红楼梦人物关系问答系统*

王树西　刘群　白硕

(中科院计算所软件室, 北京 100080)

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

摘要: 论文在分析专家系统起源、发展，现有理论技术的基础上，采用自然语言问答的人机交互方式，搭建了《红楼梦人物关系问答系统》，对专家系统现存的问题进行了有益的探索。测试结果表明，该系统知识完备，表示方法及组织方法适当，求解问题质量高，人机交互便利，效率高，可维护性好，解释能力强。

关键词: 专家系统 知识库 规则库 模式库 推理机 模式匹配

1 引言

1950 年，图灵(Turing)发表了里程碑式的论文《电脑能思考吗？》。在文中，图灵第一次提出“机器思维”的概念，并且提出了判断计算机系统是否具有智能的实验方法--著名的“图灵测试”。“图灵测试“的参加者是计算机系统，被实验的人以及主持试验的人，这里的计算机系统，可以看作是一个专家系统。

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来理解与求解问题的知识系统,是人工智能领域中最广泛、最实用、最有成就的分支。

1968 年，费根鲍姆等人研制成功第一个专家系统 DENDRAL，自此之后，各种不同功能、不同类型的专家系统相继建立，如 PROSPECTOR、MYCIN、XCON 等。专家系统的开发技术，也取得了长足的进步，体系结构由最初的单一知识库及单一推理机发展为多知识库和多推理机，由集中式发展为分布式。在知识获取方面，已逐渐用半自动方式取代原来的手工方式。在知识表示及推理方面，也由原来的精确表示及推理发展为不确定性处理理论。此外，人们建立了多种不同功能、不同类型的专家系统开发工具。

当然，专家系统还存在不少有待解决的问题。例如，知识的完备性问题、知识的自动获取问题、深层知识的表示与利用问题、分布式知识的处理问题、多专家的合作与综合问题、常识性知识的推理问题等等。本文以《红楼梦人物问答系统》为例，对相关问题进行了有益的探索。

2 知识获取、检测与存储

专家系统的基础是知识。人类专家之所以能称为“专家“，是由于他掌握了某一领域的专门知识，使得他在处理问题时能比别人技高一筹。一个专家系统为了能像人类专家那样的工作，就必须具有专家

*本文有关研究得到国家重点基础研究项目 （ G19980305(07-4 和 G19980305(10 ） 资助。

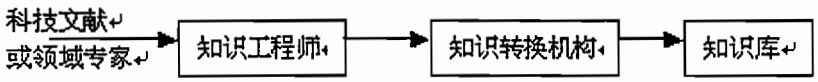
级的知识，知识越丰富，质量越高，解决问题的能力就越强。

2.1 知识获取

为了得到知识，就必须具有获取知识的能力。按知识获取的自动化程度划分，可分为自动知识获取和非自动知识获取两种方式。

所谓自动知识获取是指系统自身具有获取知识的能力，它不仅可以直接与领域专家对话，从专家提供的原始信息中“学习”到专家系统所需的知识，而且还能从系统自身的运行实践中总结、归纳出新的知识，发现知识中可能存在的错误，不断自我完善，建立起性能优良、知识完备的知识库。自动知识获取是一种理想的知识获取方式，为达到这一目的，它至少应具有识别语音、文字、图像的能力，理解、分析、归纳的能力，以及从运行实践中学习的能力。遗憾的是，目前专家系统在这些方面的能力还比较弱。就目前已经取得得研究成果而言，尚不足以真正实现自动知识获取，因此，知识的完全自动获取目前还只能作为人们为之奋斗的目标。

当前的知识获取主要是非自动知识获取，其过程大致可以分为两步：知识工程师负责从领域专家或科技文献里面抽取知识，并用合适的模式把知识表示出来；专家系统中的知识转换机构负责把知识转换为计算机可存储的内部形式，然后把它存入知识库中。整个过程实际上是一个知识转换的过程，这是因为，人类专家或科技文献中的知识通常是用自然语言、图形、表格等形式表示的，而知识库中的知识是用计算机能够识别、运用的形式表示的，两者有较大的差异。为了把从专家及有关文献中抽取出来的知识送入知识库供求解问题使用，需要进行知识表示形式的转换工作。其工作方式如图所示。



非自动知识获取

例如：《红楼梦》原著中有这么一句话：“姓甄，名费，字士隐。嫡妻封氏，情性贤淑，深明礼义“。通过这句话，可以抽取出甄士隐和封氏之间的关系，并且表示为“封氏是甄士隐的妻子”这种模式。这是知识获取的第一步，然后通过专家系统内部的知识转换机构，把这条知识转换为计算机可存储、识别的内部形式：”qizi('封氏','甄士隐').”。

2.2 知识的检测与求精

知识库的建立过程是知识经过一系列变换进入计算机系统的过程，在这个过程中存在着各种各样导致知识不健全的因素。例如：文献或领域专家提供的知识存在某些不一致、不完整、甚至错误的知识；或者由于未能正确理解领域专家或文献的意思，使得所形成的知识条款隐含着种种错误；或者采用的知识表示模式不适当，不能把领域知识准确的表示出来，等等。由于这些原因，知识库中经常会出现这样或者那样的问题，主要表现为知识冗余、矛盾、从属、环路、不完整等方面。

为了保证知识库的正确性，需要做好对知识的检测，检测分为静态检测和动态检测两种，本文采用静态检测和动态检测两种方法来保证知识库的正确性。静态检测是指在知识输入之前由领域专家及知识工程师所做的检查工作。例如，在知识输入之前，通过人工的检查，发现下面四条知识：“贾宝玉是贾政的父亲”，“贾政是贾宝玉的父亲”，“贾政是贾珠的父亲”，“贾政是贾珠的父亲”，显然，第一条知识“贾宝玉是贾政的父亲”是错误的，第四条知识“贾政是贾珠的父亲”是冗余的，都应当应该删除。动态检测是指在知识输入过程中以及对知识库的增、删、改时由系统所进行的检查。在系统运行过程中出现错误时也需要对知识库进行动态检测。例如，“北京和上海是两个直辖市”是地理方面的一条知识，由于本

系统处理的是人物关系，所以在处理过程中，会检测到这是一条无关的知识，无法进行知识转换，从而将其抛弃。

2.3 知识的存储

知识的存储形式，或者说是知识库，用于存放从文本中抽取,并经系统转换的知识。知识库的基础是知识表示，知识表示有多种方式，采用哪一种取决于是否方便系统分析问题。本文采用逻辑表示的方法，这里所指的逻辑指一阶谓词逻辑。系统存在一个亲属词对应表，亲属词和其谓词表达式相互对应，例如，“妻子“对应”qizi”，”儿子”对应“son”，等等。以下是知识库里面的两条知识：

[K1] qizi('王夫人','贾政'). 表示“王夫人是贾政的妻子“这条知识。

[K2] son('贾政','贾代善'). 表示“贾政是贾代善的儿子“这条知识。

3 推理机制

推理机是专家系统的“思维”机构，是构成专家系统的核心部分。其任务是模拟领域专家的思维过程，控制并执行对问题的求解。它能根据当前已知的事实，利用知识库中的知识，按一定的推理方法和控制策略进行推理，求得问题的答案或证明某个假设的正确性。

系统的推理规则是人工定义的，所有规则的集合形成规则库。规则可以视为事实的伸延，只不过是附加了条件，要这个规则为真，必须符合规则中的条件。规则的实质就是储存起来的查询。由于本系统处理范围有限（红楼梦人物关系），所以规则库是有限的。

规则分为两部份：第一部份和事实差不多（一个有变元的谓词）；第二部份包含其它短句（事实或规则，以逗号分隔），这是产生式的表示方法。例如,下面就是一条推理规则：

[R] erxifu(X,Y):-qizi(X,Z),son(Z,Y).

其中，erxifu(X,Y)表示 X 是 Y 的儿媳妇，qizi(X,Z)表示 X 是 Z 的妻子，son(Z,Y)表示 Z 是 Y 的儿子。[R]表示：如果 X 是 Z 的妻子，Z 是 Y 的儿子，那么 X 是 Y 的儿媳妇。

汉语是亲属词丰富程度非常高的语言（比英语要丰富，或者说颗粒度要细）。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(P)、夫妻关系(M)、长幼关系(O)。最基本的属性是性别属性(S)。

可以利用最基本的亲属关系和性别属性推理出其它亲属关系。例如，下面是“父子“这种亲属关系的推理表达式。

[1] father_son(X,Y):-son(Y,X),male(X).

[2] father_son(X,Y):-husband(X,Z),son(Y,Z).

其中，father_son(X,Y)表示 X 和 Y 是父子关系，male(X)表示 X 是男性，son(X,Y)表示 X 是 Y 的儿子，husband(X,Y)表示 X 是 Y 的丈夫。

4 人机交互方式

这里所说的人机交互指的是专家系统与一般用户之间的交互，本系统采用自然语言问答的人机交互方式。这就要求专家系统能够充分理解用户用自然语言提出的问题，为此，本文采用模式匹配的算法。

现有的模式匹配技术中，一种是关键字匹配。系统中预先存放了一定数目的含有关键词的基本模式，每一个模式都与一个或多个解释相对应。系统将句子与这些模式逐一匹配，一旦成功即可得到这个句子的解释。至于句子中那些不属于关键词的成分，系统则不考虑。这种技术的优点是处理简单、效率高，并且对一个文法上不完全正确的句子，只要它含有特定的关键字，就能做出相应的处理。另一种是句法模式匹配。它要求句子必须符合系统允许的一个句法模式，否则系统将不予接受。这种技术的缺点首先

在于要求系统中的句法模式不能太少，否则将导致系统理解效率低下；其次句子必须严格遵循句法模式的要求。

在分析现有模式匹配技术的基础上，本文提出一种新的模式匹配方法。判断字符串 sMatch 和模式 sPattern 是否匹配的具体算法是：

(1) 确定 sPattern 中所有非变量字符串及其位置、变量字符及其位置，转 (2)；

(2) 利用 sPattern 中所有非变量字符串顺序的匹配 sMatch，匹配位置开始为零，根据每次匹配的非变量字符串依次后移，如果匹配失败，错误返回；否则转 (3)；

(3) 确定 sPattern 每个变量表示的 sMatch 中字符串，转 (4)；

(4) 如果 sPattern 中的一个或多个变量表示的字符串为空，错误返回；否则转 (5)；

(5) 如果 sMatch 同时匹配到多个模式，取变量个数最多的模式作为最佳模式；

(6) 正确返回。

系统所需模式的集合形成模式库。模式库里面存放的模式有两类：一类是从文本中提取知识的模式，比如，文本中的句子“贾宝玉是贾政的儿子”与模式“X 是 Y 的儿子”相匹配；另一类是支持人机对话的模式。例如，用户的问题“谁是贾政的儿子”和模式“X 是 Y 的 Z”相匹配。客观问题的复杂性要求系统模式库中存放大量的各种各样的模式。模式越多，系统的自然语言处理能力就越强。

对于用户以自然语言提出的问题，系统通过上述模式匹配算法，转换成系统理解的模式，提交系统处理，然后将处理结果转换为自然语言的形式，提交给用户。

5 红楼梦人物关系问答系统

基于上述分析，本文构建了红楼梦人物关系问答系统。本系统允许用户通过自然语言，询问《红楼梦》里面人物之间的关系。

5.1 系统的结构和工作流程

系统结构如下图所示。它包括知识获取子系统、推理子系统、会话子系统、知识文本以及知识库。其中，规则库包含在推理子系统里面，结合 prolog 完成推理机制；模板库包含在知识获取子系统里面，结合模式匹配机制完成知识抽取和存储任务。通过会话子系统，用户和系统进行交互：用户用自然语言向系统提问；系统以自然语言的形式，将处理结果提交给用户。

系统工作流程：

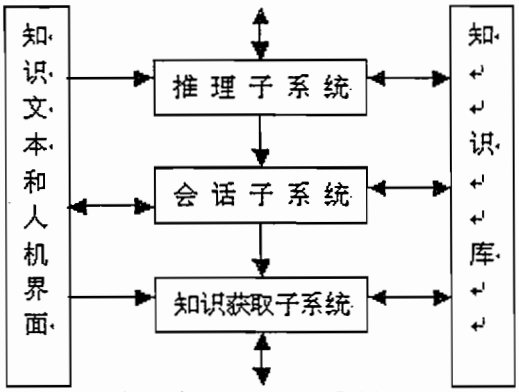
(1) 通过知识获取子系统，将知识文本中的知识转换为计算机可存储的内部形式，然后存入知识库。这是一个预处理的过程。转 (2)；

(2) 通过会话子系统和用户交互，得到用户的问题；转 (3)；

(3) 通过知识获取子系统里面的模式匹配机制，对用户问题进行处理，从中得到知识的形式化表示 λ。如果匹配失败，错误返回；如果成功，转 (4)；

(4) 通过推理子系统，对 λ 进行推理，如果推理失败，错误返回；否则转 (5)；

(5) 以自然语言的形式，将推理结果提交给用户，退出。



红楼梦人物关系问答系统结构

5.2 系统实现、演示和性能评价

我们以C++和prolog为工具，实现这个系统。系统界面和底层处理（包括模式匹配机制）由C++完成，中间的推理机制由prolog结合规则库完成，C++和prolog相互调用。在CPU PIII667、内存192MB、操作系统Windows 2000 Professional上的机器上面，系统运行状况良好。

下面是用户和系统的对话过程。【问】表示用户的问题，【答】表示机器的回答。

【问】: 贾宝玉和薛宝钗是什么关系?

【答】: 贾宝玉是薛宝钗的丈夫,薛宝钗是贾宝玉的妻子;
贾宝玉是薛宝钗的姨表弟,薛宝钗是贾宝玉的表姐;
共有 2 种关系。

在这个例子中，系统根据用户的问题“贾宝玉和薛宝钗是什么关系？”，结合模式库，利用模式匹配机制，得到匹配这个句子的最佳模式“A和B是C关系“，并且判断出是问题的类型是查询两人之间的关系。结合知识库里面的相应知识，分别判断出A(贾宝玉)和B(薛宝钗)的性别。利用推理规则near_relation(X,Y,Relation)进行判断，得知两个人有直接的亲属关系，然后，从推理规则male_female(X,Y,Relation)开始，利用一系列的推理规则，结合知识库，得到A(贾宝玉)和B(薛宝钗)之间的关系。最后，以自然语言的形式，将答案提交给用户。

【问】: 林黛玉是贾蓉的什么人?

【答】: 林黛玉 是 贾蓉 的 姨父 的 姑表姐妹
详细解释:
林黛玉 是 贾琏 的 姑表姐妹，贾琏 是 林黛玉 的 表兄弟;
贾蓉 是 贾琏 的 外甥，贾琏 是 贾蓉 的 姨父;
林黛玉是贾蓉的表姑(婶,舅妈,姨妈)。贾蓉是林黛玉的表侄(外甥)。
共有 1 种关系。

在这个例子中，系统根据用户的问题“林黛玉是贾蓉的什么人？”，结合模式库，利用模式匹配机制，得到匹配这个句子的最佳模式“A是B的什么人“，并且判断出是问题的类型是查询两人之间的关系。结合知识库里面的相关知识，分别判断出A(林黛玉)和B(贾蓉)的性别。利用推理规则far_relation(X,Y,Z,R1,R2,D1,D2)进行判断，得知两个人没有直接的亲属关系，但是分别和另外一个人“中介人”有直接的亲属关系。然后，从推理规则far_relation(X,Y,Z,R1,R2,D1,D2):-female(X),male(Y),female(Z),far_female_female(X,Z,R1,D1),far_male_female(Y,Z,R2,D2).开始，利用一系列的推理规则，结合知识库，分别得到A(林黛玉)和B(贾蓉)之间的关系与“中介人”贾琏的关系。最后，以自然语言的形式，将答案提交给用户。

软件工程中通常以正确性和健壮性为主来评加一个软件的性能。对于专家系统来说，正确性显然指的是它回答问题的正确率；而健壮性主要看它在错误数据输入的条件下的反应如何。

系统的模式库所包含的提问模式，几乎涵盖了所有用户常用的提问模式，具有相当的代表性。从上面的对话记录可以看出，对于“近亲”，系统可以给出准确、全面的回答；对于“远亲”，系统也可以通过“中间人”，把两个人的关系“连接”起来，给出合情合理的答复。系统的正确性是令人满意的；

对于无法识别的亲属词，系统给出错误的原因。但是，只要在亲属词文件（文本文件）中加入相应的亲属词，系统马上可以给出正确的回答。对于超出系统推理能力的问题，系统给出错误的原因。例如，如果系统无法通过“中间人“，把两个人的关系“连接”起来，那么给出错误的回答。

由此可以看出，系统具有令人满意的健壮性，并且因为程序跟数据分开，我们可以随时修改、添加规则、模式、知识而不用修改程序系统，系统具有和灵活性，易维护性的特点。

6 结束语

本文在分析现有专家系统理论技术的基础上，采用自然语言问答的人机交互方式，搭建了《红楼梦人物关系问答系统》，对专家系统现存的问题进行了有益的探索。通过测试表明，该系统系统具有令人满意的健壮性，并且具有灵活性，易维护性的特点。该系统人机交互便利，解释能力强。

本文在分析现有模式匹配技术的基础上，本文提出一种新的模式匹配方法。实验表明，这种模式匹配方法是比较有效的。

当然，本文还存在一些问题，对现有专家系统存在的一些问题还没有完全解决，这将是本文下一步要做的工作。

参考文献:

[1] 白硕 《计算语言学教程》(电子版)

[2] 白硕 《Reasoning Without Deep Structure 》(PowerPoint)

[3] 史忠植 《高级人工智能》 科学出版社 1998 年

[4] 陆钟万 《面向计算机科学的数理逻辑》科学出版社 1998 年

[5] 王永庆 《人工智能原理与方法》 西安交通大学出版社 1999 年

[6] 张寿权 周建峰 《专家系统建造原理及方法》 中国铁道出版社 1992 年

[7] 曹存根 建一个国家的大实验室 http://www.people.com.cn/GB/paper53/3073/410477.html)

[8] 许洪波《基于 Web 的问答系统的关键技术》(PowerPoint)

[9] 黄梯云《智能决策支持系统》 电子工业出版社 2000 年

[10] 黄雄《小灵通问答系统》(电子版)

[11] Encarta(http://encarta.msn.com/)

[12] 尤利卡智能搜索引擎(www.ulika.com)

[13] http://www.hownet.com

[14] 计算机科学家图灵(http://www.longen.com/S-Z/details~z/Turning.htm)

[15] http://www.how-net.com

[16] http://www.cyc.com/

[17] http://www.wordnet.com/

[19] AskJeeves(http://www.askjeeves.com)

[20] 程慧霞《用 C++建造专家系统》 电子工业出版社 1995 年

[21] 蔡自兴、徐光佑 《人工智能及其应用》清华大学出版社 1996 年

作者简介： 王树西，1976 年生，男，山东郓城人，博士生，主要研究领域为人工智能、计算语言学；白硕，

男, 博士, 教授, 博士生导师, 主要研究领域为信息安全系统, 人工智能, 计算语言学; 刘群, 男, 副研究员, 硕士生导师, 主要研究领域为机器翻译, 人工智能。

A QA system on Character relationship in Hongloumeng*

WANG Shuxi LIU Qun BAI Shuo

(Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, China)

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

Abstract: The paper analyzed the origin ,development and current technology of Expert System. Based on all of the above,we adopted the interactive method of natural language and put up "A QA System on Character Relationship In Hongloumeng". The result of the system indicate that the knowledge of the system is maturity and the denote method is propriety. The quality of maintenance, efficiency and explanation ability of the system is high.

Key words: Expert System; Knowledge base; Rule base; Pattern base; Reasoning machine;Pattern match;

* Supported by the National Natural Science Foundation of China under Grant No.00000000