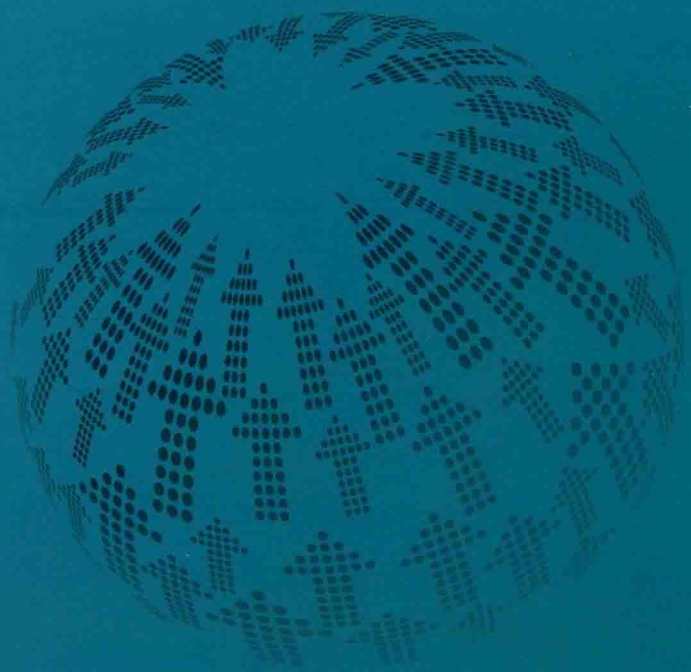
中文问答系统技术及应用

CHINESE QUESTION ANSWERING SYSTEM TECHNOLOGY AND APPLICATION

张 巍◎著



中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

中文问答系统技术及应用

张 巍 著

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

本书全面系统地介绍问答系统的基本技术及应用，不仅讨论受限领域的问答系统，而且讨论开放领域的问答系统。全书共13章：第1章为绪论，介绍问答系统的研究背景、意义、研究现状及分类；第2～12章分为两个部分，分别介绍受限域问答系统和开放域问答系统。在受限域问答系统中，讨论面向常问问题库及面向本体的问答策略，并讨论推理机制在问答系统中的应用。在开放域问答系统中，讨论中文问题分类技术及关键词扩展技术，然后讨论大规模问答对库的建立、答案推荐的技术等。各章对理论的叙述力求概念清晰、表达准确，突出理论联系实际，富有启发性，易于理解。

本书可以作为高等学校自然语言处理和计算语言学等专业本科生和研究生自然语言处理的教材，也可以作为从事自然语言处理相关领域的研究人员和技术人员的参考书。

图书在版编目（CIP）数据

中文问答系统技术及应用 / 张巍著. —北京：电子工业出版社，2016.5
ISBN 978-7-121-28270-6

I. ①中… II. ①张… III. ①自然语言处理－研究 IV. ①TP391

中国版本图书馆 CIP 数据核字（2016）第 045170 号

策划编辑：王晓庆
责任编辑：王晓庆　　文字编辑：王二华
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×1092　1/16　　印张：9.5　　字数：237 千字
版　　次：2016 年 5 月第 1 版
印　　次：2016 年 5 月第 1 次印刷
定　　价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010) 88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：(010) 88254113，wangxq@phei.com.cn。

前　言

互联网技术已经给人类社会带来翻天覆地的变化，人们已经习惯于从互联网上获取各类信息，这得益于搜索引擎技术的发展。然而，搜索引擎仍然有一些问题没有解决。首先，用户检索时，相关信息返回太多且不精确；其次，检索只能按关键字进行匹配，实际上并不能精确表达人们的检索需求。

由于以上问题的存在，问答系统应运而生。问答系统允许用户以自然语言方式进行提问，系统返回的是简洁的答案。问答系统主要分为受限域问答系统和开放域问答系统。因此本书针对这两种类型的问答系统进行讨论。第一部分介绍的是受限域问答系统，第二部分则介绍开放域问答系统。

第一部分共分 6 章，其中第 2 章介绍受限域问答系统研究的内容、本体的定义及本体语言 OWL，讨论本书用的“医院”领域本体的构建，并设计本体构建实验及本体推理实验，同时进行实验分析。第 3 章针对面向 FAQ 库的问答策略展开研究。首先建立问题库，然后设计基于常问问题集的问答策略模块，讨论基于统计和语义的方法及基于依存句法和改进编辑距离的方法计算问句相似度。通过对比实验，对这两种方法的优、缺点进行说明。

第 4 章深入研究面向本体知识库的问答策略。首先设计面向本体知识库的问答模块，提出基于语义块的问句浅层语义分析方法，建立语义块的定义规则及语义块的判定规则，并讨论语义块冲突时的处理方法。在此基础上，提出使用 SPARQL 技术在本体查询模块中抽取答案的方法。通过 FAQ 模块与本体查询模块的总体实验，验证多策略混合的问答系统方法的有效性。

第 5 章将 Jena 推理机引入问答系统中。首先介绍推理机的功能，并讨论本体推理机，然后论述 Jena 及组成结构、Jena2 的推理机制，最后提出 Jena 推理规则的构造及在问答系统中的应用，并设计 Jena 推理在问答系统中应用的实验。

第 6 章研究 SWRL 及 Jess 推理在问答系统中的应用。首先介绍 SWRL 架构及表示方式，然后研究基于本体的 SWRL 及 Jess 推理系统框架及实现框架，又讨论推理的过程，最后设计 SWRL 及 Jess 推理在问答系统中进行推理的实验，并进行结果分析。

作为受限域问答系统的应用，第 7 章研究城域医院问答系统的实现。论述系统的构建意义，提出系统的设计原则，建立系统的总体结构，最后介绍系统的实现并做出分析。

第二部分介绍的是开放域问答系统。其中第 8 章为概述，介绍开放域问答系统的特点，开放域问答系统的基本结构，最后给出本书第二部分的结构。第 9 章首先介绍重要的语义资源——《知网》，然后讨论问题的表示、问题预处理和关键词的提取。问题分类的关键有三点：一是选择合适的分类算法；二是提取合适的分类特征；三是要有好的训练语料。该章在介绍分类算法、分类体系的基础上，给出问题疑问词、核心词的主要义原、命名实体和名词单/复数等四种问句分类特征的提取方法，并给出问题分类的方法步骤和实验结果。实验结果表明，问句的疑问词和问句核心词的主要义原是分类的关键特征，特别是选择核心词的主要义原作为分类特征，有效地降低了问句特征向量的维数。

第 10 章首先介绍关键词扩展的意义，然后介绍信息检索中的同义词、《同义词词林》及其扩展版，接着详细描述利用《同义词词林》扩展，利用《知网》精简的关键词扩展的设计思路、具体算法及相关实验。第 11 章讨论答案源的获取方法。答案源的获取指的是通过信息检索和信息抽取得到包含答案信息的相关文档，为答案抽取提供数据源。对于开放域问答系统，信息检索阶段主要依赖现有的搜索引擎在互联网上进行信息检索。该章依据这一特点，介绍网页采集、网页去重、信息提取，最后介绍基于“百度知道”的问答对库的建立过程。

第 12 章首先介绍基于大规模问答对库的答案推荐的研究背景和研究现状，接着在大规模问答对库的基础上，介绍三种问题相似度计算的方法，最后通过实验方法和实验结果，证明基于关键词语义的句子相似度计算方法是最佳的。作为开放域问答系统的应用，第 13 章实现一个基于相似问题推荐的问答系统的原型系统，给出原型系统的结构，介绍原型系统的工作方式，提出通过问答对的满意度属性值实现问答对库自动优化的思路。

本书是在山西省科技攻关项目“基于 Web 的智能中文交互式甲流问答检索系统”（20110313019）资助下的一个成果，也包含了作者近年来的主要研究成果。本书的出版得到电子工业出版社的大力支持，在此致以真诚的感谢。另外尤其要感谢电子工业出版社的王晓庆编辑，她的建议大大提高了本书的质量。

由于作者的水平有限，书中难免存在错误和不妥，恳请读者批评指正。

作　者

2016 年 4 月

目　录

第二部分　开放域问答系统

第1章 绪 论

1.1 中文问答系统研究

1.1.1 研究背景

互联网是人类历史上最重要的发明，大大加速了人类文明的进程。同时，互联网也是一个巨大的信息资源库。从互联网上，人们可以得到自己需要的信息。但是，互联网也产生了信息过载的问题，就是互联网信息资源极大丰富，而其中有用的信息不多，这样就导致了现代信息检索技术的发展。

当今时代是一个信息的时代，人们可以通过互联网即时方便地浏览自己所关注的信息。特别是搜索引擎的出现，使人们能够更好地利用互联网的资源来查找信息，人们经常使用百度、Google[1]这些搜索引擎进行信息的检索。但是，由于搜索引擎返回给用户的往往是一大堆相关的文档，并不是精确的信息，所以用户仍然需要花费大量的时间到这些相关的文档中去查找精确的信息。这些缺点反映了使用关键字检索是有缺陷的[2]：首先，它不能准确地表达用户检索的真正需求，如果用户采用自然语言方式进行提问，则可以满足用户的真正需要；其次，采用关键字检索得到的答案，会包含很多语义不相关的信息。这两个问题，促使很多研究人员对问答系统展开了研究[3]，出现了大量的基于 Web 的问答系统，如 START[4]、AnswerBus[5]、Brainboost[6]等。但这些研究并没有解决上述的第二个问题，要想解决答案包含不相关信息的问题，必须寻找一种新的知识源，这种知识源可以包含语义。

当前，随着互联网技术[7~8]的不断深入，特别是语义 Web[9~10]的提出，采用本体[11~15]技术作为知识源进行检索，可以解决答案包含语义无关信息的问题。因为本体可以包含语义，而且可以进行推理。所以，以本体为知识源的问答技术将开辟信息检索技术的新方向。

本体（Ontology）这一概念来源于哲学领域，用于表示客观存在。20 世纪 80 年代以来本体被引入信息领域并取得了长足的发展，但在信息领域和哲学领域中本体扮演着完全不同的角色。信息领域的本体被重新定义为“共享概念模型的形式化规范说明”，用于提供对该领域知识的共同理解，确定该领域内共同认可的词汇。要表示本体知识，需要有本体语言。本体语言经历了十多年的发展，曾经涌现的语言有 SHOE[16]、XOL[17]、RDF[18]、RDFS[19]、OIL[20]、DAML+OIL[21]等，都以 XML 作为基础。2004 年，W3C 总结了历史上出现过的本体语言并吸取了它们的长处，将 OWL[22~23]（Web Ontology Language，即网络本体语言）设定为工业标准。本体语言标准化，大大促进了语义 Web 技术的发展，出现了很多 OWL 知识库，如啤酒本体[24]等。正是由于本体可以包含语义，而且可以进行推理，因此本书以本体作为知识源，进行问答系统的研究。和本书相关的本体的研究热点还有 OWL 本体知识查询[25]。

本体知识查询如图 1-1 所示。

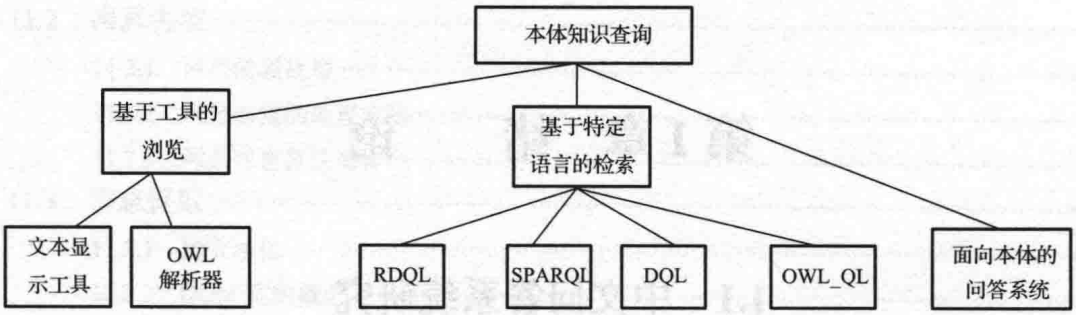


图 1-1 本体知识查询

由图 1-1 可见，可以将本体知识查询分为基于工具的浏览、基于特定查询语言的检索和问答式检索这三种查询方式。基于工具的浏览包括文本显示工具浏览和 OWL 解析器浏览。文本显示工具（如浏览器、记事本等）能将 OWL 知识显示成普通文本。OWL 解析器提供了图形化的显示方式，帮助用户直观地浏览所需信息。如 Protégé[26][27]、SWOOP[28]是其中常用的工具。

不同的查询语言可以查询不同本体语言表示的本体知识，同一种查询语言有时也可以查询不同格式的本体知识库。现阶段，能够检索 OWL 知识的查询语言主要有 RDQL[29]、SPARQL[30]、DQL[31]、OWL-QL[32]四种。RDQL 类似于 SQL，是专门用于查询 RDF 语言而设计的。SPARQL 已经于 2008 年 1 月 15 日成为 W3C 的推荐标准，也可以用于 RDF 语言的查询。DQL 是专门用于对 DAML+OIL 语言进行查询的，而 OWL-QL 是用于对 OWL 知识进行查询的。

问答系统可以分为面向本体的问答系统、面向 Web 的问答系统、面向数据库的问答系统。它们均使用自然语言进行检索，因此更加符合人类的行为方式。面向本体的问答系统对本体知识库进行检索，可以实现真正意义上的人机交互。随着语义 Web 技术的发展，网络上出现了越来越多的 OWL 格式的本体知识，这样就给智能 Agent 的共享和交换创造了条件，而这是机-机之间交流的基础。语义 Web 技术的另一个目标是实现人-机交流。实现人类和 Agent 之间的动态交互是研究人-机交流的一个重要目标。

因此，研究与 OWL 相关的本体知识具有极其重要的意义，既可以利用现有的 OWL 本体知识，又可以扩展其表达和推理能力，从而满足新一代 Web 的需要。

1.1.2 研究意义

2016 年 1 月 22 日，中国互联网络信息中心（CNNIC）发布第 37 次《中国互联网络发展状况统计报告》[33]（以下简称为《报告》）。《报告》显示，截至 2015 年 12 月，中国网民规模达 6.88 亿，互联网普及率达到 50.3%，半数中国人已接入互联网。同时，移动互联网塑造了全新的社会生活形态，“互联网+”行动计划不断助力企业发展，互联网对于整体社会的影响已进入新的阶段。

随着互联网技术的急速发展，人们遇到问题，总是喜欢用百度和谷歌等搜索引擎来寻找答案，只需要输入一些关键字，搜索引擎就会返回很多包含关键字的网页，但是人们在使用此种方法搜索时发现许多不满意的地方。

（1）返回信息太多，搜索引擎每次都返回许多和输入关键词有关的网页。其中出现许多重复的内容，这样必然会增加人们得到自己所需答案的时间。

（2）产生大量无关内容。有些网页仅仅包含了用户输入的一个或几个关键词，返回的网页常常和用户所需要的答案没有关系，用户只得改变关键词再进行检索。

根据IDC（Internet Data Center，即互联网数据中心）的调查，迄今为止，搜索仍然是上网冲浪者在网上最频繁的活动。另根据RoperStarch的调查，36%的互联网用户一个星期花了超过2个小时的时间在网上搜索；71%的用户在使用搜索引擎的时候遇到过麻烦；平均搜索12分钟以后发现搜索受挫；搜索受挫的原因46%是因为链接错误；86%的人认为搜索引擎应该可以更加有效。搜索结果不能使人满意的原因主要如下。

（1）目前用户输入的只是简单的关键词组合，而关键词组合常常不能表达用户的真正意图，返回的答案也就不能使用户满意。

（2）目前的搜索以关键词为基础，没有结合自然语言处理对语义进行研究，还是停留在词语的表层。这样要正确理解用户输入的查找意图是困难的，检索效果就不可能有大的提高。

（3）答案仅仅是一般的网页，而没有进行加工。这样就造成答案中有许多冗余及无用信息。

为了克服传统搜索引擎存在的缺陷，很多大学和一些大公司都进行了结合自然语言处理技术和语义搜索技术的研究工作。现阶段，美国AskJeeves公司的检索系统是比较成功的一例，人们可以用自然语言句子对此系统进行提问，如果系统能判断出用户意图就检索答案并返回给人们。遗憾的是，该系统返回的还是相关的网页，而不是直接的回答，因此不能算是真正意义上的自动问答系统。

人们希望可以有一个用自然语言句子提问，能够直接返回答案的检索系统出现，而这正是自动问答系统。这就促进了自动问答系统进一步的研究。

1.2 问答系统国内外研究现状

最早的问答系统，实际上是阅读理解[34]，就是对一篇文章进行问答。SAM[35]（Script Applier Mechanism）就是这样的系统，它可以读很短的故事，并回答故事中的一些问题。系统中有多种范本，可以表示一定情境下可能发生的事情。后来出现的Ms.Malaprop[36]也是使用框架来进行推理的。PAM[37]也是类似的系统，可以对发生的事件进行解释。这些系统都是对文本进行理解，其原理是建立特定框架并组建知识库，以此来回答问题，但回答问题的范围极其有限。

随着互联网的出现，FAQ Finder[38]、AnswerBus[39]及MULDER[40]扩大了回答问题的范围，从互联网上来抽取答案，因此可以回答大量的问题。

国际上最具权威的文本检索会议TREC于1999年首次设置了有关问答系统的竞赛[41~44]。TREC-8冠军自动问答系统LASSO[45]使用了问题分类技术，并对每类问题设置相应的处理策略，这样大大提高了问题分析的准确度。其缺陷是问题分类粒度太粗，因此对很多问题不能准确得到答案。

INSIGHT[46]在2001年的TREC-10的自动问答系统竞赛中，在分析问题时只采用一种知识：若干浅显的模板，便获得了冠军。此系统使用字符串构建模板，并据此抽取答案，大大提高了系统回答的正确率，但是此系统通用性较差，原因是它采用的模板粒度划分过细。为了解决此问题，研究人员Zhang和Lee[47]采用了一种模板学习算法，目标是通过增加模板的

数目提高模板库的通用性，然而依然没有调节所获得的问题模板的粒度。同时，由于普林斯顿大学创建的英语词典《Wordnet》能够包含英文的同义词、反义词、上位及下位词，所以《Wordnet》[48]被用在自动问答系统 LCC[49]中，这样就可以把问题中重要词汇的隐藏信息，如同义词等挖掘出来，并使用这些信息抽取出那些潜在的答案。

AQUA[50]是一个可以检索学术领域问题的问答系统，综合应用了本体技术、自然语言处理技术，在检索答案时，使用自定义的查询逻辑语言进行推理，从而自动地得到答案。

与国外的问答系统相比，中文问答系统不管是研究还是技术都相对滞后，目前最好的中文问答系统的正确率为 55%[51]。因为中文处理起来比英文更难，所以中文问答系统的研究更难。但因为其具有较高的研究价值，吸引了大批学者进行研究。在以往的 TREC QA Track 中，复旦大学[52]、中科院计算所[53]都取得了良好的成绩[54]。近年来，取得成绩较大的研究机构是哈尔滨工业大学[55]，他们在中文依存句法分析、中文消歧处理等方面做了很多工作。工业界则采用网络社区的方式，通过人们的提问和回答，来解决日常生活中的常见问题，这方面的代表有百度知道[56]、天涯问答[57]、奇虎网[58]、搜搜问问[59]、爱问知识人[60]等。当用户向这类系统提问后，如果答案在历史数据库中，则系统会立即返回答案。否则，若干相似的问题及答案会返回给提问者。

总而言之，中文问答系统和国外的问答系统相比较，差距很大[61]，这其中有多方面的原因，但主要原因有两点：一个原因是中文信息处理技术还不成熟；另一个原因是到目前还没有建立起一个权威的中文问答系统评测平台[62]。

1.3 问答系统的分类

根据面向的领域的差异、信息存储数据格式的不同、系统处理的答案来源的不同、问答系统实现技术的特色差异等，问答系统可以有多种分类方法。

按照面向领域的差异，问答系统可以划分为特定领域（Special Domain）问答系统、开放领域（Open Domain）问答系统、常见问题集（Frequently Asked Questions, FAQ）问答系统[63]，如图 1-2 所示。

图 1-2 问答系统按照问题领域分类

顾名思义，特定领域问答系统是指在旅游、法律、税务、餐饮、教育、房地产、银行等特定领域中，问答系统受各专业领域知识限制，答案具有专业特定的具体知识和特殊要求，因为涉及范围较小，现有的特定领域问答系统研究已经具备较好的处理效果。开放领域问答系统指的是处理领域不受限制的问答系统，以互联网为基础源的问答系统是非常典型且常用的开放域问答系统，但因为互联网本身的特性，涉及面广，处理的对象复杂多变，处理过程比较繁杂，因此目前针对开放领域问答系统的研究还没有达到理想的状况[64~66]。

目前，互联网已经逐渐成为人们日常生活中查询信息时不可或缺的一个组成部分，开放领域问答系统，特别是对针对互联网的开放领域问答系统的研究和探索逐渐引起越来越多人的重视。而常见问题集的构成实际上是一个典型的特定领域问答系统的特例，常见问题集通常将某个专业网站或某类产品使用中经常碰到的问题及答案汇集到一起，以知识库的形式形成问答内容集合，如果用户提出的问题正好与常见问题集合中的问题相同或类似时，系统能够实现快速查询并自动给出相应答案[67~70]。

按照信息存储数据格式的不同，问答系统可以划分为处理自由文本的问答系统、处理结构化数据的问答系统和处理半结构化数据的问答系统[71~73]，如图 1-3 所示。

图 1-3 问答系统按照处理的数据格式分类

所谓自由文本指的就是我们常规处理的文本数据，数据中不包含任何格式信息，对文本的自动处理只能严格从词法、语法、语义分析的角度进行，目前针对大规模文本检索已有一些成熟的算法，但将这些算法直接应用于查询问题比较简洁的问答系统还有待进一步讨论。结构化数据是指行数据，即存储在数据库里，可以用二维表结构来表达实现的数据。目前的计算机系统通常采用关系数据库来存储和管理这种类型的数据，而对于数据库中存放的结构化数据的关键信息抽取或者其中所包含的命名实体的识别，则通过目前成熟的关系数据库理论和工具可以有很高的识别精度，因此处理结构化数据的问答系统的整体性能都相对比较理想，但需要事先确定查询的关键词。

除了可以用关系数据库来处理的结构化数据外，现今社会生活中，还有一些无法用统一结构表示的信息，如文本、图像、声音、网页等，称为非结构化数据，结构化数据其实是非结构化数据的特例。而半结构化数据是指结构变化很大、不能够简单地通过结构化方式的表格来处理的一类数据，如网页，包含一定的格式，但格式变化比较大，通常采用 XML/HTML 等标记语言来表述，将不同类别的信息保存在不同的节点中，能够灵活地进行扩展以适应半结构化数据存储和管理的特点。目前开放领域搜索引擎的研究正是针对这类半结构化数据来处理的，但搜索返回的信息量太大，需要用户做二次过滤。

按照系统能够提供的答案来源的不同，问答系统可以划分为单文本问答系统、固定数据库问答系统、局域网问答系统和互联网问答系统[74~76]，如图 1-4 所示。

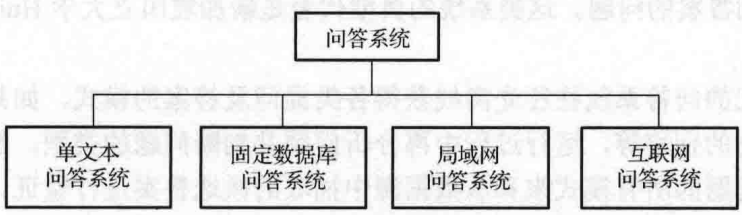


图 1-4 问答系统按照答案来源分类

所谓单文本问答系统，也称为阅读理解式的问答系统，采用类似于一般人完成阅读理解的过程，问答系统在“阅读”完一篇指定的文本以后，自动从文本中查找答案，根据问答系统在阅读过程中对文本的“理解”给出问题的答案。

固定数据库问答系统的基础是大量已经建好的、有意义的真实文本语料库或关系数据库，如大家所熟知的 TREC QA Track 评测平台。固定数据库问答系统的优势是已建好的数据库平台上通常都有标准化的、性能优良的算法评测工具，适用于针对不同类型的问答技术进行比较研究。目前这类问答系统平台主要用于各种问答技术效果的测评，或者各类改良算法的对比研究和测评；缺陷是数据库内容已经事先确定，不能根据用户需求变化，因此只能将用户问题对应到数据库中已经存在的相似问题及其答案上去，无法包含所有类型的用户提问。

局域网问答系统是指在局域网范围内所包含的数据资源上实现的问答系统，通常受局域网范围的数据量限制，问答系统处理的数据量较小，但速度和可靠性较高。而互联网问答系统指在因特网上搜寻答案，也即目前的搜索引擎所做的工作。众所周知，互联网上包含的信息数量、数据格式都非常巨大，从这类问答系统返回的结果中如何筛选出用户需要的精简答案已经成为一个新的研究课题。

按系统采用的实现技术的不同，问答系统可划分为基于信息检索的问答系统、基于模式匹配的问答系统和基于自然语言处理的问答系统三种[77~79]，如图 1-5 所示。

图 1-5 问答系统按照实现技术不同的分类

基于信息检索的问答系统，在查询基础上通过信息抽取算法获取问题答案。这种方法实施的关键技术之一是，候选答案排序方式选择的不同，会影响问答系统给出的最终答案。排序方式常常是问答系统的提问处理模块所生成的查询关键词，参考文献[1]根据既有的不同类别关键词对排序的贡献的不同，把查询关键词分为普通关键词、扩展关键词、基本名词短语、引用词和其他关键词，并给出这些常用关键词的加权方法。采用信息检索技术来实现问答系统比较简单易行，但这种方法查询技术的主要依据是关键词，而关键词之间是离散的，几乎没有联系，不涉及这些关键词在问句中的相互关系，因此这种方法给出的答案缺乏问句中所包含的句法关系和语义关系，更无法回答隐藏在答案源中、需要进一步推理才能找到答案的问题。这类系统的典型代表是新加坡国立大学 Hui Yang 等人研发的系统[80]。

基于模式匹配的问答系统往往先离线获得各类提问及答案的模式，如某人的出生日期、某人的原名、某物的别称等，运行过程中再分析问题并判断问题的类型，然后用离线阶段准备好的关于这类问题的所有模式来和从数据源中抽取的候选答案进行验证。这种类型的问答系统虽然对一些固定模式的问题（如定义，地点、出生日期提问等）具有较好的效果，但缺陷是离线模板不可能涵盖提问中可能出现的所有问题，也不可能处理答案源中没有直接包含

的答案。俄罗斯 InsightSoft-M 公司 Martin Soubbotin 等人研发的系统[81, 82]是这种技术所实现的问答系统的典型代表。

基于自然语言处理的问答系统显然是上述各种分类中最理想的自动问答系统，具有自然语言处理功能的问答系统既可以按照人的思维分析提问，又可以按照人的思维对答案源进行分析、整理，甚至推理。但这种技术目前还不是很成熟，除了一些简单的规则化文本处理技术（如词汇识别、分词、词性标注等）外，语义的处理等深层技术还只是在研究阶段中，没有很成熟的应用。因此，基于自然语言处理的问答系统只能作为前两种方法的有效补充。这类系统的典型代表是美国 Language Computer Corporation 公司的 Sanda Harabagiu 等人研发的系统[83]。