# 基于领域知网的中文自动答疑系统的设计①

## Design of Chinese Q&A System Based on Terrain-Hownet

余文利 (衢州职业技术学院 信息与电力工程系 浙江 衢州 324000)

**摘 要:** 针对传统中文自动答疑系统的不足，借鉴《知网》的思路，并在此基础上提出了一种基于领域知网的中文自动答疑系统 CQASTH。为验证系统答疑的准确性，实现了一个实验系统 CMHQAES。实验表明，CQASTH 实现了语义理解和语义计算，提高了答疑的准确率。

**关键字:** 中文自动答疑 领域知网 语义理解 语义计算

## 1 引言

近年来，远程教育在我国得到了快速的发展，各种网络教学平台不断涌现，从本质上来说远程教育是指教与学在时空分离的状态下实施的教学。因此，作为网络教学中师生沟通与交流的最主要方式的网络答疑就显得尤为重要。

目前，国内网络教学平台主要采用 E-mail、在线讨论、关键词查询、FAQ(常见问题库)四种方式答疑。它们都有很大缺陷。在 E-mail 方式[1]中教师不能及时地把答案反馈给学生，且同一个问题多次回答；在线讨论方式要求教师时时在线。这两种方式都造成了教师和答案资源的极大浪费。关键词查询方式要求学生具备一定的关键词抽取能力，查询效果也不理想[2]；FAQ 方式只考虑了关键词词形，没有对关键词进行语义理解；实现的是也只是关键词间的精确匹配。经过研究认为，对学生提问进行语义理解并在此基础上进行语义相似度计算和模糊匹配是网络准确答疑的关键。本文基于这一思想，借鉴知网的思路，给出了领域义原的定义和提取方法，提出了领域关键词的标注原则和方法，定义了领域义原树及其建立方法。通过它们，实现了对领域关键词的语义理解。在此基础上改进知网义原和概念的语义相似度计算方法，实现了对领域关键词的语义相似度计算，对于知网未能解决的未登录词的相似度计算问题提出了解决方法。并将上述内容作为有机整体，构建了领域知网。将领域知网和句子语义相似度计算引入中文自动答疑系统中，设计了一个基于领域知网的中文自动答疑系统 CQASTH(Chinese Q&A System Based on Terrain-Hownet)，实现了对领域关键词的语义理解和语义计算，并在此基础上实现句子语义相似度计算和模糊匹配。本文基于 CQASTH 的设计思想，实现了一个实验系统 CMHQAES (Chinese Modern History Q&A Experiment System)以验证 CQASTH 的语义理解和语义计算。实验表明，CQASTH 实现了语义理解和语义计算，提高了答疑的准确率。

## 2 《知网》

Ontology(本体论)也被称之为世界知识，Ontology 最早是一个哲学的范畴，后来被人工智能界给予了新的定义。Studer[3]认为 Ontology 是人工智能中一种知识表现方法，是一种形式化的对共享概念的明确表述。

《知网》[4](HowNet)是最为著名的采用汉语描述的本体论，它是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。《知网》通过义原[4]的组合来标注各种各样的单纯的或复杂的概念，以及各个概念之间、概念的属性和属性之间的关系。相对来说，新词虽然层出不穷，但义原的增加却极少。在《知网》中，词义就是定义为各种义原的组合。

① 收稿时间:2008-08-25

## 3　基于领域知网的中文自动答疑系统

### 3.1 系统结构

CQASTH 采用三层结构和 B/S 模式，采用这种结构使系统内各层之间在功能上相对独立，各自独立完成不同任务，便于系统的设计开发和管理。根据各功能模块在系统中的地位和作用，将 CQASTH 从上到下分为接口层、功能层和知识层，其结构如图 1 所示。

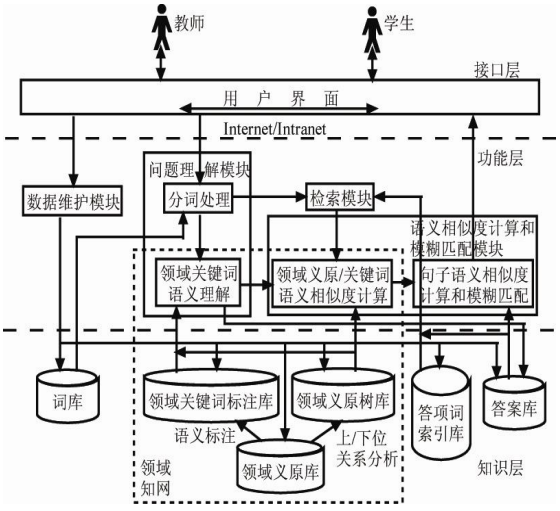


图 1　基于领域知网的中文自动答疑系统结构图

①接口层：接口层包括学生端和教师端(教师即为领域专家)，接受学生以日常习惯的句子形式的提问，并将系统返回的答案呈现给学生；接受教师指令并将其提供给系统。它完成系统与师生的交互功能。

②功能层：根据学生的提问进行问题理解、检索、语义相似度计算后找出最匹配的答案，各模块相互配合，完成系统的自动答疑功能。

③知识层：主要存储 CQASTH 中的各种资源，主要有词库、答项词索引答案库、领域知网知识库(它由领域义原库、领域义原树、领域关键词标注库构成)等。

接口层有一部分位于浏览器端，有一部分位于服务器端。功能层和知识层都位于服务器端。

### 3.2 工作流程

在领域知网模型的基础上构造的中文自动答疑系统 CQASTH 的工作流程如图 2 所示。

具体的工作流程为：

①学生利用学生端用户界面向系统提交问题后，系统调用分词处理功能模块得到问题领域关键词和答项词，通过语义理解功能块检索领域知网知识库中的领域关键词标注库，对问题领域关键词进行领域义原标注。

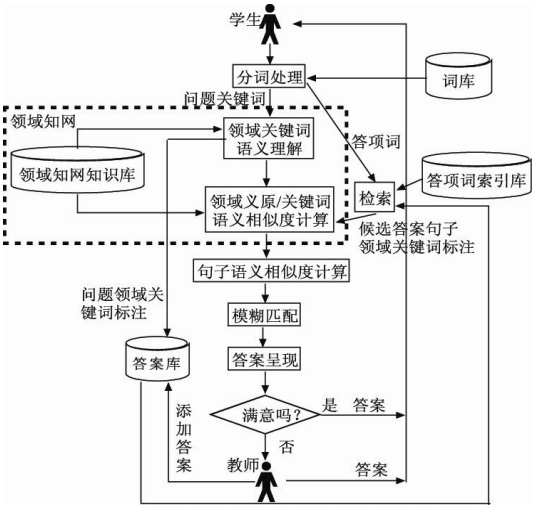


图 2　基于领域知网的中文自动答疑系统流程图

②由答项词特性调用检索功能块在答项词索引库中找到候选答案句子编号并以此在答案库中找到候选答案句子领域关键词标注。

③由句子相似度功能模块调用领域知网中语义计算功能模块并在此基础上计算问题句子与候选答案句子的语义相似度后进行模糊匹配，将问题与候选答案相似度大于某一阀值的答案序列按匹配值大小递减顺序呈现给学生。

④通过用户界面将匹配结果以送给学生端，学生判断匹配结果，如正确则接受，否则向教师求助人工答疑。

## 4　关键技术

### 4.1 领域知网知识库的构建

领域知网知识库中存储领域关键词间的语义信息和语义关系，是 CQASTH 对领域关键词进行语义理解和语义计算的语义知识资源，包含有领域义原库、领域义原树库、领域关键词标注库。

#### 4.1.1 领域义原的提取和编码规则

4.1.1.1 领域义原的提取：对于某门课程，以其领域关键词为基础，通过对领域关键词的分析来提取该门课程的领域义原集合。其流程如图 3 所示。

具体过程为：①提取领域关键词。领域关键词是指某门课程中能够代表某一特定概念的词语。教师根据本门课程特点和学生常见提问从该门课程中提取出所有的领域关键词。②对抽取的所有领域关键词进行

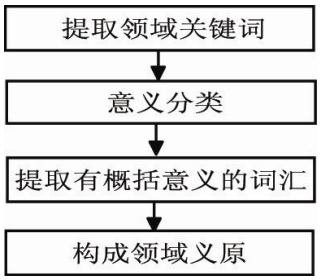


图 3　领域义原提取流程图

本体分析。本体分析是对某门课程的概念体系形式化过程，在本体分析中最重要的就是分类关系的分析，分析的方法是将该门课程中的领域关键词汇总后对它们进行意义上的分类，在从每一类中提取有概括意义的词汇，这些词汇构成领域义原。

4.1.1.2 领域义原的编码规则：为便于计算领域义原语义相似度，对领域义原进行编码，其编码规则为：①编码从根专业义原向下逐层依次进行直至所有的叶子领域义原；②根领域义原的编码为 H；③每个领域义原的下位领域义原，其左右顺序在编码前可自由调整，但一经确定后则应固定，并从左到右依次排列号为 1、2…9、A、B、C、D、E、F；④每个下位领域义原的编码为其上位领域义原的编码加上其在同层领域义原的排列号；⑤领域义原的编码长度与其所处的层数相同。

4.1.2 领域义原树

领域义原树体现了领域义原的上下位关系，它是进行语义相似度计算的语义资源。提取了领域义原后即可分析义原间的上下位关系和在领域义原树中的位置，构建领域义原树。在此，本文只利用了领域义原间的最主要的语义关系即上下位关系构建领域义原树。以《中国现代史》为例构建的领域义原树如图 4 所示。

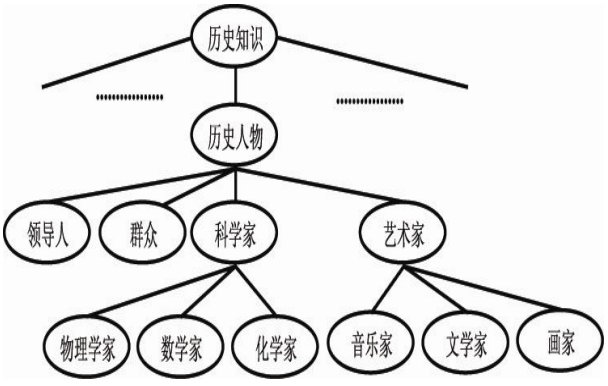


图 4　中国现代史专业义原树(部分)

4.1.3 领域关键词标注

为了准确地建立领域关键词标注，本文制定了一套领域关键词标注规则：(1)任何领域关键词的标注都以“DEF=”为开始，以“。”为结束，如有多个标注项则以“，”来分隔；(2)标注领域关键词的第一个领域义原必须指出领域关键词的最基本的意义；(3)领域关键词的标注中使用的领域义原须从叶子领域义原开始。如叶子领域义原不能准确阐释其意义，可逐层向上推移，用其上位领域义原来阐释。(4)同义词的 DEF 项应相同。以《中国现代史》为例构建了领域关键词标注如下所示：

NO. = 100010
W_C = 毛泽东
G_C = N
E_C =
DEF = 领导人，@现代，中国

其中，DEF 项表示对“毛泽东”这一领域关键词的标注。DEF 项中“领导人”、“现代”、“中国”为领域义原，“@”表示对应领域义原(“现代”)和“毛泽东”这一领域关键词间的时间关系。DEF 项是对“毛泽东”这一领域关键词的领域义原解释，体现了领域关键词与领域义原间的语义关系，是用来进行语义计算的语义资源。每个领域关键词都是由有限个领域义原表示的，而所有的领域义原构成了一个层次结构的领域义原树，所以根据领域关键词的标注，并借助领域义原树，就可以体现出领域关键词间的语义信息。

**4.2 领域知网语义相似度计算**

4.2.1 领域知网中领域义原语义相似度计算

度量两个领域义原语义相似度的指标是二者在领域义原树中的距离，二者距离越大，其语义相似度越低；反之二者的距离越小，其语义相似度越大，二者可以建立一种简单的对应关系。

为便于说明，本文作如下定义：

定义 1.领域义原深度—是指领域义原树中根领域义原到该领域义原(X)的路径边数，记为 H (X)。

定义 2.领域义原距离—是指两个领域义原 $X_1$、$X_2$ 在领域义原树中的最少边数，记为 L ($X_1$，$X_2$)。

根据以上分析定义，领域义原间的语义相似度计算公式为：

定义 3.设 $x_1$ 和 $x_2$ 是领域义原树上的两个领域义原，它们间的语义相似度记为 Sim ($x_1$, $x_2$)，其计算公

式为：

$$Sim(x_1,x_2)=1-\frac{L}{H_1+H_2} \quad (1)$$

其中，L 为领域义原 $x_1$ 和 $x_2$ 的领域义原距离。$H_1$、$H_2$ 为领域义原 $x_1$ 和 $x_2$ 的领域义原深度。

4.2.2 领域知网中领域关键词语义相似度计算

在领域义原语义相似度计算的基础上，可以实现领域关键词语义相似度计算。领域关键词语义相似度是指两个领域关键词在上下文中相互替换而不改变其语义的程度，反映领域关键词在语义上的符合程度，语义相似度是一个数值，一般取值范围在[0，1]之间，一个词与其自身和同义词的语义相似度为 1，如果两个词在语义上没有任何关系，则二者语义相似度为 0。在某门课程中，领域关键词间及其领域义原间的语义关系少得多且比较简单，用领域义原和关系标注符号就能标注领域关键词，所以领域关键词的标注分成三个部分：

①第一基本领域义原描述式，DEF 项中的第一个领域义原；

②其他基本领域义原描述式，DEF 项中除第一独立领域义原以外的所有其他独立领域义原；

③关系领域义原描述式，DEF 项中用“关系领域义原 基本领域义原”描述领域关键词的部分。两个领域关键词这三部分对应的语义相似度分别记为 $Sim_1(C_1,C_2)$、$Sim_2(C_1,C_2)$、$Sim_3(C_1,C_2)$。同时，令领域关键词的整体语义相似度为：

$$Sim(C_1+C_2)=\beta_1 Sim(C_1+C_2)+\sum_{i=2}^{3}\beta_1\beta_i Sim(C_1,C_2) \quad (2)$$

其中，$\beta_i(1\leqslant i\leqslant 3)$是一个可以调节的参数，各部分的重要程度通过 $\beta_i$ 值进行限定，并满足：$\beta_1+\beta_2+\beta_3=1$，$\beta_1\geqslant\beta_2\geqslant\beta_3>0$[5]。

以下讨论每一部分的语义相似度计算：

①第一基本领域义原描述式：假设两个领域关键词 C1 和 C2 的第一基本领域义原分别是 $x_1$ 和 $x_2$，则二者的语义相似度由公式(1)决定。

②其他基本领域义原描述式：假设两个领域关键词 $C_1$ 和 $C_2$ 的其他基本领域义原分别为集合 $Set_1=\{x_{11}, x_{12}, \cdots x_{1m}\}$ 和 $Set_2=\{x_{21}, x_{22}, \cdots x_{2n}\}$，则该部分的语义相似度计算如下：

令 SIZE=max{m,n}，$|Set_1|$和$|Set_2|$分别表示两个集合当前拥有的领域义原数量，score=0.0 表示两个集合当前语义相似度之和；

while($|Set_1|>0$ or $|Set_2|>0$) {

求出两个集合所有组合中语义相似度最大的一组领域义原 $x_i\in Set_1$ 和 $x_j\in Set_2$，$x_i$ 或 $x_j$ 有可能是空值，但不会同时为空；如其中一个为空，则二者的语义相似度为 0。

score=score+Sim($x_i$,$x_j$);

Set1=Set1 −{$x_i$};

Set2=Set2 −{$x_j$};

}

Sim($C_1$,$C_2$)=score/SIZE.

③关系领域义原描述式：该部分的语义相似度计算也是一个集合运算的问题，基本流程与(2)类似。不同地方在于，集合在两两配对分组计算关系领域义原的语义相似度时，需要首先判断关系类型(即等左侧的关系领域义原)是否相同，如果不同，则两个领域义原的语义相似度为 0，否则，取出右侧的领域义原名称，并根据公式(1)计算其语义相似度。

4.2.3 句子语义相似度计算

句子语义相似度计算是在领域知网义原、领域关键词语义相似度计算的基础上计算学生提问句子和候选答案句子的语义相似度。其方法如下：

定义：设有两个句子 A 和 B，设句子 A 由领域关键词 $A_1$、$A_2$、…、$A_m$ 组成，句子 B 由领域关键词 $B_1$、$B_2$、…、$B_n$ 组成。则句子 A 和 B 的语义相似度记为 S(|AB|)，可用相似矩阵[6]计算得到。其计算方法如下：

①构造 A、B 的相似矩阵：

$$M(A,B)=\begin{bmatrix} s(A_1,B_1),s(A_1,B_2),\ldots s(A_1,B_n) \\ \ldots\ldots \\ s(A_m,B_1),s(A_m,B_2),\ldots s(A_m,B_n) \end{bmatrix} \quad (3)$$

其中，$s(A_i, B_j)$表示领域关键词 $A_i(1\leqslant i\leqslant m)$和 $B_j(1\leqslant j\leqslant n)$之间的语义相似度。通过式(3)计算获得矩阵中的第 i 行表示句子 A 中的领域关键词 $A_i$ 与句子 B 中所有领域关键词的语义相似度。

②计算 A 句子与 B 句子间的语义相似度 s(A，B)：

$$s(A,B)=\frac{\sum_{i=1}^{m}\max(s(A_i,B_1),s(A_i,B_2),\ldots s(A_i,B_n)}{m} \quad (4)$$

即取出句子 A 中每个领域关键词与句子 B 中所有领域

关键词语义相似度的最大值，然后对这些最大值进行相加，求平均值，便得到句子 A 与句子 B 之间的语义相似度。

由于矩阵的不对称性[7]，还要进一步计算句子 B 与句子 A 之间的语义相似度。用以上相同的算法，可以得到句子 B 和 A 间的语义相似度 s(B，A)。

③计算句子 A 和 B 之间平均加权语义相似度

S(|AB|)=(S(A,B)+S(B,A))/2　　(5)

4.2.4 未登录词相似度计算

未登录词的相似度计算一直是自然语言处理领域的难点，尽管在领域知网的构建过程中，做了较全面的的考虑，但是仍有许多未被考虑的词汇。由于领域知网没有对未登录词进行语义描述，因此需要将未登录词转化为领域知网可以理解的形式。在领域知网的未登录词中，有大量的组合词出现。组合词的特点在于它是由若干个词语组合而成，因此考虑对未登录词的计算方法在于先将其分解为字或词的集合，然后计算两个字、词集合间的相似度，在此综合考虑字和语义两种因素对未登录词进行处理，其计算过程如下所示：

①将未登录词 a、b 分解为字和词的集合 A、B；

②将未登录词的相似度 Sim(a，b)分解为计算基于字、词的相似度 $Sim_w(A,B)$ 和语义的相似度 $Sim_s(A,B)$；

③ $Sim_w(A,B)$、$Sim_s(A,B)$ 的计算方法如下所示。其中 sizeof(A)表示集合 A 中字或词的个数，sizeof(B)表示集合 B 中字或词的个数。

```
for(i=0;i<sizeof(A);i++)
    for(j=0;j<sizeof(B);j++)
        if(Ai = = Bj)
            Simw(Ai，Bj)++;
        else
            Sims(Ai，Bj)二 Sim(DEF(Ai)，DEF(Bj))
Sim(A，B)=Simw(A，B)+Sims(A，B);
Sim(A，B)/=Length(A，B);
```

其中，$Length(A,B)=Length(A)+\frac{Length(B)-Length(A)}{2}$，在这里假设 Length(B) >Length(A)，否则将词集 A、B 互换。

④基于语义计算有时候会夸大未登录词的相似程度，为此这里设定惩罚系数 r 并规定：当两个词集中某个词或字进行语义相似度计算时，如果它们字词集合中没有相同的字、词存在的话，将计算的结果乘以一个惩罚系数 r=0.2。

## 5 实验结果与讨论

为验证 CQASTH 模型的可行性和准确率，本文设计并实现了一个实验系统—中国现代史答疑系统 CMHQAES，分别用没有领域知网的答疑和有领域知网的答疑做了对比实验，选择的资源主要有：领域关键词词典、答项语词典、停用词词典、常见答案库、领域知网(标注了72个领域关键词)。实验数据来自《中国现代史》课程中的 125 个问题（Bl、B2、B3 的取值分别为 0.6、0.25、0.15）。以下是对其中两个问题测试的对比结果：

**问题 1**：“社会主义全面建设时期有哪些重要会议?”

①没有领域知网的情况。找出的是社会主义全面建设时期所有有关“会议”的知识，但不能找出诸如“八大”、“七千人大会”、“庐山会议”等具体内容。

②有领域知网的情况。能弥补关键词查询和 FAQ 方式的不足。因“八大”、“七千人大会”、“庐山会议”等和“社会主义全面建设时期”、“会议”之间有很高的语义相似度，系统经过分词、语义理解和语义相似度计算，找到了恰当的答案。

**问题 2**：“1968 年的大事”

①没有领域知网的情况。由于 1949 年后，海峡两岸分离，大陆采用公历纪元，而台湾仍用民国纪元。因为纪元方法不同，造成同一年代在两岸有不同的说法。如大陆的公元 1968 年即为台湾的民国 57 年。在关键词查询中，“1968 年的大事”是无法找到“民国 57 年的大事”的，反之亦然。

②有领域知网的情况。在中国近现代史领域知网中，本文将它们标注为相同的专业关键词，这个问题就迎刃而解了。

测试中，如果句子间语义相似度大于某个阀值(阀值可以根据实验结果设定(本实验设为 0.6)，且正确答案出现在语义相似度最大的前三位则认为匹配成功。对随机测试的 125 个问题的测试结果表明，答案的准确率为 96.5%，已达到令人满意的程度。因此，基于领域知网的中文答疑系统是切实可行的。

## 6 结束语

本文着眼于提高中文答疑系统的准确率和智能性，借鉴知网的思路，根据中文答疑系统的特点，设

(下转第 13 页)

(上接第 5 页)

计了领域知网模型，将语义理解和语义计算作为一个整体引入到中文答疑系统中，并对其加以改进。通过实验验证，基于领域知网的构建方法是非常有效的，较好地完成了中文自动答疑任务，并且具有很强的可推广性。

## 参考文献


1 Watanabe Y, Sono K. A question answer system using mails posted to a mailing list. Proceedings of the 2004 ACM Symposium on Document Engineering, 2004:67－73.

2 季永华,许华虎,沈敏,等.自动答疑系统的研究与实现.计算机工程与应用, 2005,41(14): 224－226.

3 Gruber T R. Toward principles for the design of ontologies used for knowledge sharing. Int Journal of Human and Computer Studies, 1995:907－928.

4 董振东.知网.http://www.keenage.com,2000-05.

5 Dong ZhengDong. HowNet and Computation of Meaning. Singapore: World Scientific Press, 2006:187－195.

6 Li SJ. Semantic Computation in a Chinese Question Answer system. Comput. Sei. & Technol.，2002，17(16): 933－939.

7 程云鹏.矩阵论(第 2 版).西安:西北工业大学出版社，2001.