

# 基于动态知识库的问答系统研究\*

王树西 刘群 白硕 王斌 程学旗 姜吉发

中国科学院计算技术研究所 软件研究室 北京 100080

E-mail: wangshuxi@software.ict.ac.cn

**摘 要:**问答系统有着较长的历史。本文在综述现有问答系统的基础上,提出“动态知识库”的概念,并基于此,搭建了“亲属关系问答系统”,在知识获取、问答系统发展趋势等方面,进行了一定的探索。

**关键词:** 问答系统, 动态知识库, 知识获取

## The Research On QA System Based on Dynamic KB\*

Wang Shuxi Liu Qun Bai Shuo Wang Bin Cheng Xueqi Jiang Jifa

Software Division, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080

E-mail: wangshuxi@software.ict.ac.cn

**Abstract:** Question Answering System(QA) has a long history. In this paper, we first gave a survey on the current research of QA, then, we put forward the concept of “Dynamic Knowledge-Base”. Based on which, we put up “The Question Answering System of Kinship”, which is an explore of the Knowledge-Acquisition and the developing trend of QA.

**Keywords:** Question Answering System(QA), Dynamic Knowledge-Base, Knowledge Acquisition

### 1 引言

问答系统的设计目标是:以自然语言交互的方式,准确回答用户的问题。问答系统有着较长的历史。1950年,图灵(A. M. Turing)发表了里程碑式的论文:《Computing Machinery and Intelligence》,在文中,图灵第一次提出“机器智能”的概念,并提出判断“机器智能”的实验方法——“图灵测试”。本文在综述现有问答系统的基础上,提出“动态知识库”的概念,并基于此,搭建了“亲属关系问答系统”,在知识获取、问答系统发展趋势等方面,进行了一定的探索。

---

\*本文有关研究得到国家重点基础研究项目(G1998030507-4和G1998030510)的资助。王树西,男,1976年生,博士研究生,主要研究方向是人工智能、问答系统、自然语言处理。刘群,男,副研究员,主要研究方向是自然语言处理。白硕,男,博士,研究员,博士研究生导师,主要研究方向是人工智能、信息安全。王斌,男,博士,副研究员,主要研究方向是信息检索和内容安全。程学旗,男,副研究员,主要研究方向是信息安全、信息检索。姜吉发,男,博士研究生,主要研究方向是信息抽取。

## 2 问答系统综述

当前研究的问答系统主要包括：聊天机器人(Chat Bot)、自然语言界面的专家系统、基于知识库的问答系统、基于传统 IR(+IE)的问答系统，等等。

### 2.1 聊天机器人(Chat Bot)

这里所说的聊天机器人,是指一个计算机系统,它通过自然语言的方式与人交互,对用户的提问,给出尽可能合理的回答。

1956 年 Joseph Weizenbaum 实现的 ELIZA, 是第一个聊天机器人。ELIZA 的原理是, 根据用户提问中的关键词, 检索数据文件, 找到与之匹配的答案。Eliza 算法简洁, 但由于其知识库过于短小(只有几百条知识模板), 所以, 无法满足用户知识查询的需求。

1991年, “Loebner奖”设立, 奖励首次通过图灵测试的人。此奖项设立以来, 许多著名的系统参加了比赛, ALICE就是其中一例。但迄今为止, 没有任何一个系统通过“图灵测试”。

对 ALICE 等参赛系统的测试结果表明, 这类系统能够比较准确的理解用户意图, 对常识性问题, 往往给出合理的回答, 但由于其知识库规模有限, 所以面对专业性的问题, 显得力不从心。所以, 系统仅仅具备常识性知识问答能力, 而不具备专业知识问答能力。

### 2.2 自然语言界面的专家系统

专家系统(Expert System, ES), 是以计算机为工具, 利用专家知识以及知识推理等技术, 理解与求解问题的知识系统, 是人工智能应用研究的主要领域之一。1968 年, 费根鲍姆等人研制成功第一个专家系统——DENDRAL。

专家系统的人机接口模块, 将用户的输入, 转换为系统可接受的内部形式; 将系统的输出, 转换为人可理解的外部形式。人机接口的方式有多种, 如果采用自然语言的人机交互方式, 则系统整体表现为一个问答系统。

专家系统的优点是: 技术较成熟、开发工具较多、易于开发, 并且答案比较准确。但是, 大部分专家系统推理方法单调、固定, 只能做演绎推理, 不具备常识推理能力。并且, 专家系统的知识库严重不足, 自动获取知识能力差, 存在知识获取的瓶颈问题。所以, 目前专家系统的适用范围非常狭窄, 一旦超出这个范围, 系统性能很快下降到零。

### 2.3 基于知识库的问答系统

基于知识库的问答系统, 包括 CYC、NKI(US)、NKI(China)等。这类系统的优点是, 回答准确, 可以进行一定的推理计算; 缺点是, 需要建立大规模知识库, 消耗大量的人力物力。下面对 NKI(China), 做一个简要介绍。

根据 NKI 建设者自己的定义, NKI 是一个庞大的、可共享的知识群体。它不仅集成了各个学科的公共知识, 而且还融入了各学科专家的个人知识。

NKI 问答系统(<http://www.nki.net.cn>), 是基于 NKI 海量知识库的重要应用, 可以对国家地理知识库、城市天气预报知识库、人物知识库等 23 个知识库的知识进行查询。用户可以通过自由的自然语言的提问方式获取所需要的知识, 输入形式可以多样化。

## 2.4 基于传统IR(+IE)的问答系统

基于传统 IR(+IE)的问答系统，与目前的搜索引擎（如 GOOGLE）有所不同。搜索引擎通过用户输入的关键字（词），检索出相关网页；而基于传统 IR(+IE)的问答系统，则允许用户输入完整的句子。这类问答系统，又可分为两类：

第一类，将多个 Web 页面或者链接提交给用户，让用户自己寻找答案。典型的系统有 AskJeeves([www.ask.com](http://www.ask.com))、Encarta([encarta.msn.com/](http://encarta.msn.com/))等。这类系统相对简单，但仅仅以页面和链接作为用户问题的答案显然不够准确。严格地说，这不能算是一个完全意义上的问答系统。

第二类，从大量网页中检索到答案，然后以自然语言的方式提交给用户。比较典型的，是 TREC 比赛中的 QA (Question Answering) Track。自从 1999 年举办第一届 TREC QA Track 以来，许多系统参加了比赛。这些系统采用的方法不尽相同，但其工作流程大都分为三个阶段：（1）处理用户的查询；（2）检索相关文本；（3）抽取答案。这类系统技术成熟，易于开发；但是答案准确性一般，基本上是一个检索过程，对文本理解和推理涉及较少。

## 3 动态知识库

通过分析现有各类问答系统，可以看出：为了得到令人满意的答案，系统必须具备完善的推理(检索)机制以及尽可能完备的知识库。现有问答系统的知识库，都是作为一个独立的模块，在系统运行前预编译好，在系统运行过程中，知识库不再变更。这种做法的好处是，知识库与推理机分离，易于维护；缺点是，在系统运行过程中，由于知识库固定，不具备实时交互能力，不能接受新的知识，所以无法进行动态知识处理，具体表现为系统灵活性差。

基于上述分析，我们提出“动态知识库”的概念。所谓的“动态知识库”，是指在系统运行过程中，知识库不是固定的，而是可以实时的接受新知识，进行扩展与更新。

这种做法的好处是，系统的知识库变得非常灵活，可以在系统运行过程中，实时的进行扩展。并且，系统推理过程中所得到的中间结果，也可以实时的插入知识库，作为进一步推理所需要的知识，直到得出最终结果。

这种做法的另外一个特点是，将无结构的原始文本作为知识来源，不需要人工形式化的过程，知识转换工作由系统相应的模块完成，大大节省了人力物力。

基于“动态知识库”的问答系统，允许用户在线地、交互式地以文本方式输入知识。换言之，文本知识可以动态的插入知识库，即插即用，不需要“消化”或者人工加工。例如，系统知识库中原本是空的，在系统运行过程中，通过人机界面，用户插入两条文本表示的新知识：

“李纨是贾宝玉的嫂子”、“贾宝玉是王夫人的儿子”，然后紧接着问：“李纨是王夫人的什么人？”。系统实时接受这两条新知识，进行知识库扩展与更新，然后通过推理机制，经过推理，得到准确的答案：“李纨是王夫人的儿媳妇”。

## 4 基于动态知识库的问答系统——“亲属关系问答系统”

基于“动态知识库”的概念，我们搭建了“亲属关系问答系统”。在系统运行过程中，用户可以实时的插入新知识，进行知识库的扩充与更新。在这个系统中，提交（获取）知识、处理知识和查询知识都是在知识的自然语言表示下进行的。

## 4.1 知识转换

在系统运行过程中,用户插入的知识,是以文本形式表示的,而知识库所存储的,则是形式化和结构化的知识。所以必须进行知识转换工作,将用户插入的文本形式的知识,转换成系统可以接受的内部表示形式。

我们采用模板匹配的方法,进行知识转换。

首先建立模板库。我们采用手工建立模板、并给每个模板分配唯一 id 号的方法,建立模板库。模板库的存储结构如下所示:“X 是 Y 的父亲##id4(X,Y)”。其中,“X 是 Y 的父亲”是一条知识模板,“id4”是分配给这条模板的唯一 id 号,“id4(X,Y)”是这条模板的内部表示形式。客观问题的复杂性,要求模板库中存放大量的模板,模板越多,系统的功能就越强。

现有的模板匹配技术,一种是关键字匹配,另一种是句法匹配。本文提出的模板匹配方法,不对模板作句法分析,所以不同于句法匹配;并且,由于考虑模板中的非关键词成分,所以也不同于关键字匹配。具体算法如下(String 表示字符串,Pattern 表示模板):

- (1) 确定 Pattern 中所有变量字符、非变量字符串及其在模板中的位置。转(2);
- (2) 将 Pattern 中所有非变量字符串,顺序的匹配 String。  
如果匹配失败,错误返回;否则转(3);
- (3) 通过 Pattern 中非变量字符串,界定其每个变量代表的字符串。转(4);
- (4) 如果 Pattern 中,变量代表的字符串为空字符串,那么错误返回;否则转(5);
- (5) 如果多个模板同时匹配到 String,取变量个数最多的模板作为最佳模板;
- (6) 正确返回。

例如,系统运行过程中,用户插入下面一条知识:“贾政是贾宝玉的父亲”。这是一条文本形式的知识,必须经过知识转换,才能够被系统所接受。下面是知识转换过程:通过模板匹配,得到与之匹配的模板:“X 是 Y 的父亲”,其中,“X”代表“贾政”,“Y”代表“贾宝玉”。当前模板的内部表示形式为“id4(X,Y)”,变量还原,得到:“id4(‘贾政’,‘贾宝玉’)”,这就是一条经过转换的、系统可以接受的知识,可以加入动态知识库中了。

## 4.2 系统的动态知识库

系统运行之前,知识库是空的。系统运行过程中,可以实时的接受用户输入的新知识;系统推理所得到的结果,也可以实时的插入知识库。

例如,系统运行过程中,通过人机界面,用户输入文本表示的知识:“贾政是贾宝玉的父亲”、“王夫人是贾宝玉的母亲”。通过知识转换,上述两条知识形式化为:“id4(‘贾政’,‘贾宝玉’)”、“id2(‘王夫人’,‘贾宝玉’)”,并插入知识库。系统推理过程中,得到推理结果:“id8(‘贾政’,‘王夫人’)”(贾政是王夫人的丈夫)、“id9(‘王夫人’,‘贾政’)”(王夫人是贾政的妻子)。这两条推理结果,也被插入知识库中。至此,系统的知识库,已经有了四条知识:“id4(‘贾政’,‘贾宝玉’)”、“id2(‘王夫人’,‘贾宝玉’)”、“id8(‘贾政’,‘王夫人’)”、“id9(‘王夫人’,‘贾政’)”。

这样,系统的知识库变得非常灵活,能够实时的扩展与更新,称动态知识库。

## 4.3 亲属词、亲属关系推理规则库

汉语是亲属词丰富程度非常高的语言。亲属词本质上表示的是关系,复杂的关系可以还原为基本的关系和属性。最基本的关系是:亲子关系、夫妻关系、长幼关系;最基本的属性是性别属性。汉语的亲属词均为参考人 X 的函数。



不同的亲属词，可能表示相同的亲属关系。例如，“妈妈”和“母亲”这两个不同的亲属词，表示同一种亲属关系——“母子关系”。为了处理方便，我们把表示相同亲属关系的不同亲属词，归结为同一个亲属词。

亲属关系推理规则，是人工定义的，推理规则的集合，形成推理规则库。下面就是一例推理规则：“Y 是 Z 的父亲，Z 是 X 的父亲  $\therefore$  Y 是 X 的祖父”。

上例中的推理规则，是文本形式的，必须转换为形式化和结构化的内部表示形式，才能被系统所接受。我们采用下述步骤，转换推理规则。

首先，分离出推理规则中的模板；

其次，将模板转换为系统的内部表示形式；

最后，将推理规则中的模板，替换为转换之后的模板。

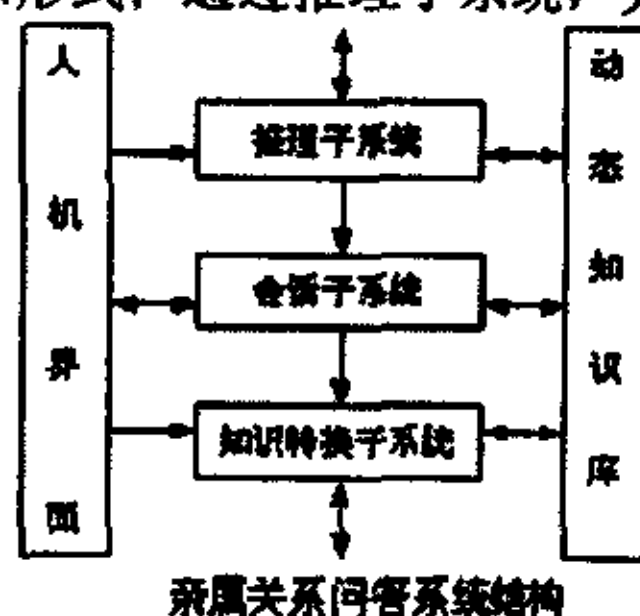
仍以上述推理规则为例。首先，分离出推理规则中的三个模板：“Y 是 Z 的父亲”、“Z 是 X 的父亲”、“Y 是 X 的祖父”。其次，将这三个模板，分别转换为系统的内部表示形式：

“id4(Y,Z)”、“id4(Z,X)”、“id17(Y,X)”。最后，将推理规则中的模板，替换为转换之后的模板，得到：“id4(Y,Z), id4(Z,X)  $\therefore$  id17(Y,X)”。这是一条系统可以接受的、形式化的推理规则。

推理是一个思维过程，即问题求解的过程。在本系统中，模板库、规则库和动态知识库，结合 prolog，作为推理机，完成推理机制。

#### 4.4 系统的结构与工作流程

系统的结构如下图所示。它包括会话子系统、知识转换子系统以及推理子系统。其中，系统通过会话子系统，和用户交互；通过知识转换子系统，将用户输入中的文本知识形式化，转换为系统可以接受的内部表示形式；通过推理子系统，完成推理机制，得到问题答案。



系统工作流程：

- (1) 通过会话子系统，接受用户的输入。转 (2)；
- (2) 通过知识转换子系统，处理用户的输入：将其中的文本知识形式化，并插入动态知识库；得到用户查询的问题，并将其形式化。转 (3)；
- (3) 通过推理子系统，结合动态知识库，对用户查询的问题进行推理。转 (4)；
- (4) 如果推理成功，那么通过知识转换子系统，将形式化的推理结果，转换为自然语言的形式；否则，也以自然语言的形式，说明推理失败，无法得到查询的答案。转 (5)；
- (5) 通过会话子系统，将推理结果提交给用户。

#### 4.5 系统测试

测试环境：CPU 为 PIII667、内存为 192MB、操作系统为 Windows 2000 Professional。  
下例是用户和系统的对话。【问】表示用户的提问，【答】表示系统的回答。

**【问】：**贾宝玉是贾政的儿子，贾政是林黛玉的舅父，贾政是薛宝钗的姨父，薛宝钗是女性，贾宝玉是薛宝钗的丈夫，薛宝钗比贾宝玉大，贾宝玉比林黛玉大，林黛玉是女性，贾政、贾宝玉、林黛玉、薛宝钗之间存在哪些关系？

**【答】：**贾政是贾宝玉的父亲，贾宝玉是贾政的儿子，贾政是林黛玉的舅父，林黛玉是贾政的外甥女，贾政是薛宝钗的姨父，贾政是薛宝钗的公公，薛宝钗是贾政的儿媳，薛宝钗是贾政的外甥女，贾宝玉是林黛玉的表哥，林黛玉是贾宝玉的表妹，贾宝玉是薛宝钗的丈夫，贾宝玉是薛宝钗的表弟，薛宝钗是贾宝玉的妻子，薛宝钗是贾宝玉的表姐，林黛玉是薛宝钗的远房表妹，薛宝钗是林黛玉的远房表姐。

本例中，系统通过会话子系统，获取用户的输入；通过知识转换子系统，将其中文本表示的知识：“贾宝玉是贾政的儿子”等，实时转换为形式化的、推理机可以接受的内部表示形式：“id5(‘贾宝玉’,‘贾政’)”等，并插入知识库；通过推理子系统，结合动态知识库，进行推理，得到推理结果：“id4(‘贾政’,‘贾宝玉’)”等；通过知识转换子系统，将推理结果转换为自然语言表示的形式：“贾政是贾宝玉的父亲”等，并通过会话子系统，提交给用户。对系统的反复测试表明，系统不但具有灵活性、易维护性的特点，而且具有令人满意的鲁棒性。

## 5 结束语

本文的主要贡献在于：在综述现有问答系统的基础上，提出“动态知识库”的概念，并基于此，搭建了“亲属关系问答系统”，在知识获取、问答系统发展趋势等方面，进行了一定的探索。但是，本文的工作还仅仅是探索性的，尚需进一步的深入。今后的工作，重点将在如下方面：(1) 自动模板抽取；(2) 模板之间的匹配、合一与推理。在这方面，我们已经取得了阶段性的成果。

## 参 考 文 献

- [1] 陆汝铃 《人工智能》 科学出版社 2000 年
- [2] 陆钟万 《面向计算机科学的数理逻辑》 科学出版社 1998 年
- [3] 蔡自兴、徐光佑 《人工智能及其应用》 清华大学出版社 1996 年
- [4] 白硕 《计算语言学教程》(电子版)
- [5] 史忠植 《高级人工智能》 科学出版社 1998 年
- [6] 白硕 《Reasoning Without Deep Structure 》(PowerPoint)
- [7] 王永庆 《人工智能原理与方法》 西安交通大学出版社 1999 年
- [8] A.M.Turing “COMPUTING MACHINERY AND INTELLIGENCE”
- [9] Deepak Ravichandran “Learning Surface Text Patterns for a Question Answering System”
- [10] Rohini Srihari and Wei Li “Information Extration Supported Question Answering”
- [11] 《专家系统》 武波 马玉祥 北京理工大学出版社 2000 年
- [12] 黄梯云 《智能决策支持系统》 电子工业出版社 2000 年

致谢： 本文作者对评审人员提出的宝贵意见和建议表示衷心的感谢！