method is adopted for those specific words such as contrastive words, contradictory words and synonymous words, which might result that two opposite questions have wrong high similarity. For all this, an improved questions similarity computing method based on Hownet is proposed. In this method, not only orientation inference between Chinese words but also orientation inference between Chinese word and English word can be computed correctly. And, contrastive or contradictory words have a negative effect on questions similarity computation,

which reduces the possibility that opposite questions have high similarity. Empirical results show the validity of the method.

(3) A prototype of FAQ answering system is designed to provide some related algorithms from our research, and a platform for future in-depth studies.

Keywords: FAQ Question Answering System; Questions Similarity; Question Classification; Hownet

致 谢

时间过的真的很快,从2010年7月份第一次进入实验室学习,到现在已经将近3年了,研究生生涯即将结束,新的生活也即将开始。研究生生涯受到很多老师的潜心指导、许多师兄师姐的细心的帮助和其他一些同学的意见,都对我的论文写作有着很大的帮助。尽管感觉到还有好多东西要学,还有好多东西不会,但是我并不害怕,因为我学习到了怎么去发现问题,然后去解决问题,这才是我研究生期间最大的收获。

记得2010年暑假还在家等待着最终被哪个导师录取,有一天收到姜海秋师兄的电话,说我被田卫东导师录取,还告诉我可以准备进实验室学习。当时就非常兴奋,没几天就到了实验室。从那时开始田老师就给我安排任务,告诉我怎么完成任务和一些做人的道理,中间没少给田老师惹麻烦,也吸取了教训,会继续努力。可以这样说,这三年来和导师打交道的最多,学到的最多,真的要十分感谢我的导师田老师。

刚进实验室,也多亏了实验室张娟、姜海秋、吴芳和赵利等师兄师姐的帮忙,少走了不少弯路,我也会继续把这种好的品德传承下去。研究生期间的休息时间,有几位好友和室友的陪伴,也使研究生生活更加丰富多彩,更加回味无穷。

研一的时候,幸运的是旁听了吴信东老师的海清项目组的报告,渐渐的了解是怎么做研究。真的非常感谢吴信东老师、郭丹老师和谢飞老师,他们对我做的小论文都给予了很大的帮助,教会了我不少东西,懂得了做研究不是那么容易的事,真的需要脚踏实地。

要想有好的学习,好的学习环境和学习氛围必不可少,这真的要感谢胡学钢老师。有一次开会结束,教我们做了一套锻炼脊椎的运动,学习到了做研究也不是那么枯燥,而是要劳逸结合,对我以后的工作和生活都会有所启迪。

此外,我还要感谢学院的各位领导和老师们,感谢他们给我们提供的教学环境和便利条件。更加感谢那些评阅我的硕士论文和出席硕士论文答辩会的各位老师,感谢他们给我提出的批评和建议。

最后,我要感谢我的父母和兄弟姐妹们。没有父母的支持,我不可能读研究生,我一定不会辜负你们,从现在开始一定会常回家看看,好好孝敬你们。还有我哥哥、姐姐和弟弟的支持,时不时给予我的零花钱和买的几件衣服,都让我体会到了亲情的重要,真的太感谢你们了。

强继朋
2013 年 4

目 录

第 1 章 绪论	1
1.1 研究背景与意义	1
1.2 国内外研究现状	2
1.3 本文研究的主要内容	4
1.4 本文使用的平台和数据集	5
1.5 本文的组织结构	5
1.6 本章小结	6
第 2 章 语句相似度计算	7
2.1 问答系统的体系结构	7
2.2 词语的相似度计算	8
2.2.1 基于《同义词词林》的词语相似度	8
2.2.2 基于《知网》的词语相似度	10
2.2.3 词语相似度计算评价	11
2.3 语句的相似度计算	11
2.3.1 语句相似度的概念	11
2.3.2 基于向量空间模型的 TF-IDF 方法	12
2.3.3 基于语义依存的语句相似度计算	12
2.3.4 基于语义距离的语句相似度计算	13
2.3.5 编辑距离计算方法	13
2.3.6 语句相似度计算的评价	14
2.4 小结	14
第 3 章 基于问句特征的问句相似度计算	15
3.1 引言	15
3.2 问句相似度	16
3.2.1 问句相似度计算与语句相似度计算的差别	16
3.2.2 问题分类	17
3.2.3 否定型中文问题	18
3.3 问句相似度计算	18
3.3.1 问句相似度计算方法	18
3.3.2 问句相似度计算的步骤	20
3.3.3 时间复杂度分析	21
3.4 实验结果及分析	22
3.4.1 简单几个问句的计算	22
3.4.2 面试常问问题集	22
3.4.3 哈工大数据集	23
3.4.4 问题分类准确度的影响	23
3.4.5 答案数目的影响	24
3.5 本章小结	25
第 4 章 基于词语褒贬倾向的问句相似度计算	26
4.1 引言	26
4.2 词语相似度计算	27
4.3 问句相似度计算	29

4.4 实验结果及分析	30
4.5 本章小结	33
第 5 章 FAQ 自动问答原型系统	34
5.1 系统总体设计介绍	34
5.1.1 关键词的扩展	34
5.1.2 FAQ 库候选问题集的查找	35
5.1.3 问句相似度的计算	36
5.2 系统功能介绍	36
第 6 章 总结和展望	39
6.1 总结	39
6.2 展望	39
参考文献	41
攻读硕士学位期间参加研究的课题和发表的论文	46

插图清单

图 2-1 问答系统的体系结构	7
图 3-1 具体各个类上分类精度比较	24
图 3-2 问句分类准确度对问句相似度计算准确度的影响	24
图 3-3 答案的数目对问句相似度计算准确度的影响	25
图 4-1 《知网》中海岸(coast)的概念定义	27
图 4-2 根节点为“entity 实体”的树中的一个分支	28
图 4-3 问句相似度计算实验结果	32
图 4-4 不同方法实验结构的对比	32
图 5-1 系统的总体框架图	34
图 5-2 FAQ 库数据结构示意图	35
图 5-3 系统的主界面	37
图 5-4 FAQ 库的打开	37
图 5-5 采用多特征问句相似度计算方法的结果	38
图 5-6 采用基于问句特征的问句相似度计算结果	38

表格清单

表 2-1 《哈工大同义词词林扩展版》编码规则表	9
表 2-2 几种语句相似度计算方法对比	14
表 3-1 问句相似度计算的例子	16
表 3-2 比较有代表性的 TREC 分类体系	17
表 3-3 中文问题分类体系	18
表 3-4 否定型的中文问题	18
表 3-5 问句相似度计算的伪代码	20
表 3-6 FAQ 问答系统的执行伪代码	21
表 3-7 简单几个问句的计算结果	22
表 3-8 面试常问问题集的实验结果	23
表 3-9 哈工大数据集的实验结果	23
表 4-1 中文词语相似度计算结果	29
表 4-2 中英文词语相似度计算结果	31
表 4-3 用户问句与 FAQ 库中问句的语句相似度计算结果	31

第1章 绪论

随着Internet的快速发展,网络中的资源在爆炸式的增长,如何在这些海量的数据中快速挖掘出用户所需要的信息也越来越困难,所以如何进行准确的信息定位成为用户关心的问题。虽然各种搜索引擎的出现给用户搜索信息带来了很大的便利,但返回的是与关键词相关的全部文档,用户必须自己从这些相关文档中寻找答案。如何用计算机从这些相关文档中寻找答案,然后返还给用户,就是目前得到广泛关注的自动问答系统。

1.1 研究背景与意义

自动问答技术^[1,2]是自然语言处理领域和信息检索领域的热门研究方向。为了推动自动问答系统的研究,文本检索会议^[3,4,5](Text Retrieval Conference, TREC)从1999年开始对问答系统的性能进行测评。许多研究单位参加会议进行交流,对自动问答^[6,7,8]的各方面研究起了积极的作用。在2000年10月召开的ACL2000国际计算语言学学术会议上,有一个关于“Open-Domain Question Answering”的讨论会,进一步推动了自动问答领域的研究。

自动问答系统根据用户用自然语言提出的问题,进行问题理解、信息检索和答案抽取等一系列步骤,找到简洁的答案返回给用户。例如问题“中国的首都是哪个城市?”,返回答案“北京”。对于这个看似简单的问题,计算机处理起来都很困难。不仅需要自然理解的技术,还需要信息检索和答案抽取的技术。可以这样说,由于问题形式的多样性和答案抽取的复杂性,快速的找到合适的答案是非常困难的。

现实中,由于有些问题会被多次问到,进行反复的信息检索和答案抽取,会大大降低问答系统的效率。所以,为了提高问答系统的效率,常常会把用户经常提到的问题及其对应的答案收集起来,形成常问问题集(Frequently asked question, FAQ)。FAQ库^[9,10,11]做为问答系统的一个子模块,这样进行问题检索,首先从FAQ中进行检索,如果检索到类似于用户的问题,就直接返回给该问题对应的答案,否则,进行原来的步骤。

问答系统中使用了FAQ后,就存在匹配用户提交的问题和FAQ库的更新两个主要问题,而解决这两个问题的关键,则在于问题(或称问句)相似度的准确计算。现有文献中,问句相似度的计算主要采用普通陈述句的相似度计算方法。普通陈述句的相似度计算方法只考虑了两个问句之间语义的相似,或者结构上的相似等等,并不关心两个问句是否是类似的问题。而实际情况是往往两个语义和句法结构都相似的问句却不是类似的问题,对应的答案也差别很大,比如,问句“哪个国家的货币最值钱”和“哪个国家的货币不值钱”。所以,计算问句相似度的过程中,不仅要考虑问句语义上的相似性,也要结合问句与答案的信

息运用到问句相似度计算中。

FAQ问答系统的实质就是从FAQ库中找到类似于用户提交的问题的“问题-答案”对，如果存在，把对应的“答案”返回给用户。因此需要对用户提交的问题与FAQ库中问题进行相似度计算，然后选择大于阈值且相似值最高的问句对应的答案返回给用户。可知，问句相似度计算的研究对问答系统具有很重要的意义，问句相似度计算的准确率对问答系统性能也有很重要的作用。

1.2 国内外研究现状

目前，文档或者长文本之间的相似度度量得到了广泛的研究，而短文本或者语句的相似度度量有着少量的研究。语句相似度与文本相似度的不同在于，语句是一种短文本形式，语句中包含的信息比较少，而且没有上下文环境，导致了语句相似度计算非常困难。因此，语句相似度需要对语句做深层次的分析，例如，句法分析，语义分析等，才能取得很好的精度。

(1)英文语句相似度计算

目前，基于英文语句相似度计算的方法可以大致分为四类^[12]：词共现的方法、基于语料的方法、混合的方法和基于关键词的方法。

基于词共现的方法^[10]也叫做“词袋”方法。通常用于信息获取系统^[13]当中，用来计算文档的相似度。系统拥有预先编译的 n 个词。为了能包含自然语言中有意义的词， n 的大小一般都是成千上万。然后，每个文档都可以用这 n 个词进行空间向量表示。文档的相关度就可以基于两个文档向量之间的相似度。词共现的技术基于越相似的文档共享着更多相似的词的思想。该方法有着显著的缺陷： n 一般远远大于语句的长度，导致向量空间都是空元素；相似的语句不一定共享着许多词。后来一系列改进的方法被提出。Okazaki等人^[14]获取语句的相似度是通过聚集所有词对的相似值。Chiang等人^[15]引入了模式匹配的方法，考虑了语句结构信息中词的共现。

另一个研究语句相似度比较热门的方法是基于统计信息的方法。基于语料资源比较熟知的方法有潜在的语义分析方法(LSA)^[16,17]和超空间模拟语言模(HAL)型^[18]。在LSA中，首先要从大量的上下文当中找出代表性词的集合。然后构建一个基于词的上下文的矩阵。当用LSA来计算语句相似度时，每个语句都用简化的向量进行表示，然后通过计算两个向量的相似度。HAL不像LSA需要通过文本中文档或者段落的单元来建立词的矩阵，HAL建立的是基于滑动窗口词共现的矩阵。窗口(通常宽度是10个词)的移动在语料的整个文档中。通过给定的 N 个词，形成一个 $N*N$ 的矩阵。当在文档中移动窗口时，矩阵记录着词共现的权重。通过结合矩阵中的行和列，这样一个词的意思就可以用一个 $2N$ 维的向量表示。然后，语句的向量通过增加语句中词的向量形成。两个语句的相似度就可以用矩阵来计算，例如欧几里得距离。

第三种方法是基于特征的方法。语句用设定的一些特征来表示，一般情况下，选用一些名词来作为特征。后来一些改进的特性向量表示方法，引入了主特征和复合特征^[19,20]。主特征是那些主要的特征相对于每个文本中单个项目。复合特征是成对的主特征的联合。这样一个文本就可以用这些主特征和文本特性来表示。文本之间的相似度就可以通过训练的分类器得到。该方法的困难之处是怎样自动的从语句中获得特征值。训练向量集合是不切合实际的和费时的任务。

由于语句信息的缺少，基于语义的方法得到广泛的关注^[21,22,23]。常用的语义资源是WordNet^[24]，可以用来计算英文词语的相似度。语句与语句之间的相似度通过用WordNet来计算词与词之间的相似度，进而得到语句之间的相似度。Li等人^[22]还结合了词语之间的顺序来计算语句的相似度。

(2)中文语句相似度计算

中文语句相似度计算和英文语句相似度计算存在着相同之处，如语句简短包含的信息量少和没有上下文环境。但是，中文语句相似度计算也有自身的特点和难点，如中文语句的自然语言处理更难^[25]、语句结构更加复杂等。从而，中文的语句相似度计算如果直接借鉴英文语句相似度的计算方法，效果并没有原来的好，所以需要根据中文语句的特点重现设计方法。目前，中文语句相似度计算大致有下面几种方法：基于关键词信息的方法、基于语义信息的方法和基于句法特征的方法。

基于关键词信息的方法将语句看成词的组合，然后通过对比两个语句中相同词的数目和词的顺序计算语句的相似度。详细划分可以分为基于统计信息的方法和基于规则方法。

基于统计信息的方法^[26]通过训练语料找出代表意义的关键词，然后用这些关键词作为特征，把语句表示成向量形式。语句之间的相似度可以通过两个向量间的夹角余弦值计算得到。该方法的优点是简单易实现，效率高。但是它是对关键词的词频进行统计，为了更好的反应统计特征，适合于处理长文本的数据。另外，该方法不能很好的处理一次多义的情况，主要是因为只引入了词的统计信息，没考虑蕴含的语义信息。

基于规则的方法是利用词的相似度的不同组合计算语句之间的相似度，其中词的相似度是根据词之间的语义距离、词的形态变化和反义词等等。张等人^[27]计算两个语句的相似度是通过同义词表得到两句之间的语义距离，从而用语义距离衡量语句相似度。钱等人^[28]是根据词形相似度和词序相似度计算两句的相似度。基于规则的方法在特定领域的问题系统中采用的比较多，可以根据不同情况设置不同的规则。而在广泛的情况下，很难人为的设定规则，就需要更多考虑语句之间的语义关系。

基于语义信息的语句相似度更能深入的了解语句内容的信息。要做语义相

似度计算必须得构建语义知识资源。目前,进行中文语句相似度计算,主要利用董振东和董强先生创建的《知网》^[29](HowNet)进行词的语义相似度计算,然后利用最大匹配法得到语句的相似度值^[30]。基于语义信息的方法考虑了语句中词之间的深层语义信息,不同于基于关键词信息的方法只考虑表层信息的局限性。但是,基于语义的方法会因使用词典的不全面从而使计算带来一定的误差。另外,基于语义信息的语句相似度计算没有考虑语句的结构信息,准确度有待进一步提高。

为了进一步提高语句之间相似度计算的准确度,一些基于句法结构的方法也得到了大家的关注,如骨架依存方法^[31]、语义依存方法^[32]。句法分析的方法目前采用哈工大的语言技术平台(Language Technology Platform, LTP)的居多。LTP平台^[33]是哈工大社会计算与信息检索研究中心历时十年开发的一整套开源的中文语言处理系统。LTP提供了包括分词、词形标注、命名实体识别、语义消歧和语义角色标注等功能。由于汉语语言的复杂性和结构的复杂性,基于句法结构的方法的研究还不够完善,准确度还不能达到满意的要求。

综上所述,上面介绍的所有算法都是用来计算语句相似度,直接拿来用于问句相似度计算,并一定十分合适,主要因为FAQ问答系统中的问句相似度计算不仅仅要求语句之间的相似,还要求问句与答案的匹配。

1.3 本文研究的主要内容

问句相似度计算是问答系统的一个重要模块,问句相似度计算的精度对问答系统的整体性能具有重要的作用。本文研究的主要目的是希望在已有的问句相似度计算的基础上,通过分析问句的特征和语义来提高问句相似度计算的准确度。主要工作包括以下两点:

(1)问句特征。由于问答系统中问句相似度计算的特殊性,如两个问句相似度值高就不一定是类似的问题。而且,目前已有的问句相似度的计算方法,采用的是普通陈述句的相似度计算方法,并没有深入的考虑问句本身的特征。本文的问句相似度计算综合考虑了问句之间的匹配和问句与答案的匹配。但是,结合答案句的特征信息来协助获取问句的深层语义信息,困难在于,在FAQ的问句匹配阶段,无法获知答案句。所以,怎么分析问句间接的刻画答案句的信息是本文的研究重点。

(2)词语的褒贬倾向对问句相似度计算的影响。尽管问句相似度计算采用的是基于《知网》的词语间的相似度计算,但是在计算词语过程中对义词、反义词的两个词语的相似度值也很高,导致两个截然相反的问候句,相似度值也极高。所以,如何利用对义词、反义词对问句相似度计算的作用是本文主要研究的内容之一。另外,中文问句中含有英文词语越来越常见,如果忽略两个问句之间中文词语与英文词语之间的相似度计算,势必为降低两个问句的相似度值。所

以，如何处理中文问句中中文词语与英文词语之间的相似度也是本文研究的内容之一。

1.4 本文使用的平台和数据集

本文的实验中，使用的第三方平台有ICTCLAS3.0分词系统和知网。

ICTCLAS3.0 分词系统^[34]是中国科学院计算技术研究所研制的汉语词法分析系统。该系统分词速度单机可以达到 996KB/S，和分词的准确度可以达到 98.45%。目前，该系统是公认的世界上最好的汉语词法分析系统之一，并且支持多种开发语言和多种操作系统，免费开放和使用。

《知网》(英文名称为HowNet)是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。

本文实验中采用的问题集一部分是我们自己整理的常用的面试问题集。还有一部分是哈尔滨工业大学相关研究人员整理并进行标注的(简称为 HQ 问题集)，在中文问题研究中具有一定的代表性。在实验过程中，会根据实验需要的具体情况，对 HQ 问题集进行了进一步的修改。

对文本分类处理结果的评价标准中最常见的是准确率(precision)和查全率(recall)，由于FAQ自动问答系统的特殊性，在选择评价标准中忽略查全率。本文实验过程中，问句相似度计算方法准确率的估算采用下面两个公式，

$$Prec_1 = \frac{1}{n} \sum_{i=1}^n a_i \quad (1-1)$$

$$Prec_2 = \frac{1}{n} \sum_{i=1}^n \frac{1}{r_i} \quad (1-2)$$

其中， n 为测试问题的总数， a_i 为第 i 个问题的值，如果正确为1，否则为0。 $Prec_2$ 是参照其他FAQ问答系统论文中的评价标准^[35]。该标准要求最多可以给予大于阈值的相似度最大的三个结果，其中 r_i 为正确答案的位置。若给出的结果中存在答案，如果是第一个， $1/r_i$ 取值为1，第二个和第三个分别取值为1/2和1/3，若都不正确，则 $1/r_i$ 取值为0。若FAQ库中没有答案，系统也没给予答案， $1/r_i$ 也取值为1。

1.5 本文的组织结构

本文共分为六章：

第1章：简单介绍了问答系统的背景和进展、语句相似度计算的国内外现状和主要技术、本文的研究内容、使用的平台和评价标准及本文的组织结构。

第2章：主要是对常用的语句相似度计算方法进行了详细的介绍。首先，介绍了问答系统的体系结构，然后给出了主流的词语之间的相似度计算方法。最

后，介绍了常用的语句相似度计算方法，包括基于统计学的方法、基于语义依存的方法、基于语义的方法和编辑距离的计算方法。

第3章：提出了一种基于问句特性的问句相似度计算方法。问句的特征，这里分析了两个方面，一个方面是问句的类型，另一方面是问句的肯定还是否定问法。通过引入问句的特征，结合原来的问句相似度计算，达到提高FAQ问答系统性能的目的。最后，给出了实验结果，并对实验结果进行了详细的分析。

第4章：主要探讨基于语义的问句相似度计算。在已有的利用语义信息进行问句相似度计算的基础上，没有考虑到对义词、反义词对问句相似度的影响，也没有考虑中文问句中出现英文单词对问句相似度计算的影响。为了解决上述情况，提出了一种新的基于语义的问句相似度计算方法。在新的方法中，弥补了以前没有考虑到词与词之间反义、对义和英文单词的情况，准确率得到进一步提高。

第5章：给出了FAQ自动问答原型系统。详细介绍了系统的总体框架和每一步的执行原理，也给出了运行结果。

第6章：总结与展望。总结了本文的研究工作，并探讨下一步的研究工作。

1.6 本章小结

首先，本章介绍了问句相似度计算的研究背景和意义，并给出了目前国内外已有的这方面的相关工作。然后，通过分析问句相似度计算中还存在的问题，给出了本文主要的研究工作。接着，介绍了本文中使用的平台和数据集，和选择的评价标准。最后，给出了本文的组织结构。

第2章 语句相似度计算

本章介绍了问答系统的体系结构，和目前已有的词语相似度计算方法和语句相似度计算方法，也对不同方法的各个方面进行了对比。

2.1 问答系统的体系结构

在自动问答系统中引入基于FAQ的辅助模块是满足常见问题回答的一种有效的手段，可以大大缩减检索的时间。典型的自动问题系统结构如图2-1所示。

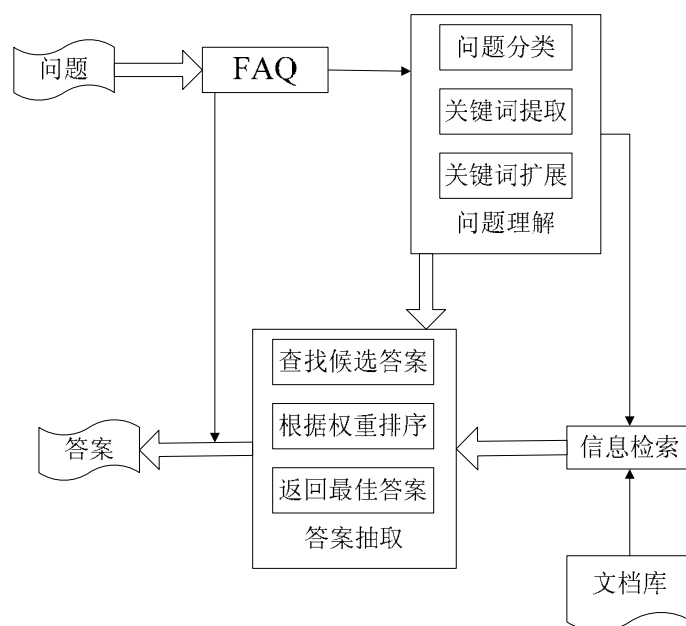


图 2-1 问答系统的体系结构

该结构主要包括以下四个模块。

(1)FAQ模块

FAQ库主要用于提高自动问答系统的效率。对自动问答系统进行检索，首先从FAQ库中寻找是否包括用户问题，如果包含问句，则直接返回问句对应的答案给用户，从而省去后面一系列的步骤；否则，就得进行问题理解、信息检索和答案抽取。FAQ库的主要技术就是问句相似度计算和候选问句的选择。

(2)问题理解模块

该模块主要是让计算机理解用户的问题，确定问题的关键词和问题类型从而为后面的信息检索和答案提供服务。一般可分为下面几个部分：问题预处理、问题分类、关键词提取和关键词扩展等。其中最困难的是问题分类，问题分类主要是确定问题的类别，方便于信息的检索和答案的抽取。该模块用到的技术主要包括分词、同义词词典、分类方法等。

(3)信息检索模块

该模块主要从互联网或者知识库中找到与问题相关的文档，作为答案提取

的原材料。主流的检索技术，一部分是利用普通的搜索引擎提供的原材料，如 Google 和百度等，还有一部分是建立特定的知识库，然后根据知识库建立索引模块，方便快速的找到相关文档，并根据特点的排序算法对文档进行排序。包括的主要技术有查询扩展、语料库的构建技术、词汇索引、文档排序等。

(4) 答案抽取模块

利用问题的类型，构建相应的答案抽取策略，从信息检索后的文档中排序比较高的文档的前 N 个文档进行答案的定位和输出。涉及到技术主要有答案抽取模版的制定、模式匹配、聚类等方法。

2.2 词语的相似度计算

汉语语句中，字是组成语句的最小单位，但表达语句具体含义的是由字构成的词语，所以词语才是语句语义和语法的基本单位。因此，词语的相似度计算是语句相似度计算的基础。词语的相似度计算也需要结合其应用背景，本文的词语相似度是已自动问答系统为背景，研究词语语义的相似度。

词语语义的相似度计算模型可以分为下面两类：一类是基于大规模语料库统计的方法^[36]；一类是基于语义词典的方法^[37]。下面将详细介绍本文采用的基于语义词典的方法。

基于语义词典的词语相似度计算方法是以前语义词典作为语义分类体系。语义词典中概念间存在上下位、反义和同义等多种关系。词语相似度关系就是利用这些关系计算得到概念之间的语义距离，进而得到概念间的相似度，最后用概念相似度去衡量词语相似度。

汉语语义词典中具有代表性的有《知网》、《同义词词林》^[38]等；英语语义词典有 WordNet 等。国内的学者多是利用《知网》和《同义词词林》计算词语相似度。

2.2.1 基于《同义词词林》的词语相似度

《同义词词林》是梅家驹等人于1983年完成的一部语料库。这本词典中不仅包括了词语的同义词，也包含了一定数量的同类词，即与其相关的词。考虑到年限的久远，后来由哈尔滨工业大学信息检索实验室对《同义词词林》进行更新，新的版本为《同义词词林扩展版》。《同义词词林扩展版》收录词语近77343条，剔除14706个罕用词和非常用词的一部同义词词典。

《同义词词林》按照树状的层次结构把所有收录的词条组织到一起，把词汇分成大、中、小3类，大类有12个，中类有97个，小类有1400个。最后，小类根据同义原划分成许多词群作为第四级，接着，词群下面划分出原子词群为第五级。这样，《同义词词林扩展版》就表示成了一个由上到下，由粗到细的五级层次的语义分类体系。为了更好的对收录的词语进行计算，所有词都通过

一个编码进行描述，编码格式如表2-1所示，

表 2-1 《哈工大同义词词林扩展版》编码规则表

编码位	1	2	3	4	5	6	7	8		
符号举例	D	a	1	5	B	0	2	=	#	@
符号性质	大类	中类	小类		词群	原子词群				
级别	第一级	第二级	第三级		第四级	第五级				

《同义词词林》一共有5层编码：第1级用大写英文字母；第2级用小写英文字母；第3级用二位十进制整数；第4级用大写英文字母；第5级用二位十进制整数表示。例如：“Ad01A01=居民 居者 定居者 居住者”，“Ad01A01= ”是编码，“居民 居者 定居者 居住者”是该类的词语。

由于第5级包含有同义词、相关词和只有一个词的情况，所以分类结果需要特别说明，可以分为下面3种情况。由于语义层面上的意思是不同的，所以在使用过程中，这3种情况必须分别对待。《同义词词林》对第8位的标记有3种，分别为“=”、“#”、“@”。“=”代表相等和同义的意思；“#”代表不等或者同类的意思；“@”代表独立的意思，即在词典中没有相同和相关的词。

基于《同义词词林扩展版》的编撰特点，田久乐等人^[39]设计了一种基于《同义词词林》的方法来计算词语相似度。

首先判断在《同义词词林扩展版》中作为叶子节点的两个义项分别在每一层分支，即两个义项的编号在哪一层不同。例如：Ad01A04与Ad01B03在第4层分支。从第1层开始判断，相同则乘1，否则在分支层乘以相应的系数，然后乘以调节参数 $\cos(n*\pi/180)$ ，其中 n 是分支层的节点总数，该调节参数的功能是把义项相似度控制在 $[0, 1]$ 之间。

词语所在树的密度，分支的多少都直接影响到词之间义项的相似度。例如，密度较大的义项相似度的值相比密度小的相似度的值要更精确，所以需要再乘以一个控制参数 $(n-k+1)/n$ ，其中 k 是两个分支间的距离。这样把原本计算出的值更加细化，计算的结果也更加精确。

若两个义项的相似度用 $Sim(A,B)$ 表示

(1)若两个义项不在同一棵树上

$$Sim(A,B)=f \quad (2-1)$$

(2)若两个义项不在同一棵树上：

若在第2层分支，系数为 a

$$Sim(A,B)=1*a*\cos(n*\pi/180) (n-k+1)/n \quad (2-2)$$

若在第3层分支，系统为 b

$$Sim(A,B)=1*1*b*\cos(n*\pi/180) (n-k+1)/n \quad (2-3)$$

若在第4层分支，系统为 c

$$Sim(A,B)=1*1*1*c*\cos(n*\pi/180) (n-k+1)/n \quad (2-4)$$

若在第5层分支，系统为 d

$$Sim(A,B)=1*1*1*1*d*\cos(n*\pi/180) (n-k+1)/n \quad (2-5)$$

论文中，层数初值分别设置为 $a=0.65$ ， $b=0.8$ ， $c=0.9$ ， $d=0.96$ ， $e=0.5$ ， $f=0.1$ 。

2.2.2 基于《知网》的词语相似度

关于任意两个事物的相似度，Dekang Lin^[40]从信息论的方面进行总结，认为事物的相似度取决于它们的共性和个性，从而给出了任意两个事物的相似度计算公式，如公式2-6所示。

$$Sim(A,B)=\frac{\log p(common(A,B))}{\log p(description(A,B))} \quad (2-6)$$

其中分子是描述A、B共同拥有的信息量；分母是A、B总共包含的信息量。

刘群等人^[37]认为词的相似度就是两个词语在不同的上下文中可以互相替换使用而不改变文本的句法语义结构的程度。两个词语，如果在不同的上下文中可以互相替换且不改变文本的句法语义结构的可能性越大，二者的相似度就越高，否则相似度就越低。刘群等人的描述相对Dekang Lin描述的更具体化，主要以基于实例的机器翻译这一研究背景的。

在《知网》中有两个最重要的概念：“概念”与“义原”。概念是词语语义的一种描述。每一个词可以表达为多个概念。概念是用一种“知识表示语言”来描述的，这种“知识表示语言”所用的词汇叫做义原。义原是用于描述一个“概念”的最小意义单位。

给定两个词 W_1 和 W_2 ，假设 W_1 有 n 个概念 $C_{11}, C_{12}, \dots, C_{1n}$ ， W_2 有 m 个概念 $C_{21}, C_{22}, \dots, C_{2m}$ 。 W_1 和 W_2 之间相似度的公式定义如下：

$$Sim(W_1, W_2) = \max_{i=1 \dots n, j=1 \dots m} Sim(C_{1i}, C_{2j}) \quad (2-7)$$

概念是用义原来描述的。义原相似度是概念相似度的基础。义原相似度的计算依据义原的层次体系(上下位关系)来计算。两个义原的语义距离的计算公式如下：

$$Sim(p_1, p_2) = \frac{\alpha}{d + \alpha} \quad (2-8)$$

其中 p_1 和 p_2 代表两个义原； d 表示 p_1 和 p_2 在义原层次体系中的路径长度，是一个正整数； α 是一个可调节的参数，通常被设置为1.6。

《知网》中收录的词包含实词和虚词。虚词的描述比较简单，用“句法义原”或“关系义原”进行描述。实词的描述比较复杂，在《知网》中表示为一个特征结构，该特征结构含有以下四个特征：(1)第一基本义原描述：用 $Sim_1(S_1, S_2)$ 表示；(2)其他基本义原描述：除第一基本义原描述式以外的其他基本义原描述

式，用 $Sim_2(S_1, S_2)$ 描述；(3)关系义原描述：语义表达式中所有用关系表示的义原称做关系义原，用 $Sim_3(S_1, S_2)$ ；(4)符号义原描述：语义表达式中所有的用符号表示的义原称做符号义原，用 $Sim_4(S_1, S_2)$ 。最后两个实词相似度的计算公式定义如下：

$$Sim(S_1, S_2) = \sum_{i=1}^4 \beta_i \prod_{j=1}^4 Sim_j(S_1, S_2) \quad (2-9)$$

其中 $\beta_1, \beta_2, \beta_3$ 和 β_4 都是调价参数，且有 $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$ ， $\beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$ ，反映从 $Sim_1(S_1, S_2)$ 到 $Sim_4(S_1, S_2)$ 这四个相似度值对总体相似度所起到的作用依次递减。

2.2.3 词语相似度计算评价

词语的相似度计算方法主要从语料库和语义词典两个方面。语料库方法需要大量的语料库数据，由于数据原因，可能会导致数据稀疏性问题。而基于语义词典的方法，由于词典不可能收录所有的词语，会出现一些词检索不到，称为未登录词。《知网》相对于《同义词词林》描述的更具体，更常用于词语相似度计算。所以，比较常用的一种计算词语相似度的方法是对登录的词语用语义词典的方法进行计算，未登录的词语用语料库的方法进行计算。

2.3 语句的相似度计算

2.3.1 语句相似度的概念

相似度是一个比较抽象的概念，在很多领域都有着广泛的研究，例如哲学、语义学和信息理论等。目前，由于相似度涉及到的知识面比较宽，导致很难给相似度一个明确的定义。在这里，着重强调语句之间的相似度，语句相似度的定义必须结合具体的应用背景。

在自动问答系统的信息检索中，语句相似度更多地反映语句间的语义上的匹配符合程度，当语句的相似度达到某个设定的阈值时，就认为这两个语句相似。所以，语句相似度的计算就是计算两个语句在语义上的符合程度，即相似值越大认为语句越相似，相似值越小则认为两个语句越不相似。相似度的值一般取 $[0, 1]$ 之间的实数，如果相似值为 1 时，认为两个语句是完全相同的两个语句。当相似值取值为 0 时，认为两个语句是完全不同的两个语句。

根据对语句的分析理解程度，语句相似度计算可以分为下面四种方法：(1)基于关键词信息模型的方法：只对语句中的词进行处理，即词的词频和词性等信息。(2)基于语义依存的计算方法：对两个语句进行深层的句法结构分析，根据句中的语句依存关系进行语义相似度计算。(3)基于语义分析的计算方法：需要用已构建的语义知识库对句中的词的相似度计算，然后根据词的相似度结果计算语句的相似度值；(4)编辑距离的相似度计算：分析两个语句转变成相同语

句需要的最小步骤，一种常用于模式匹配中的方法。

2.3.2 基于向量空间模型的 TF-IDF 方法

这种方法只需要语句表层的关键词信息，不需要对语句内容的深层理解。目前最常用的基于关键词信息的方法是基于向量空间模型(Vector Space Model, VSM)的 TF-IDF(term frequency-inverse document frequency)方法。

TF-IDF 是统计方法的一种，统计关键词在语料库中的出现频率，通常建立在大量真实的文本语料的基础之上。在问句相似度中，词频(term frequency, TF)指某个词在这个问句中的出现次数；逆向文档频率(inverse document frequency, IDF)由总问句数目除以含有该词的问句数目，再取对数得到，该值是描述某个词重要性普遍存在的度量。这样，只有某个词在某一特定问句中出现的频率高，但在整个问句集中出现的频率较低，该词的 TF-IDF 值会很大。

若用户提问的问句与 FAQ 中间句包含的所有词为 w_1, w_2, \dots, w_n ，则问句用一个 n 维向量 $T = \langle T_1, T_2, \dots, T_n \rangle$ 来表示。其中 $T_i (1 \leq i \leq n)$ 为 w_i 的 TF-IDF 值。TF-IDF 计算公式如下：

$$T_i = n * \log(M / m) \quad (2-10)$$

其中 n 是词 w_i 在该问句中出现的次数， M 表示 FAQ 中间句的总数目， m 表示包含词 w_i 的问句数目。

同样，FAQ 中每一个问句都可以用一个 n 维向量表示。假设一个问句用一个 n 维的向量 $T' = \langle T'_1, T'_2, \dots, T'_n \rangle$ 来表示。计算 T 和 T' 两个问句之间的相似度就可以利用 T 和 T' 这两个向量之间的夹角的余弦值来表示，如公式 2-11 所示。

$$Sim(T, T') = \frac{\sum_{i=1}^n T_i * T'_i}{\sqrt{\sum_{i=1}^n T_i^2 \sum_{i=1}^n T'^2_i}} \quad (2-11)$$

TF-IDF 方法是一种基于词频信息的统计方法，用于大规模的文档库中文本向量表示会产生较好的效果，主要因为语句含有词的数目比较少，没有上下文信息。

2.3.3 基于语义依存的语句相似度计算

依存句法是由 L.Tesniere 在其《结构句法基础》^[41]一书中第一次提出。依存句法促进了后来的语言学的发展，例如，在计算机语言学界中依存句法得到了广泛的重视。依存句法主要分析语句内部成分之间的语义依存关系，找出其中的内在联系和修饰关系。利用依存句法来用于语句相似度计算^[42]，就是要获取内部之间的依存关系。

目前，中文的依存句法分析器采用哈工大信息检索实验室开方的比较多。该依存体系中包含了 24 种依存关系，用于语句相似度计算时，就是要考虑有些配对之间的相似度。所谓有效配对是指全句核心词和直接依存于它的有效词组成的配对，这里有效词定义为动词、名词以及形容词，它是由分词后的词性标注决定的。

相似度计算公式如 2-12 所示，

$$Sim(S_1, S_2) = \frac{\sum_{i=1}^n W_i}{Max\{P_1, P_2\}} \quad (2-12)$$

$\sum_{i=1}^n W_i$ 为语句 1 和语句 2 有效配对匹配的总权重， $Max\{P_1, P_2\}$ 为语句 1 和语句 2 的有效配对数目。

这种方法从句法深度进行考虑，考虑到了词与词之间的依存关系，对语句的理解更加充分，从而更准确的得到语句相似度的值。但是，现有的句法分析技术还不够成熟，还无法将所有的句法信息特征全部考虑进来，所以会产生一定误差。

2.3.4 基于语义距离的语句相似度计算

该方法是利用了基于语义距离的词语相似度，能够充分考虑语句中每个词的深层信息。最常用的基于语义距离的语句相似度计算方法是基于《知网》的问句相似度计算。调用基于《知网》的词与词的相似度计算结果，然后去计算问句的相似度计算结果。但由于词典的不全面和一些未登录词的词义代码的缺失，也会给计算带来一定误差。

2.3.5 编辑距离计算方法

编辑距离，又称为 Levenshtein 距离，指两个字符串之间，从一个转成另一个所需的最少编辑操作次数，包括插入，删除和替换三种操作。

编辑距离最初处理的是不考虑语义的字符，其在字符串相似度计算、拼写检查、自动文摘、音乐识别、图像识别、数据清理、语音识别等众多领域都有着广泛的应用背景。在语句的相似度计算方面，也有一定的应用。例如，Leusch 等人^[43]利用编辑距离计算语句之间的相似度，并用于机器翻译评价。

编辑距离计算中文语句相似度中，车等人^[44]提出了一个新方法。该方法通过改进编辑距离方法来对中文相似语句进行检索，其主要思想是：以普通编辑距离算法为基础，采用词而不是单个的汉字作为基本的编辑单元，使用《知网》和《同义词词林》计算得到的词语之间的语义距离作为词语之间的替代代价，并且赋予插入，删除和替换三种操作不同的权重。该方法不需要句法分析，就

考虑了词汇的顺序和语义等信息，使之计算和实现都比较简单，也有不错的相似度计算效果。

2.3.6 语句相似度计算的评价

下面将从几个方面来比较上面提到的几种方法，对比结果如表 2-2 所示。

表 2-2 几种语句相似度计算方法对比

方法名称 评价指标	TF-IDF	词性和词 序结合的 方法	语义依 存	编辑距 离	语义距 离
复杂程度	难	易	难	难	易
可实现化	难	难	难	难	一般
对义、反义词	没考虑	没考虑	没考虑	没考虑	没考虑
匹配答案	没考虑	没考虑	没考虑	没考虑	没考虑
语义信息	没考虑	没考虑	考虑	适当考 虑	考虑
处理英文词语	不能	不能	不能	不能	不能

从表 2-2 可以看出，TF-IDF 和词性和词序结合没有考虑语义信息，用来处理问句相似度计算效果不会很好。对于是否考虑匹配答案和词语的褒贬性，上面几种方法都不能做到。由于中文问句中含有英文词语也越来越常见，不能处理英文词语也必定会影响问句相似度计算的精度，上面的几种方法也不能做到。所以，要选择一种方法更加通用和更适合用于 FAQ 系统的问句相似度计算，尽可能要考虑问句与答案的匹配，问句的褒贬倾向和包含英文词语是否能够得到处理。

2.4 小结

本章首先给出了问答系统的体系结构；接着，介绍了常用的词语相似度计算方法，并对词语相似度计算的不同方法的优缺点进行论述；然后，给出了语句的相似度计算方法，并对比了已有的方法，分析了各自的优缺点。

第3章 基于问句特征问句相似度计算

在介绍词语相似度的相关理论和方法的基础上，本章将分析问句的特征，并将其应用于问句相似度的计算中。目前，FAQ问答系统中的问句相似度计算主要借鉴普通陈述句的相似度计算方法，因为仅考虑问句间的相似因素而忽略答案句之间的相似因素，存在相似度计算准确性欠佳的问题。本章提出了一种综合考虑问句及答案信息的问句相似度计算方法。该方法不仅利用问句之间的语义和语法特征考察问句之间的匹配程度，还利用问句的问题类型等信息来间接考察答案句之间的相似程度，以提高问句相似度计算的准确性。实验结果表明，与现有方法相比，本文方法在问句相似度计算上具有较高的准确度。

3.1 引言

现有文献中，问句相似度的计算主要借鉴普通陈述句的相似度计算方法，例如，语义、语法、句法结构以及多种方法结合^[45,46]的方法。但是在FAQ问答系统的背景之下，问句之间相似性计算与普通陈述句之间相似性计算既有相似地方，也有所区别。这是因为衡量普通陈述句间的相似性仅须考虑两个陈述句在语义上的相似程度，而问句间的相似性的则须同时考虑问句及其答案句之间的相似程度。实际上，问句间相似度的细微差别常会导致答案的大相径庭，例如，句法结构完全不同的两个语句也许是相同问题，而语法句法结构完全相同的语句也许是不同问题，亦即问句间相似度计算需要考察语句的深层语义信息，如“北京有什么好吃的”和“北京有什么好玩的”。对此，现有普通陈述句的相似度计算方法是难以完成的。

其次，问句是否含有否定词，对问句的相似度计算影响也比较大。例如，用户问句为“什么货币最不值钱？”，如果FAQ问句中，包含有“什么货币最值钱”。用语义计算的结果为0.9，用语义结合语法的结果为0.918，都是比较高的相似度。很有可能就把“什么货币最值钱”的结果返回给用户，而达不到用户的满意。如果把问句的肯定性与否定性考虑到问句的相似度计算中，就有可能得到正确的结果。

但是，结合答案句的特征信息来协助获取问句的深层语义信息，困难在于，在FAQ的问句匹配阶段，无法获知答案句。为此，本章引入问题分类的相关研究成果^[47,48,49]，利用问题分类的结果，如问题类别、中心词等信息，来间接刻画答案句的特征信息，从而丰富问句相似性计算时的考虑因素。在此基础上，与传统的问句相似度计算方法所考量因素如语法和语义等相结合，设计了一种新的问句相似度计算方法。该方法虽然考虑了问句信息但并未增加多少额外的计算开销，这是因为在问答系统中，问题分类工作本身是需要首先完成的。

3.2 问句相似度

3.2.1 问句相似度计算与语句相似度计算的差别

普通陈述句之间的相似度更多的反映语句间的语义上的匹配符合程度，只要达到某个事先设定的阈值，就认为相似。而FAQ问答系统中的问句相似度计算，不仅要求两个语句的相似度很高，也要求问句之间对应答案的一致性，所以说，普通陈述句相似度计算方法如果用于问句相似度计算，就没有考虑到问句与答案的对应，必定会影响FAQ问答系统的性能。

为了更明显的显示问句相似度计算和普通陈述句相似度计算的不同，给出了几个简单问句的对比，采用了语义相似度计算方法^[50]和多种特征结合的相似度计算方法^[35]分别计算了问句的相似度值，结果如表3-1所示。

表 3-1 问句相似度计算的例子

序号	问句	语义	多特征
1	清朝一共有哪几位皇帝？	0.88	0.86
	清朝总共有几位皇帝？		
2	哪个国家的货币最值钱？	0.93	0.94
	哪个国家的货币最不值钱？		
3	上网对人有什么好处？	0.97	0.92
	上网对人有什么坏处？		
4	白宫在哪？	0.51	0.47
	白宫的地址是多少？		

从结果可以看出，前3对问句都有很好的相似度，但都不是相同的问题，都会影响FAQ问答系统的性能。第4对问题，虽然有很低的相似度，但确是相同的问题。原有的算法不能很好的计算这些问句，主要是下面几方面原因，

(1)没有考虑两个问句的类型是否一致。第1对问句就是两个不同的问题，一个是问人，另一个是问数字。第4对问句是相同类型的问题，都是关于地址方面的问题，但是没有考虑类型问题，导致相似度很低。

(2)没有考虑问句的句型。第2对问句，下面一句比上面一句多一个否定词，导致了结果的截然不同。

(3)没有考虑词语之间的褒贬倾向。第3对问句中“好处”和“坏处”分别一个是褒义词和贬义词，而采用的基于《知网》的词语相似度计算方法有很高的相似值^[37]。

为此，本文着重从问句的类型和句型两方面来间接刻画答案句的信息。为了解决此问题，在计算问句相似度的过程中，引入了问句的类型和句型作为问句相似度计算的一个因素。

3.2.2 问题分类

问题分类的目的主要是为了增加约束条件,便于信息检索和答案的提取。目前,并没有统一的问题分类体系。为此,很多问答系统都采用了较为复杂的问题分类体系^[51,52]。表3-2给出了几种在TREC会议上有代表性的几种英文问题分类体系,也是评测比较高的几种分类体系。

表 3-2 比较有代表性的 TREC 分类体系

系统名称	大类数量 /小类数量	大类类型名称
Cymfony	18/22+	AGE、AREA、DATE、DAY、DURATION、FREQUENCY、LENGTH、LOCATION、MONEY、NAME、NUMBER ORGANIZATION、PERSON、PRODUCT、REASON、TIME、WEIGHT
ISI	7/140+	Abstract Qtargets、Combinations of targets、Lexical Qtargets、Role targets、Semantic Qtargets、Slot Qtargets、Syntactic targets
SMU	16/30+	DATE、DEFINITION、DISTANCE、LOCATION、MANNER MONEY、NAME、NNP、NUMBER、ORGANIZATION、PERSON PRICE、REASON、TIME、TITLE、UNDEFINED
UIUC	6/50	ABBREV、DESCRIPTION、ENTITY、HUMAN LOCTIRON、NUERIC

注：“+”表示不少于标注的类型

由于本文用到的是中文的问题分类,因此将介绍中文问题分类体系。目前一些中文问题研究的问题分类体系都是从TREC会议上的英文分类体系直接翻译过来,采用的都是从粗类到细类的层次分类体系。虽然在某些类型上可能造成歧义,总体来说可也得到较高的分类精度,还可以增加约束条件便于分类。哈尔滨工业大学在UIUC的分类体系^[53]的基础之上,加上中文问题本身的特点,也采用了层次分类的方法制定了中文的分类体系。

本文就采用了哈工大的问题分类体系。其中包括7大类和65个小类。但未知类 UNKNOWN 在问题集中的数量非常少,对分类结果几乎没有影响,因此为了方便起见,在后面的实验中是不包括该类的,即采用6大类64小类。

具体的分类体系如表3-3所示。

表 3-3 中文问题分类体系

大类名称	小类名称
描述类	简称 定义 表达 判断 方法 意义 其它 原因
时间类	日 时代 月 其它 范围 时间 年
实体类	理论 动物 艺术 器官 民族 颜色 货币 事件 食物 工具 语言 文字 材料 其它 植物 宗教 运动 物质 职业
数字类	面积 号码 数量 距离 频率 金钱 其它 百分比 周期 范围 顺序 速度 温度 重量
人物类	描述 团体机构 人物
地点类	地址 建筑 城市 大陆 国家 岛屿 湖泊 山脉 大洋 其它 省 河流 星球
未知类	

3.2.3 否定型中文问题

汉语常用的否定词有不、非、否、无、没、莫、罔、弗、勿和毋等。汉语中的问句常被分为三种，肯定型、否定型和不确定型。有些问句就是因为多了一个否定词，导致答案的截然相反。但并不是有否定词的问句都是这样的。例如，“有没有有特异功能的比较好看的电影？”，这个问句虽然含有否定词，但却不是一个否定型的问句，这样的问句被称之为不确定型问句。在这里我们关注的问题是类似如表3-4所示的问题。

表 3-4 否定型的中文问题

序号	问题
1	哪两个南美国家不与巴西接壤？
2	什么货币最不值钱？
3	人在什么情况下会没有攻击性？

本文识别哪些问句是否定型中文问题，是必须包含下面两个条件：问句分词后含有否定词；否定词前后两个词不能一样。两个条件都不满足，认为属于肯定型中文问题。如果含有否定词，且否定词前后两个词一样，认为属于不确定型问题。

3.3 问句相似度计算

3.3.1 问句相似度计算方法

(1) 语法方法

语法相似度主要从词形、词序、句长和距离四个方面进行考虑。

词形相似度反映两个问句中词语在形态上的相似程度，用两个问句中含有的共同词的个数来衡量。用 $WordSim(A,B)$ 表示问句 A 和 B 的词形相似度，计算

如公式 3-1 所示。

$$WordSim(A, B) = 2 * \frac{Same(A, B)}{Len(A) + Len(B)} \quad (3-1)$$

其中, $Same(A, B)$ 表示 A 和 B 中共同词的数目, 如果 A, B 中一个词出现的次数不同时, 以出词次数少的那句的数目为准。 $Len(A)$ 和 $Len(B)$ 分别表示 A 和 B 中词的数目。

词序相似度反映两个语句中词语在位置关系上的相似程度, 计算如公式 3-2 所示。

$$OrderSim(A, B) = \begin{cases} 1 - \frac{ReWord(A, B)}{|OrderOccur(A, B)| - 1}, & |OrderOccur(A, B)| > 1 \\ 1, & |OrderOccur(A, B)| = 1 \\ 0, & |OrderOccur(A, B)| < 1 \end{cases} \quad (3-2)$$

其中 $OrderOccur(A, B)$ 表示在 A, B 中都出现且都只出现一次的词, $PFirst(A, B)$ 表示 $OrderOccur(A, B)$ 中的词在 A 中的位置序号构成的向量, $PSecond(A, B)$ 表示 $PFirst(A, B)$ 中的分量按对应词在 B 中的词序排序生成的向量, $ReWord(A, B)$ 表示 $PSecond(A, B)$ 各相邻分量的逆序数。

句长相似度反映两个问句在长度形态上的相似程度。用 $LenSim(A, B)$ 表示问句 A 和 B 的句长相似度, 计算如公式 3-3 所示。

$$LenSim(A, B) = 1 - \left| \frac{Len(A) - Len(B)}{Len(A) + Len(B)} \right| \quad (3-3)$$

距离相似度用相同关键词在问句上的距离来衡量语句的相似度。用 $DisSim(A, B)$ 表示问句 A 和 B 的距离相似度, 计算如公式 3-4 所示。

$$DisSim(A, B) = 1 - \left| \frac{SameDis(A) - SameDis(B)}{Dis(A) + Dis(B)} \right| \quad (3-4)$$

其中, $SameDis(A)$ 表示 A, B 中相同的词在 A 中的距离, 若相同的词出现多次, 以产生的最大距离为准。 $Dis(A)$ 表示 A 中非重复词中最左到最右词之间的距离, 若该词出现多次, 以产生的最小距离为准。

由以上四部分可以加权得到问句的语法相似度计算如公式 3-5。

$$SyntaxSim(A, B) = \alpha * WordSim(A, B) + \beta * OrderSim(A, B) + \gamma * LenSim(A, B) + \mu * DisSim(A, B) \quad (3-5)$$

其中 $\alpha, \beta, \gamma, \mu$ 是常数, 且满足 $\alpha + \beta + \gamma + \mu = 1$, 显示 $Sim(A, B) \in [0, 1]$ 。

在语法相似度中, 一般不难理解词形相似度起着决定性的作用, 其他三个方面起着次要的作用, 因此 $\alpha, \beta, \gamma, \mu$ 取值时应该有 $\alpha \gg \beta, \gamma, \mu$, 本文选择的值为 $\alpha = 0.7, \beta = \gamma = \mu = 0.1$ 。

(2) 语义方法

问句语义相似度计算还是以词的语义计算为基础。词的计算采用了刘等人

介绍的基于《知网》的词语的相似度计算。

设两个问句 A 和 B , A 包含的词为 $w_{11}, w_{12}, \dots, w_{1n}$, B 包含的词为 $w_{21}, w_{22}, \dots, w_{2m}$, 则词语 $w_{1i} (1 \leq i \leq n)$ 和 $w_{2j} (1 \leq j \leq m)$ 之间的相似度表示为 $\text{sim}(w_{1i}, w_{2j})$ 。问句 A 和 B 之间的语义相似度可以根据公式 3-6 进行计算 ,

$$\text{SemanticSim}(A, B) = \frac{1}{2} \left(\frac{1}{n} \sum_{i=1}^n \max \{ \text{sim}(w_{1i}, w_{2j}) \mid 1 \leq j \leq m \} + \frac{1}{m} \sum_{j=1}^m \max \{ \text{sim}(w_{1i}, w_{2j}) \mid 1 \leq i \leq n \} \right) \quad (3-6)$$

(3)问句类型的方法

用 $\text{ClassSim}(A, B)$ 表示两个问句类别相似度 , 计算如公式 3-7 所示。

$$\text{ClassSim}(A, B) = \begin{cases} 1, & A.\text{smallClass} = B.\text{smallClass} \\ 0.5, & A.\text{bigClass} = B.\text{bigClass} \\ 0, & \text{Others} \end{cases} \quad (3-7)$$

其中 , $A.\text{smallClass}$ 表示问句 A 的细类类别 , $A.\text{bigClass}$ 表示问句 A 的粗类类别。

3.3.2 问句相似度计算的步骤

给予两个问句 A 和 B , 问句相似度的计算如公式 3-8 所示。在原有的语法方法和语义方法的基础之上 , 结合了问句类型的方法。两个问句的相似度计算的伪代码如表 3-5 所示 , $\text{hownet}(w_{1i}, w_{2j})$ 是调用基于《知网》的词语相似度计算方法^[37]。

$$\text{Sim}(A, B) = \alpha * \text{SemanticSim}(A, B) + \beta * \text{SyntaxSim}(A, B) + \gamma * \text{ClassSim}(A, B) \quad (3-8)$$

其中满足 $\alpha + \beta + \gamma = 1$, 且满足 $\alpha > \beta, \gamma$ 。

表 3-5 问句相似度计算的伪代码

输入 : 两个问句 A 和 B
输出 : $\text{Sim}(A, B)$
(1) 对 A 和 B 进行分词处理 , 去除一些停用词 , 得到 A 包含的词为 $w_{11}, w_{12}, \dots, w_{1n}$, B 包含的词为 $w_{21}, w_{22}, \dots, w_{2m}$;
(2) 计算 WordSim , OrderSim , LenSim 和 DisSim , 得到 SyntaxSim ;
(3) $\text{sis} \leftarrow 0$;
(4) $\text{Semantic} \leftarrow 0$;
(5) for ($i \leftarrow 1$; $i \leq n$; $i++$)
(6) for ($j \leftarrow 1$; $j \leq m$; $j++$)
(7) if $\text{sis} < \text{hownet}(w_{1i}, w_{2j})$
(8) $\text{sis} \leftarrow \text{hownet}(w_{1i}, w_{2j})$;
(9) $\text{Semantic} \leftarrow \text{Semantic} + \text{sis}$;
(10) $\text{SemanticSim} \leftarrow \text{Semantic} / n$;

```

(11)   $sis \leftarrow 0$  ;
(12)   $Semantic \leftarrow 0$  ;
(13)  for ( $i \leftarrow 1$ ;  $i \leq m$ ;  $i++$ )
(14)      for ( $j \leftarrow 1$ ;  $j \leq n$ ;  $j++$ )
(15)          if  $sis < hownet(w_{2i}, w_{1j})$ 
(16)               $sis \leftarrow hownet(w_{2i}, w_{1j})$ 
(17)           $Semantic \leftarrow Semantic + sis$  ;
(18)   $SemanticSim \leftarrow SemanticSim + Semantic/m$  ;
(19)   $SemanticSim \leftarrow SemanticSim/2$  ;
(20) 采用问题分类算法对  $A$  和  $B$  进行问题分类，得到  $ClassSim$  ;
(21) 根据公式 3-8，可以得到  $Sim(A, B)$  ;

```

下面将给出，用户输入问句，怎么从FAQ问句中得到最相似度的问句。

假设 Q 为用户输入问句， $corpus$ 为常问问题库FAQ。要找出与 Q 最相似的且大于阈值 w 的语句 S 。

如果FAQ问句的数目比较庞大，首先要进行的是问句特征词的扩展和候选问句的选取，这里不作为重点详述。主要介绍从候选问句选取最相似度的问句，步骤如表3-6所示。

表 3-6 FAQ 问答系统的执行伪代码

输入：用户问句为 Q ，FAQ 库中的问句 $Corpus$ ，和相似度阈值 w
输出：最相似的问句
(1) 对用户问句进行分词处理，选取关键词；
(2) 对关键词词进行扩展；
(3) FAQ 库候选问句集的查找，得到候选问题集 $\{h_1, h_2, \dots, h_n\}$ ；
(4) 分别计算得到 Q 与 h_i 的相似度值 $Sim_i(Q, h_i)$ ；
(5) 按照相似度值 $Sim_i(Q, h_i)$ 从高到低进行排序。如果存在大于阈值 w 的问句且 Q 和 h_i 不是一个肯定性问句和一个否定性问句，返回相似度值最高的问句对应答案给用户，否则返回 null；

3.3.3 时间复杂度分析

时间复杂度包括这三个部分，语法相似度、语义相似度和问句类型相似度分别的时间复杂度。

词形相似度的复杂度为 $O(nm)$ ，其中 n 和 m 分别为两个问句中词的数目。词序相似度的复杂度为 $O(c^2)$ ，其中 c 为问句中共同词的数目。句长相似度的复杂度为 $O(1)$ 。距离相似度的复杂度为 $O(c)$ 。

调用基于语义的方法的时间复杂度为 $O(nmh)$ ，其中 h 为调用基于知网计算的两个词语相似度的复杂度。

调用基于问题类型的方法，需要的时间复杂度为 $O(2q)$ ，其中 q 为调用的问题分类算法的时间复杂度。

最后，两个问句的时间复杂度为 $O(nm + c^2 + nmh + q)$ 。

3.4 实验结果及分析

实验用的数据集主要包括两个部分，第一个是整理的面试常问问题集，第二个用哈工大整理的问题集。

面试常问问题集是从网上下载的200句作为FAQ库，并手工标注问题类别。然后找了10个学生，每个学生列举出10个面试中被问到的问题，总共100句作为测试集。

哈工大整理的问题集中的训练集4966句作为FAQ库，都已标注过类别。从这4966句中随机选择2000句，然后对这些问句进行一系列的改变。比如，问句中词语进行同义词的替换，改变问法，和该主题的其他类似的问法等等。

实验过程中参数的取值， $\alpha=0.6$ ， $\beta=\gamma=0.2$ ，阈值 $w=0.65$ 。对比的算法：语义的方法采用金等人^[50]论文上的算法，多特征方法采用张等人^[35]论文上的算法，参数都是参照原论文。

3.4.1 简单几个问句的计算

以问句Q：“我们去北京的原因是什么？”为例，比较Q与FAQ库中的3个问题Q1、Q2和Q3，结果如表3-7所示。

表 3-7 简单几个问句的计算结果

FAQ 库中间句	语义	多特征	本文
Q1：为什么我们去北京？	0.728	0.683	0.778
Q2：什么时候我们去北京？	0.836	0.760	0.666
Q3：北京有什么好玩的地方？	0.724	0.651	0.63

从计算结果可以看出，不论是采用语义的方法，还是采用多特征的方法，都选用了Q2作为最后的结果。以人的主观性进行判断，不难发现问句Q1可能是我们想要的结果。采用了本文的方法，最后选择了Q1作为最后的结果，主要是由于只有Q1和用户问句的问句类项是一样的。

3.4.2 面试常问问题集

选用了面试常问问题集。选用的实验评价指标如1.4节所示。得到的实验结果如表3-8所示。

可以看出，由于本文的算法的值要比其他算法好，说明了考虑了问题的类型和句型对准确率的提高是有一定的帮助的。每个算法的 $Prec_2$ 都比 $Prec_1$ 要高，

说明方法的准确度跟选择问句返回的数目有关，返回的问句数目越多，准确度越高。同时，

表 3-8 面试常问问题集的实验结果

实验方法	Prec ₁	Prec ₂
语义	0.55	0.60
多特征	0.68	0.72
本文	0.79	0.81

3.4.3 哈工大数据集

选用了哈工大整理的数据集。考虑到FAQ库和测试集中的问句都已标注了问句类别，同时为了证明问题分类准确性对问句相似度计算的影响。第一种方法选择目前已有的算法进行问题分类^[49]，然后用于相似度计算，简称“方法1”；第二种方法，直接用标注的类别进行问句相似度计算，简称“方法2”。实验结果如表3-9所示。

从结果可以看出，方法1和方法2都比已有算法的准确度要高，说明了引入了问题类型对问句相似度的计算是有用的。同时由于方法2比方法1的准确度要高，说明了问题分类的准确对问句的相似度是有影响的。问题分类的准确度越高，问句相似度的准确度也越高，因此选择分类准确度高的问题分类算法也十分必要。

表 3-9 哈工大数据集的实验结果

实验方法	Prec ₁	Prec ₂
语义	0.52	0.54
多特征	0.65	0.68
方法1	0.74	0.75
方法2	0.79	0.80

3.4.4 问题分类准确度的影响

采用哈工大数据集，验证每个细类上的问句相似度计算的准确度，实验结果如图 3-1 所示，可以看出，不同细类上，准确度是有差别的，主要由于不同类别问题分类的准确度的不同。

实验过程中，把第 3.4.2 节用到的测试集中的一定比例的问题类型不变，其他问题的类型随机的设置为其他问题类型，从而用来验证问题分类算法总体情况的准确度对问句相似度计算准确度的影响。实验结果如图 3-2 所示。

从图可知，在问句分类的准确度从 0.2 到 0.4 的过程中，随着问句分类准确度的提高，问句相似度计算准确度增加的很慢，主要是因为这一段问句分类的

准确度都比较低，问句类型对问句相似度计算的过程的影响比较小，又由于综合了多特征的方法，多特征方法占据着主要的权重。随后，问句相似度计算准确度增加的较快，这是因为问句分类正确可以缩小 FAQ 库中其他类型问句的干扰。总之，问句相似度计算的准确度随着问句类型准确度的提高而提高。

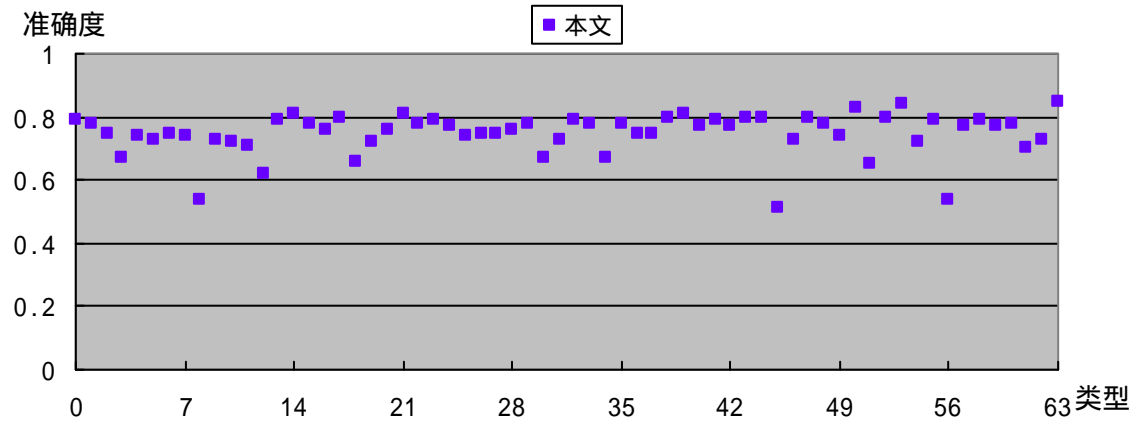


图 3-1 具体各个类上分类精度比较

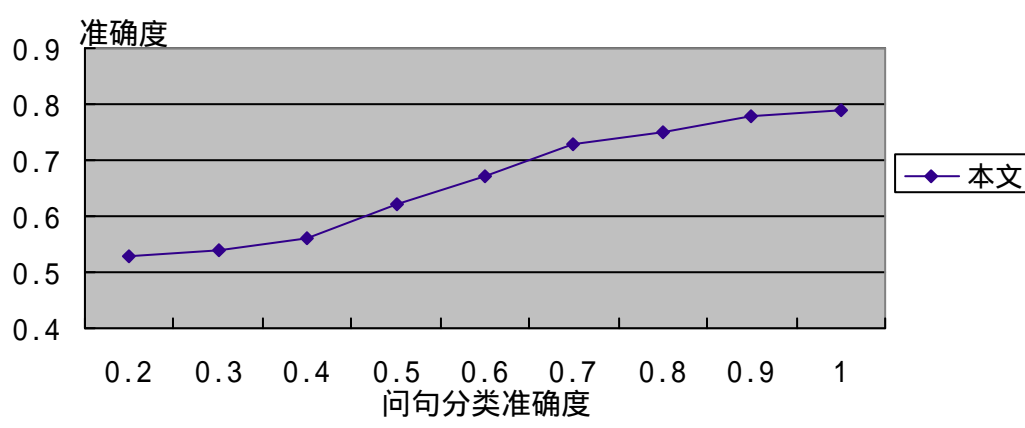


图 3-2 问句分类准确度对问句相似度计算准确度的影响

3.4.5 答案数目的影响

研究最终选择答案的数目对不同方法的影响。数据选择 3.4.2 节使用的数据集。答案数目取值从 1 到 6，实验结果如图 3-3 所示。可以看出，不同方法都随着答案数目的增加，问句相似度计算的准确度增加，然后不变。但是，语义的方法和多特征的方法不变后的准确度值与初始值的差值要大于本文方法差值，主要是由于本文的方法考虑到问题类型，缩小了候选问题答案的规模。

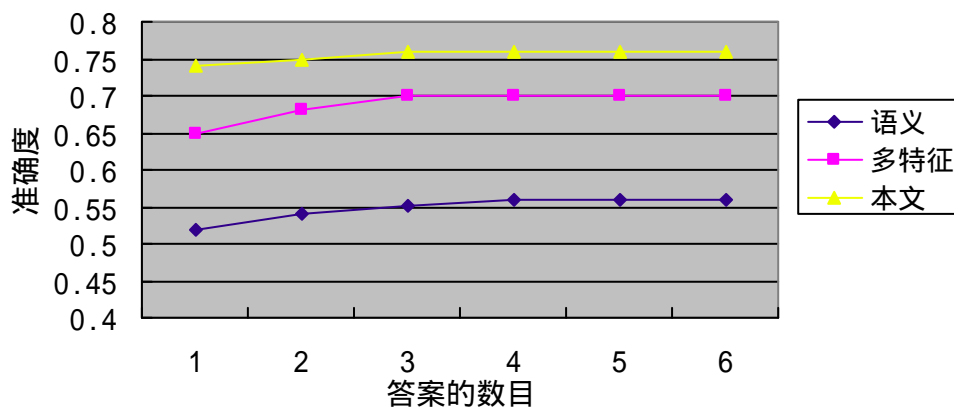


图 3-3 答案的数目对问句相似度计算准确度的影响

3.5 本章小结

本文通过分析，发现普通陈述句的相似度计算方法并不适用于FAQ问答系统中的问句相似度计算，从而，引入问句特征来间接刻画问句与答案句的匹配，提出了一种新的问句相似度计算方法，实验结果也证明了该方法的有效性。由于，该方法会受到问题分类准确度的影响，在考虑结合特定的领域设计FAQ问答系统时，设计针对该领域的问题分类算法从而用于问句相似度计算，作为下一步研究的内容。

第4章 基于词语褒贬倾向的问句相似度计算

上一章讲述了 FAQ 问答系统中问句相似度计算和普通陈述句相似度计算的不同之处，引入了问句的特征信息来计算问句的相似度计算。本章着重解决目前的基于词语褒贬性的问句相似度计算方法，采用的词语语义计算方法中，对义或反义的词语与同义或近义的词语对问句相似度计算结果都起着正作用，从而有可能导致截然相反的两个问句都有着很高的相似度，而这对于 FAQ 自动问答系统就是两个截然不同的问题，为此，提出了一种基于词语褒贬倾向的问句相似度计算方法。

4.1 引言

网络等信息技术的快速发展，很多语句都在一定程度上体现了人们的情感色彩，语句中表现出了明显的褒贬性，所以，许多学者提出了语句情感倾向识别的判别方法^[54,55,56]。目前，在分析语句的情感倾向，都是以分析词的褒贬性为基础。语句的褒贬性都是以词的褒贬性为基础，所以，词的褒贬性研究也得到了广泛的研究^[57]。而目前的问句相似度计算方法都没有考虑到词的褒贬性对计算结果的影响，所以本章以 FAQ 系统为研究背景，探讨词的褒贬性对问句相似度计算的影响。

例如，采用目前的基于《知网》的问句相似度计算方法，问句“上网对人有什么好处”和“上网对人有什么坏处”的相似度值为 0.9，但是，可以看出这是两个不同的问题。主要是因为采用的计算方法中“好处”与“坏处”的相似度为 0.814815，导致了最后的语句相似度值很高。本章引入了江等人^[57]的基于《知网》的词语相似度计算方法运用到问句相似度计算中，该方法考了中文词语之间的褒贬性，计算的词语取值范围为 $[-1,1]$ ，相近的词语取值为正数，相反的词语取值为负数。

目前，在因特网上有很多多语言内容的文档。如，在因特网上至少有 150 种语言被广泛的使用，创造了海量的 web 网页。其中，68.4%的 web 网页都是用英文写的，和大约 16%是用中文、日语、德语和法语。这说明了拥有处理多语言的信息获取系统是非常必要的。比如，问句“人们 love 北京的原因是什么？”和“人们讨厌北京的原因是什么？”。采用的目前的词语相似度计算方法，得到的“love”和“讨厌”的值都有很高的相似度，必定为影响两个问句的相似度计算结果。

为了使问句相似度计算方法处理的更广泛，本章根据《知网》的双语特点，结合江等人的词语相似度计算方法，给出了计算中文词语与英文单词和英文词语之间的褒贬性计算，得到的相似度取值范围也是从 $[-1,1]$ ，考虑到了中文词语与英文词语的褒贬性。最后，选取对问句相似度值影响比较大的词语作为组合，

来得到两个问句的相似度值 ,从而达到反义或者对义的词语对问句起到负作用 ,同义或者近义的词对问句起到正作用的目的。实验表明 , 与传统的问句相似度计算方法 , 该方法更符合现实情况和更能准确的计算问句相似度。

本文第4.2节给出了词语的相似度计算 ; 第4.3节给出问句相似度的计算方法和步骤 ; 第4.4节从不同方面来验证算法的有效性 ; 最后 ,4.5节给出结论。

4.2 词语相似度计算

《知网》的基本形式是对中文词语和英文词语的双语言解释和描述。词语的意义不是通过其他词语来解释、说明 , 而是通过义原来确定。例如 , 在《知网》中关于“ 海岸(coast) ”的解释如图 4-1 所示。“ 海岸 ”的概念定义指出海岸是一个陆地和靠近水域。同时 , 可以看出 , DEF 结构包含了两部分 : (1)基本义原 , 如 , “ land|陆地 ” ; (2)其他义原 , 如 , “ {BeNear|靠近 : existent={~} , partner={waters|水域}} ”。

NO. =048973	//概念 ID
W_C=海岸	//中文词
G_C =N [hai3 an4]	//词性和拼音
W_E=coast	//相一致的英文词语
G_E=N	//英文词性
DEF={land 陆地: {BeNear 靠近 : existent={~},partner={waters 水域}} //概念定义	

图 4-1 《知网》中海岸(coast)的概念定义

图 4-2 截取了以“ entity|实体 ”为根节点的树中的一个分支(英文没标注出来) , 李峰等人^[58]在刘群公式的基础上考虑了义原的层次深度 , 如公式 4-1 所示。

$$sim(p_1, p_2) = \frac{a * \min(depth(p_1), depth(p_2))}{dist(p_1, p_2) + \alpha * \min(depth(p_1), depth(p_2))} \quad (4-1)$$

其中 , p_1, p_2 表示两个义原 , $depth(p_1)$ 表示 p_1 距离根节点的层次 , $dist(p_1, p_2)$ 表示它们的路径长度 , $\min(depth(p_1), depth(p_2))$ 表示 p_1 和 p_2 两者在义原树中深度的最小值 , α 是一个调节参数。

如图 4-2 所示 , 用此来讨论义原的相似度。公式 2-3 给出了刘群等人^[37]的关于义原相似度的计算公式 , 该公式只考虑了义原的距离 , 没有考虑义原的深度。例如 “ 人 ” 和 “ 兽 ” 的相似度等于 “ 生物 ” 与 “ 非生物 ” , 这明显是不合理的。同时李峰等人计算的公式也是不合理的 , 如 “ 动物 ” 和 “ 植物 ” 的义原深度都为 2 , “ 动物 ” 和 “ 牲畜 ” 的深度分别为 2 和 4 , 从人的直观可以看出 , 后者的相似度要大于前者 , 但是用公式 4-1 没有了区分度。主要原因是 $\min(depth(p_1), depth(p_2))$ 只考虑了义原深度的较小值。

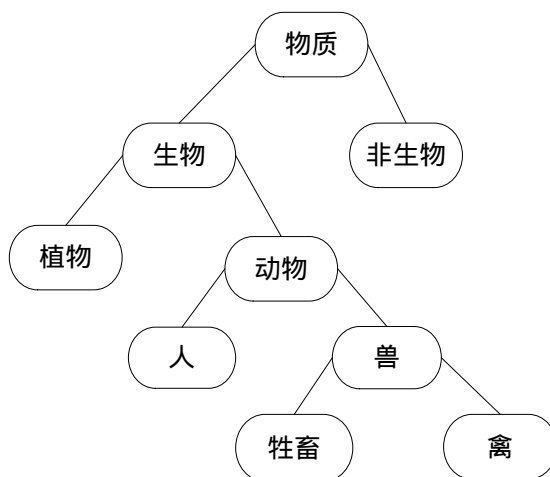


图 4-2 根节点为“entity|实体”的树中的一个分支

为此，吴等人^[59]认为词语的相似度的大小也与节点所处的深度有关：相等距离的两个词语，词语相似度应该随着所处层次的总和的增加而增加，随着他们之间层次差的增加而减少。如从人的主观角度上理解，“动物”和“禽”的相似度要大于“动物”和“物质”，尽管两对词语之间的距离都是 2。

$$sim(p_1, p_2) = \alpha * (depth(p_1) + depth(p_2)) / (\alpha * (depth(p_1) + depth(p_2)) + dist(p_1, p_2) + |depth(p_1) - depth(p_2)|) \quad (4-2)$$

由于，本章需要考虑到词语的褒贬倾向，采用了江等人^[57]论文中提到的词语相似度计算方法，其中计算义原 p_1 和 p_2 的相似度，如公式 4-2 所示。该方法在原有的基础之上，进一步考虑了义原的对义、反义关系，主要分下面两种情况：如果两个义原直接是对义、反义关系，相似度为-1；如果两个义原路径中存在对义、反义关系，则相似度为 $-1 * s_sim(p_1, p_2)$ 。

$$s_sim(p_1, p_2) = \alpha * (depth(p_1) + depth(p_2)) / (\alpha * (depth(p_1) + depth(p_2)) + dist(p_1, p_2)' + |depth(p_1) - depth(p_2)|) \quad (4-3)$$

其中， $dist(p_1, p_2)'$ 把距离义原 p_1 和 p_2 最近的一对对义、反义义原节点看作同一个节点后得到的路径长度。基于此，该方法把词语相似度的取值范围从原来的 $[0, 1]$ 变换到 $[-1, 1]$ 之间：意思越接近的词相似度越接近 1，意思越相反的词相似度越接近 -1。例如，部分词的结果表 4-1 所示，可以看出，一些看似存在一定程度上反义的词语，实际结果也确实都为负值。该方法在词性极性识别中，得到很好的实验效果。

对于中文词语与英文词语的相似度计算也有许多工作。不过，Xia 等人^[60]计算中文词语与英文词语相似度的范围不是取值到 1 之间的值，不利于用于计算问句相似度，更没考虑词极性对相似度值得影响。同样，Dai 等人^[61]的方法可以用于英文词语之间的相似度，也没考虑词极性对相似度值的影响，不能很好的用于计算问句相似度。

下面介绍中文与英文词语的相似度和英语与英文词语的相似度。Hownet 提

供了 API 接口用于寻找中文词语或者英文词语的概念。例如 ,英文单词 coast(海岸) ,可以找到图 4-1 所示的概念 ,找到对应的中文词语“ 海岸 ” ,这样计算“ coast ”和其他词语的相似 ,就可以采用“ 海岸 ” 进行替换。如果一个英文单词对应着多个中文词语 ,这是需采用公式 4-4 进行计算。

表 4-1 中文词语相似度计算结果

词语一	词语二	刘等人	公式 4-1
高尚	卑鄙	0.788360	-0.912500
医生	患者	0.788360	-0.547805
美丽	贼眉鼠眼	0.814815	-1.000000
美丽	优雅	0.788360	0.7500000

设英文单词 W_1 对应的多个中文词语为 w_1' 、 w_2' 、...、 w_n' , 中文词语为 W_2 。

$$sim(W_1, W_2) = \eta \max_{i=1 \dots n} |sim(w_i', W_2)| \quad (4-4)$$

其中 η 表示 $sim(w_i', W_2)$ ($1 \leq i \leq n$) 中绝对值最大的数所带的符号。

如果两个问句中都存在着英文词语 , 就会出现英文词语与英文词语的相似度计算 , 采用公式 4-5 计算。

设英文单词 W_1 对应的多个中文词语为 w_{11}' 、 w_{12}' 、...、 w_{1n}' , 英文单词 W_2 对应的多个中文词语为 w_{21}' 、 w_{22}' 、...、 w_{2m}' 。

$$sim(W_1, W_2) = \lambda \max_{i=1 \dots n, j=1 \dots m} |sim(w_{1i}', w_{2j}')| \quad (4-5)$$

其中 λ 表示 $sim(w_{1i}', w_{2j}')$ ($1 \leq i \leq n$, $1 \leq j \leq m$) 中绝对值最大的数所带的符号。

4.3 问句相似度计算

通过对语句结构的分析和先前介绍的中英文词语相似度的计算来达到中英文问句相似度的计算。首先是通过分词对问句进行分词 , 去除停用词 , 选择实词 , 包含名称、动词、形容词、数词、量词和代词^[62]等 , 和一些英文字符。

设两个问句 A 和 B , A 包含的词为 w_{11} 、 w_{12} 、...、 w_{1n} , B 包含的词为 w_{21} 、 w_{22} 、...、 w_{2m} 。

设 $Sim(A, B)$ 为问句 A 、 B 相似度的特征矩阵 ,

$$AB = A * B^T = \begin{pmatrix} w_{11}w_{21} & \dots & w_{1n}w_{21} \\ \vdots & & \vdots \\ w_{11}w_{2m} & \dots & w_{1n}w_{2m} \end{pmatrix} \quad (4-6)$$

其中 , $w_{1i}w_{2j} = sim(w_{1i}, w_{2j})$ 。

计算问句相似度 , 首先遍历相似度特征矩阵 , 找出绝对值最大的元素 , 即

$|w_{1i}w_{2j}|$ 最大的值，再删除该值所在的行和列中的其他元素，循环执行，直到矩阵中没有元素为止，此时可得到词语最大组合序列。

$$\max L = \{ \text{sim}_{\max_1}, \text{sim}_{\max_2}, \dots, \text{sim}_{\max_k} \} \quad (4-7)$$

其中， k 等于 n 和 m 之间的小者。

最后，可以获得问句 A 和 B 的相似度值，如公式 4-8。

$$\text{Sim}(A, B) = \frac{1}{k} \sum_{i=1}^k \text{sim}_{\max_i} \quad (4-8)$$

为了更好的理解问句相似度的计算过程，给出了一个例子描述两个问句的计算过程。两个问句分别为“网络对人有什么好处”和“Network 对人造成什么坏处”。分词后的结果“网络、对、人、有、什么、好处”和“网络、对、人、造成、什么、坏处”。然后，计算词与词的相似度，得到相似度矩阵。

1	0	0.24	0.074	0.266	0.2
0	1	0	0	0	0
0.24	0	1	0.768	0.175	0.616
0.074	0	0.074	0.126	0.044	0.044
0.266	0	0.175	0.044	1	0.175
0.044	0	0.618	0.044	0.048	-0.833

由矩阵里面的相似度值，根据值绝对值的大小进行选择，得到 $\max L = \{1, 1, 1, 1, -0.833, 0.126\}$ 。

最后，由公式 4-8 进行计算，可以得到两个问句的相似度值为 0.549。

4.4 实验结果及分析

首先，给出了中英文词语之间的词语褒贬性实验。然后，验证把词语褒贬倾向的计算用于问句相似度计算后的效果。最后，结合第 3 章提出的算法综合考试问句相似度计算的准确度。

基于《知网》的中文词语与英文词语和英文词语之间的词语褒贬倾向的计算，实验结果如表 4-2 所示。该方法不同于其他一些基于《知网》的用于语句相似度计算的词语相似度计算方法，因为考虑到对义词、反义词对中英文词语相似度的影响，其取值范围为 $[-1, 1]$ 。

从前三组词语对的计算结果，可以看出对中文词语与英文词语的相似度计算大致满足了人的正常理解。从“医生”和“patient”的相似度为 -0.547805，可以看出考虑了对义词、反义词后，对相似度值的影响。同理，处理英文词语之间的相似度也大致符合人的主观理解。

表 4-2 中英文词语相似度计算结果

词语一	词语二	相似度值
男人	man	1.000000
男人	job	0.057017
男人	monk	0.8333333
医生	patient	-0.547805
run	leap	0.818182
jewel	treasure	0.500000
health	sick	-0.875000
health	deaf	-0.277230

用户提出的问句为 Q：“上网对人有什么好处？”为例，假设 FAQ 库中有 3 个问题，分别为 Q1，Q2，Q3。表 4-3 中列出了相似度计算结果。

表 4-3 用户问句与 FAQ 库中间句的语句相似度计算结果

FAQ 库中间句	语义	本文
Q1：网络对人有什么坏处？	0.9	0.63
Q2：网络对人有什么好处？	0.98	0.96
Q3：Internet 对人有什么好处？	0.81	0.93

从人的主观可以看出，Q1 和 Q 是两个不同的问题，但是相似度值极高，很容易造成 FAQ 问答系统的误差。而本文采用的方法，相似度值仅为 0.63，相对语义的方法，得到的值更合理。主要是因为语义的方法计算的词“坏处”和“好处”为 0.81，而本文的计算的结果为-0.833，所以导致最后结果的差距很大。本文的方法考虑了词的极性，并用于问句的相似度计算对计算问句是有用处的。

Q3 和 Q2 其实意思差不多，但是，用语义的方法，与 Q 的相似度值差距就很大。采用本文的方法，两个计算差值只为 0.03，更加符合实际的需求。主要是因为“语义”的方法不能够处理中文词语与英文词语的相似度，采用的相似度值为 0。实际上，Internet 在汉语中也是很常用的词语，Internet 和网络的意思差不多。由于本文的方法能够处理中文词语和英语词语的相似度，所以取得较接近实际的结果。

为了更充分的验证本文方法的准确度，本文采用了第 3 章用到的整理后的哈工大的数据集 data1。还对 data1 进行进一步的修改，主要是由于这个数据集中几乎没有英文单词的出现。从测试集中随机抽取 300 个问句找出一个词用意思相似的英文词进行替换，改后的数据集称为 data2。评价标准只选择了 $Prec_1$ 作为评价标准。实验结果如图 4-3 所示。

“语义/data1”和“本文/data1”：(1)两个方法的准确率差别很小，是因为

data1 数据集中褒贬倾向的干扰句很少；(2)准确度略微提高 1%，说明了考虑了词的褒贬倾向有利于问句相似度计算。

“语义/data2”和“本文/data2”：两者的准确率差别有点大，因为数据集中部分问题包含了英文词语，而由于语义方法不能处理英文词语褒贬性导致准确率下降。

“语义/data1”和“语义/data2”：在 data2 数据集中准确度降低，主要是因为对 data1 数据集进行了修改，增加了许多英文词语，导致准确度的下降。

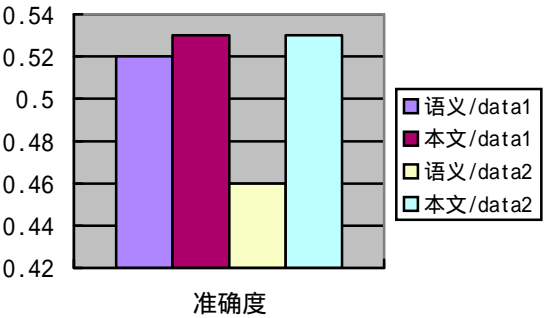


图 4-3 问句相似度计算实验结果

为了结合第 3 章中提到的方法，在数据集 data2 上做了进一步的实验。方法 3 是方法 1 中的语义方法用本章提到的语义方法替换。方法 4 是方法 2 中的语义方法用本章提到的语义方法替换。

从图 4-4 可以看出，方法 1 和方法 3 的准确度值低于原来的在 data1 中的准确度值，因为采用的问题分类算法对问题中包含有英文词语的问题进行分类，分类准确度降低所致。

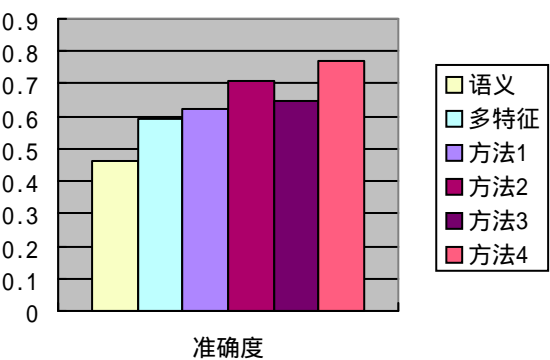


图 4-4 不同方法实验结构的对比

方法 3 的值高于方法 1 和方法 4 的值高于方法 2，说明引入了本文的语义方法能够提高了相似度计算的准确度。主要是由于考虑了问题中英文词语的褒贬倾向，原来的语义方法并不能够处理英文词语的褒贬倾向。

总之，可以看出，都比原来的语义方法和多特征方法的准确度要高，说明新方法可以提高从 FAQ 库中找到匹配问题的准确度。

4.5 本章小结

本章的主要内容是研究问句中存在反义词、对义词对问句相似度计算的影响。在 FAQ 问答系统中，如果不考虑反义词、对义词，有可能导致两个相反的问候句也有着很高的相似度，势必会影响 FAQ 问答系统的性能。所以，本文运用了新的基于《知网》的词语相似度计算方法到问句的相似度计算中，可以降低相反的问候句成为相似的问候句的可能。实验结果表明本文中的方法具有更好的准确度，说明本文中的语义方法发挥了较好的作用。

第5章 FAQ 自动问答原型系统

前面两章详细介绍了基于问句特征的问句相似度方法和改进的基于《知网》的问句语义相似度计算方法，并分别在多个数据集上进行了大量的实验。本章主要通过应用这两个方法设计了一个可以返回相似问句的 FAQ 自动问答原型系统，详细论述了构建 FAQ 自动问答系统的每一个步骤。

5.1 系统总体设计介绍

本节主要介绍系统所需的技术和总体框架。由于选用的是哈工大整理的 HQ 问题集，对应的问题集并没有对应的答案，运行结果中返回的是满足要求的相似问句，而不是对应的答案。

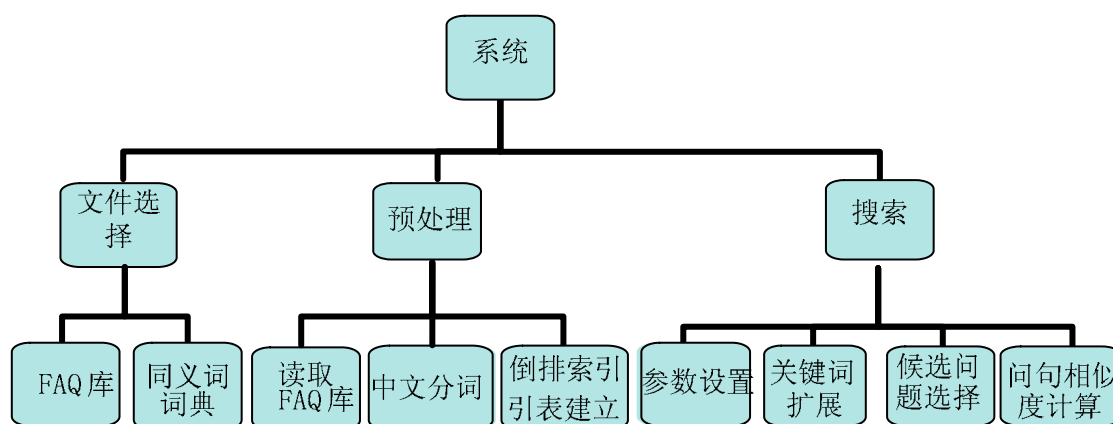


图 5-1 系统的总体框架图

本系统使用 Java 语言实现，总体框架如图 5-1，主要分为两大部分：文件选择、预处理和搜索，其中文件选择包括 FAQ 库和同义词词典，预处理包括读取 FAQ 库、中文分词和倒排索引表的建立，搜索阶段包括参数的设置、关键词扩展、候选问题的选择和问句相似度计算四个部分。

假设 Q 为用户输入的问句， $corpus$ 为常问问题库FAQ。要找出与 Q 最相似的且大于阈值 w 的语句 S 。该过程可以用下面公式进行描述，

$$S = \max_{S \in corpus} sim(Q, S) > w \quad (5-1)$$

其中， S 表示 $corpus$ 中任意的语句。

5.1.1 关键词的扩展

由于汉语语义的丰富性，好多词语都有很多同义词和近义词，如果直接用问句中的词进行检索，有可能导致相近的问句没有检索到。例如，问句“电脑是谁发明的？”，和问句“计算机是谁发明的？”，可以看出，两个问句一个使用的是“电脑”，另一个使用的是“计算机”，如果不进行关键词的扩展可能会降低系统的召回率。

适当的关键词扩展可以提高系统的召回率，但是不适当的扩展可能影响系统的准确度和效率。因此，目前通常采用的基于《同义词词林》的关键词扩展。后来哈尔滨工业大学信息检索实验室对《同义词词林》进行扩展，得到了《同义词词林的扩展版》，大约收录词语近 7 万条。本系统主要针对《同义词词林的扩展版》进行关键词的扩展。

5.1.2 FAQ 库候选问题集的查找

FAQ 库的容量一般都比较比较大，如果采取用户提交的问句与 FAQ 中的问句都进行语句相似度计算，虽然可以取得较好的效果，但会大大增加查询开销的时间。为此，需要进行问句相似度之前筛选 FAQ 库，提取候选问句集，然后基于候选问题集进行语句相似度计算，从而减少检索的时间，提高检索的效率。

假设用户问句扩展后的所有词为 $W=\{w_1, w_2, \dots, w_n\}$ 。FAQ 库中共有 m 个问句，第 $i(1 \leq i \leq m)$ 个问句含有 n_i 个词为 $Q_i=\{q_1, q_2, \dots, q_{n_i}\}$ 。将 W 和 Q_i 之间重叠词的个数记为 $Num_i=|\{w_1, w_2, \dots, w_n\} \cap \{q_1, q_2, \dots, q_{n_i}\}|$ ，最后，根据 Num_i 对 FAQ 进行筛选提取候选问题集。

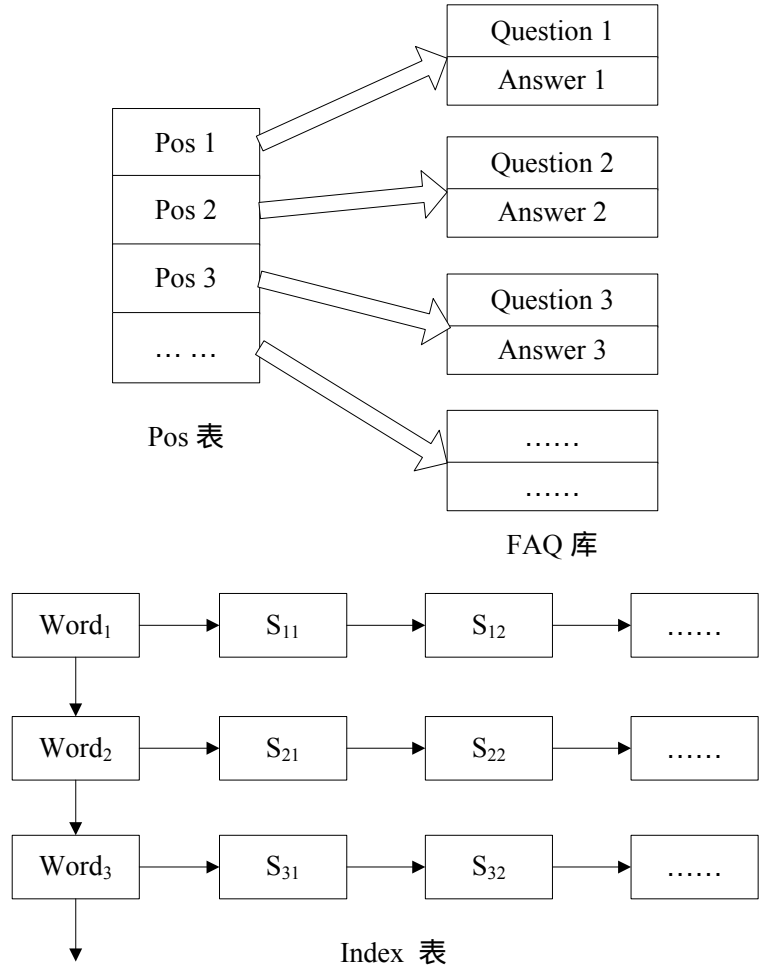


图 5-2 FAQ 库数据结构示意图

为了加快 Num_i 的计算,本文中构建了基于 FAQ 库的以词语为中心的倒排索引表,如图 5-2 所示。首先,把所有的常问问题集中的问题答案对 Question-Answer 都存储于 FAQ 库当中。接着建立 Pos 表,表中每一项都包含一个指针,分别指向 FAQ 库中一个问句。然后,对 FAQ 库中的问句进行分词和去停用词处理,建立以词为单位的索引(Index 表)。Index 表中 $Word_1$ 、 $Word_2$ 、.....是 FAQ 库中的问句进行处理和经过排序后生成的链表,其中每个 $Word_i$ 指向一个 S 链表,该链表中的每个节点是 FAQ 库中包含该词的问句的索引号。

在实际的检索过程中,可以采用折半查找的方法寻找用户问句中的词 w_i 在 Index 链表中中的位置。假设找到的节点为 $Word_k$,根据 $Word_k$ 所指向的 S 链表,就能够知道哪些问句包含 $Word_k$ 。对 W 中的每一个词都采用这样的计算方式,就可以得到 Num_i 的值,从而得到候选问题集。

5.1.3 问句相似度的计算

假设用户问句为 q , 候选问句包含 n 个问句,用 $\{h_1, h_2, \dots, h_n\}$ 表示。首先,对用户问句 q 分别和候选问句进行相似度计算。然后,把计算出来的结果按照相似度值的大小进行排序,找出大于阈值且相似度值最大的一个候选问句对应的答案返回给用户。

5.2 系统功能介绍

图 5-3 是系统的主界面,其中最上面的是菜单栏,菜单栏下面是用户提出问题的文本输入框,其右边是搜索按钮,中间是用来显示返回的问句,最下面是参数设置,包括阈值和返回问句的数目。

用户在使用本文中介绍的系统时,首先需要从菜单“文件”中,选择好 FAQ 库,如图 5-4 所示。然后,点击“预处理”,就会对 FAQ 库中的问句建立倒排索引表。接着,用户设置好“参数”,选择“问句相似度计算方法”,输入想要提出的问题,点击“搜索”即可。这一步,首先对问题进行分词,然后是关键词的扩展,接着是从倒排索引表中选择候选问句集合,最近是问句相似度计算,根据参数选择合适问句显示出来。问句相似度计算给出了四种方法,分别是基于语义的、多特征的和本文提出的两种相似度计算方法。

下面给出搜索实例,用户输入:

- (1)设置 FAQ 库:浏览并上传数据,本实例中选择的是 HQ 数据集。
- (2)预处理:对 FAQ 库的问句进行处理。
- (3)输入问题:用户自行设定想要问的问题,如本例中,“马克思哪年出生”。
- (4)参数和算法选择:设置阈值和返回问句的数目,如本例中,分别为“0.75”和“1”;选择问句相似度计算方法,如图 5-5 选择的是多特征的问句相似度计

算方法，图 5-6 选择的是基于问句特征的问句相似度计算方法。

(5)搜索：选择大于阈值，最相似的问句及计算的相似度值显示到界面上。

如图 5-5 和图 5-6 所示的结果。



图 5-3 系统的主界面

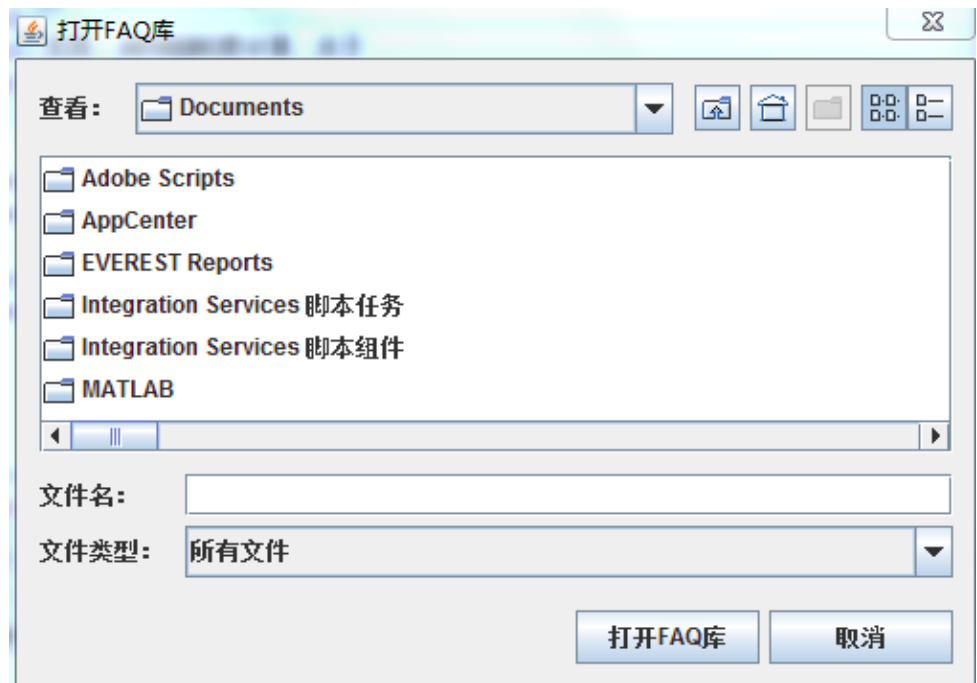


图 5-4 FAQ 库的打开

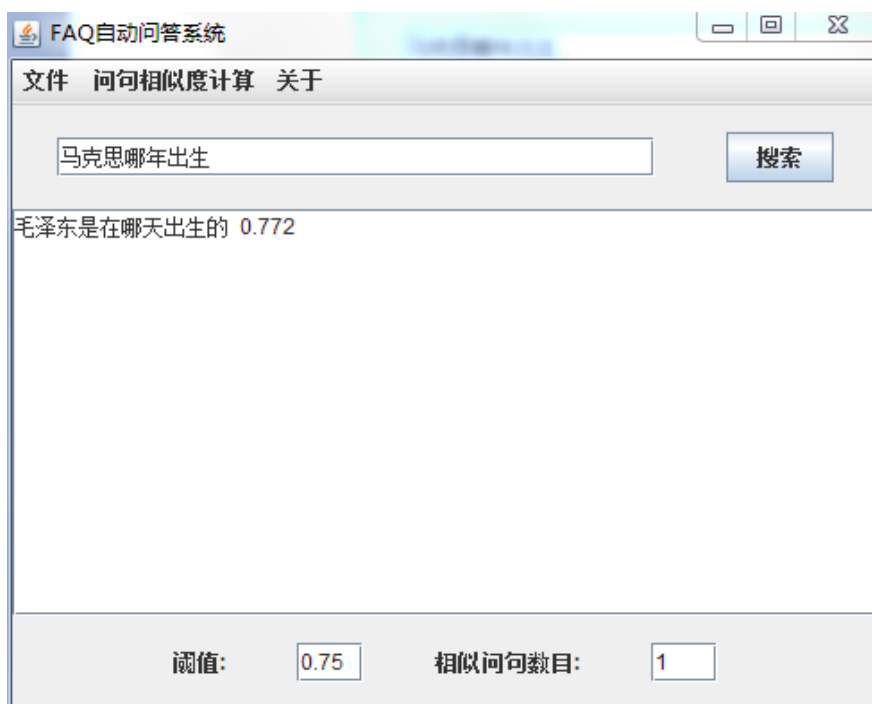


图 5-5 采用多特征问句相似度计算方法的结果



图 5-6 采用基于问句特征的问句相似度计算结果

第6章 总结和展望

6.1 总结

FAQ问答系统作为问答系统的重要模块，主要目的就是已有的“问答-答案”FAQ库中直接搜索问题，如果存在相似问题就返回对应的答案，从而可以省去复杂的检索过程。FAQ问答系统的核心技术是问句相似度计算，其直接关系到FAQ问答系统的性能和效率。因此本文研究工作的出发点是提高问句相似度计算的准确度。

目前，FAQ问答系统中的问句相似度计算只考虑了语句的语义相似，可能导致高相似的问句但确不是相似的问题。所以，需要深层次的挖掘语义信息，从而减少FAQ自动问答系统的误判。本文打算从问句所对应的答案来协助获取语义信息，但是，为在问题的匹配阶段，答案是未知的，要想深层的挖掘语义信息，是非常困难的。所以，本文为了间接刻画答案句的特征，选择了从问句的类型和问句的句型着手，在原有的多特征方法的基础上，提出一种新的问句相似度计算方法。问句的特征主要考虑了以下两个方面，一方面是问题分类算法分类后的类型，另一方面是问句的否定性识别。然后，给出了新方法和其他方法的准确度进行对比，实验结果表明，引入问题类型可以提高问句相似度计算的准确度。

随着网络等信息技术的快速发展，很多语句表现出了明显的褒贬性。如果FAQ自动问答系统不能够区分问句之间的反义、对义词，有可能会把截然相反的两个问句当成相似的问句，影响FAQ问答系统的性能。所以，本文研究的另一个重点就是对义或者反义词语对问句相似度的影响。为此，提出了一种基于词语褒贬性的问句语义相似度计算方法，方法中运用了新的词语相似度计算方法，同时也扩展到处处理中文词语与英文单词和英文单词之间的对义或者反义词语之间的相似度和同义或者近义词语相似度对问句相似度结果的影响。实验表明，该方法能够提高问句相似度计算的准确度。

最后，设计并实现了FAQ自动问答原型系统，把本文中提到了两种问句相似度计算方法运用到系统中，为以后更加深入的研究提供一个平台。

6.2 展望

本文方法虽然在问句相似度计算准确度上得到了一定的提高，但是仍存在一些不足：

(1)第3章提到的方法会受到问题分类准确度的影响。在具体的FAQ问题领域，可以考虑设计特定领域的问题分类算法，然后用于问句相似度计算。问句相似度计算方法的准确度有所提高，可能还不能达到满意的精度，具体方法的完善仍需要一定的研究。可以考虑把更多问句处理的方法加入进来，如句法结

构信息，以获得更加有效的特征进行问句相似度计算，使得问句相似度计算的精度得到进一步的提高。

(2)如果《知网》里面不包含问句中词语，就不能够计算词的相似度，就会影响问句的相似度，所以处理语义词典中的未登录词语也应该考虑。同时，本文中主要是为了提高问句相似度计算的精度，并没有对方法的时间性能进行比较。

(3)实验中没有合适的进行中文问句相似度计算的问题集，不能很好的评价问句相似度计算的精度。

因此，下一步，可以从下面两方面考虑：《知网》中未登录词的相似度计算运用到问句相似度计算中；整理更多的问题集，以便在不同问题集上进行实验，尽量减少由于问题集不合适造成的误差。问句相似度的精度仍有很大的上升空间，以后的研究工作仍任重道远。

参考文献

- [1] Dang H T, Lin J, Kelly D. Overview of the TREC 2007 question answering track[C]. Proceedings of TREC. 2007 (5.3):3.
- [2] 郑实福,刘挺,秦兵等.问答系统综述[J].中文信息学报,2002,16(6):46-52.
- [3] Voorhees E, Tice D M. The TREC-8 question answering track evaluation[C]. Proceedings of The Eighth Text REtrieval Conference (TREC-8), http://trec.nist.gov/pubs/trec8/t8_proceedings.html. 1999.
- [4] Wu L, Huang X, Zhou Y, et al. FDUQA on TREC2003 QA task[C]. Proc. of TREC, 2003.
- [5] Xu H, Zhang H, Bai S. ICT Experiments in TREC-11 QA Main Task[C]. Proceedings of Text REtrieval Conference, 2002.
- [6] Huang G T, Yao H H. A system for Chinese question answering[C]. IEEE/WIC International Conference on Web Intelligence, 2003: 458-461.
- [7] Huang G T, Yao H H. Chinese question-answering system[J]. Journal of Computer Science and Technology, 2004, 19(4): 479-488.
- [8] 张刚,刘挺,郑实福等.开放域中文问答系统的研究与实现[C].中国中文信息学会二十周年学术会议,2001,11.
- [9] 张琳,胡杰.FAQ问答系统语句相似度计算[J].郑州大学学报:理学版,2010,42(1):57-61.
- [10] 钟敏娟,万常选,刘爱红.基于词共现模型的常问问题集的自动问答系统研究[J].情报学报,2009,28(2):242-247.
- [11] Chen L, Shen R M. FAQ System in Specific Domain Based on Concept Hierarchy and Question Type[C]. International Conference on Computational and Information Sciences, 2011:281-284.
- [12] Islam A, Inkpen D. Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity[J]. ACM Transactions Knowledge, Discovery Data, 2008, 2(2):10-34.
- [13] Meadow C T, Kraft D H, Boyce B R. Text information retrieval systems[M]. Academic Press, Inc., 1999.
- [14] Okazaki N, Matsuo Y, Matsumura N, et al. Sentence extraction by spreading activation through sentence similarity[J]. IEICE TRANSACTIONS on Information and Systems, 2003, 86(9): 1686-1694.
- [15] Chiang J H, Yu H C. Literature extraction of protein functions using sentence pattern mining[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(8): 1088-1098.

- [16] Jurafsky D, James H. Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech[J]. 2000.
- [17] Landauer T K, Laham D, Rehder B, et al. How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans[C]. Proceedings of the 19th annual meeting of the Cognitive Science Society, 1997: 412-417.
- [18] Landauer T K, Foltz P W, Laham D. Introduction to Latent Semantic Analysis[J]. Discourse Processes, 1998, 25(2-3): 259-284.
- [19] Burgess C, Livesay K, Lund K. Explorations in Context Space: Words, Sentences, Discourse[J]. Discourse Processes, 1998, 25(2-3): 211-257.
- [20] Hatzivassiloglou V, Klavans J L, Eskin E. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning[C]. Proceedings of the 1999 joint sigdat conference on empirical methods in natural language processing and very large corpora, 1999: 203-212.
- [21] Hatzivassiloglou V, Klavans J, Eskin E. Detecting Similarity by Applying Learning over Indicators[C]. Proc. 37th Ann. Meeting of the Assoc. for Computational Linguistics, 1999.
- [22] Li Y H, Mclean D, Bandar Z A, et al. Sentence Similarity Based on Semantic Nets and Corpus Statistics[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(8):1138-1150.
- [23] Dao T, Simpson T. Measuring Similarity between Sentences. Technical report, WordNet.Net, 2005.
- [24] Sebt A, Barfrous A A. A new word sense similarity measure in WordNet[C]. Proceedings of the International Multi-conference on Computer Science and Information Technology. Washinton D C: IEEE Computer Society, 2008:369-373.
- [25] Yang S C, Chen J J. Research on Sentence Similarity Computing in Chinese Automatic Answering[J]. Journal of The China Society For Scientific And Technical Information, 2008, 27(1):35-40.
- [26] 秦兵,刘挺,王洋等. 基于常问问题集的中文问答系统研究[J]. 哈尔滨工业大学学报, 2003, 35(10):1179-1182.
- [27] 张民,李生,赵铁军等. 一种汉语语句问相似度的度量算法及其实现[C]. 计算语言学进展与应用,北京:清华大学出版社,1995.
- [28] 钱丽萍,汪立东. 基于中心短语及权值的相似度计算[J]. 郑州大学学报:理学版, 2007, 39(2):149-152.
- [29] 董振东,董强. 知网简介[EB/OL]. <http://www.keenage.com>, 2001-02-11.

- [30] Liu Q L, Gu X F, Li J P. Researches of Chinese sentence similarity based on HowNet[C]. The 2010 International Conference on Apperceiving Computing and Intelligence Analysis, 2010:26-29.
- [31] 穗志方, 俞士汶. 基于骨架依存树的语句相似度计算模型[C]. 中文信息处理国际会议(ICCIP'98), 1998: 458-465.
- [32] 李彬, 刘挺, 秦兵等. 基于语义依存的汉语语句相似度计算[J]. 计算机应用研究, 2003, 20(12):15-17.
- [33] 哈尔滨工业大学研究的汉语语言处理平台, <http://ir.hit.edu.cn/>.
- [34] Zhang H P, Yu H K, Xiong D Y, et al. HHMM-based Chinese lexical analyzer ICTCLAS[C]. Proceedings of the second SIGHAN workshop on Chinese language processing-Volume Association for Computational Linguistics, 2003: 184-187.
- [35] 张亮, 冯冲, 陈肇雄等. 基于语句相似度计算的FAQ自动回复系统设计与实现[J]. 小型微型计算机系统, 2006, 27(4):720-723.
- [36] 关毅, 王晓龙. 基于统计的汉语词汇间语义相似度计算[C]. 全国第七届计算语言学联合学术会议. 哈尔滨, 2003:221-227.
- [37] 刘群, 李素建. 基于《知网》的词汇语义相似度计算[C]. 台北: 第三届汉语词汇语义学研讨会, 2002:8-15.
- [38] 梅家驹, 竺一鸣, 高蕴琪等. 同义词词林[M]. 上海: 上海辞书出版社, 1993:106-108.
- [39] 田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报(信息科学版), 2010, 28(6):602-608.
- [40] Lin D K. An information-theoretic definition of similarity[C]. Proceedings of the 15th international conference on Machine Learning, 1998, 1: 296-304.
- [41] Tesnière L, Fourquet J. Eléments de syntaxe structurale[M]. Paris: Klincksieck, 1959.
- [42] 郭艳华, 周昌乐. 一种汉语语句依存关系网协同生成方法研究[J]. 杭州电子工业学院学报, 2000, 20(4):24-32.
- [43] Leusch G, Ueffing N, Ney H. A novel string-to-string distance measure with applications to machine translation evaluation[C]. Proceedings of MT Summit IX, 2003: 33-40.
- [44] 车万翔, 刘挺, 秦兵等. 基于改进编辑距离的中文相似语句检索[J]. 高技术通讯, 2004, 14(7):15-19.
- [45] Wang R B, Wang X H, Chi Z R, et al. Chinese Sentence Similarity Measure Based on Words and Structure Information[C]. International Conference on Advanced Language Processing and Web Information Technology, 2008:27-31.

- [46] Nan X G. The Research of Sentence Similarity Computation based on Multi-Level Fusion[C]. The 7th International Conference on System of Systems Engineering, 2012:617-619.
- [47] 孙景广,蔡东风,吕德新等.基于知网的中文问题自动分类[J].中文信息学报,2007,21(1):90-95.
- [48] 文勛,张宇,刘挺等.基于句法结构分析的中文问题分类[J].中文信息学报,2006,20(2):33-39.
- [49] 李方涛,张显,孙建树等.一种新的层次化结构问题分类器[J].中文信息学报,2008,22(1):95-100.
- [50] 金博,史彦军,腾弘飞.基于语义理解的文本相似度算法[J].大连理工大学学报,2005,45(2):291-297.
- [51] Han K S, Chung H J, Kim S B, et al. Korea University Question Answering System at TREC 2004[C]. Proceedings of the 10th Text Retrieval Conference (TREC2004), 2004.
- [52] Hovy E, Hermjakob U, Lin C Y, et al. Using Knowledge to Facilitate Factoid Answer Pinpointing[C]. Proceedings of the COLING-2002 Conference, Taipei, Taiwan, 2002: 1-7..
- [53] Li X, Roth D, Small K. The Role of Semantic Information in Learning Question Classifiers[C]. Proceedings of the International Joint Conference on Natural Language Processing, Hainan Island, China, 2004: 451-458.
- [54] 宋艳雪,张绍武,林鸿飞.基于语境歧义词的语句情感倾向性分析[J].中文信息学报,2012,26(3):38-43.
- [55] 施寒潇,厉小军.主观性语句情感倾向性分析方法的研究[J].情报学报,2011,30(5):522-529.
- [56] 熊德兰,程菊明,田胜利.基于HowNet的语句褒贬倾向性研究[J].计算机工程与应用,2008,44(22):143-145.
- [57] 江敏,肖诗斌,王弘蔚等.一种改进的基于《知网》的词语语义相似度计算[J].中文信息学报,2008,22(5):84-89.
- [58] 李峰,李芳.中文词语语义相似度计算--基于《知网》2000[J].中文信息学报,2007,21(3):99-105.
- [59] 吴健,吴朝晖,李莹等.基于本体论和词汇语义相似度的web服务发现[J].计算机学报,2005,28(4).
- [60] Xia Y Q, Zhao T T, Yao J M, et al. Measuring Chinese-English Cross-Lingual Word Similarity with HowNet and Parallel Corpus[M]. In A. Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, volume 6609 of Lecture Notes in Computer Science, Springer Berlin, Heidelberg, 2011:

221-233.

- [61] Dai L L, Liu B, Xia Y N, Wu S K. Measuring Semantic Similarity between Words Using Hownet[C]. International Conference on Computer Science and Information Technology, 2008:601-605.
- [62] 陆善采. 实用汉语语义学[M]. 上海:学林出版社,1993.

攻读硕士学位期间参加研究的课题和发表的论文

发表的论文:

- [1] Qiang J P, Tian W D, Guo D, et al. Online pattern matching with wildcards[C]. In Proceedings of 2012 IEEE International Conference on Granular Computing (GrC), 2012, pp. 394-399. (EI 检索)
- [2] 田卫东, 强继朋. 基于问句类型的问句相似度计算. 计算机应用研究(已投稿).
- [3] Tian W D, Wu F, Qiang J P, et al. An ensemble learning algorithm based on generalized attribute value partitioning[C]. Seventh International Symposium on Multispectral Image Processing and Pattern Recognition (MIPPR2011). International Society for Optics and Photonics, 2011: 80041F-80041F-7. (EI 检索)


参与项目:

- [1] 参与国家自然科学基金课题“基于特征发现的数据流概念漂移问题研究”(项目编号 60975034)相关研究。
- [2] 参与安徽省烟草公司 2010 年度科技项目计划“安徽烟草商业信息系统风险评估及灾备方案研究”(项目编号 104-038163)相关研究。
- [3] 参与国家自然科学基金课题“带有通配符和长度约束的模式匹配和挖掘”(项目编号 60828005 61229301)相关研究。

特别声明

本学位论文是在我的导师指导下独立完成的。在研究生学习期间，我的导师要求我坚决抵制学术不端行为。在此，我郑重声明，本论文无任何学术不端行为，如果被发现有学术不端行为，一切责任完全由本人承担。

学位论文作者签名：



签字日期：2013年05月01日

厚德 笃学 崇实 尚新