基于章节本体和转移网络的自动答疑系统

王晓艳,赵政文,田振刚
WANG Xiao-yan,ZHAO Zheng-wen,TIAN Zhen-gang

西北工业大学 计算机学院,西安 710072
School of Computer,Northwestern Polytechnical University,Xi'an 710072,China
E-mail:xiaoyanw06610@gmail.com

WANG Xiao-yan,ZHAO Zheng-wen,TIAN Zhen-gang.Question answering system in network education based on chaptered ontology and transfer network.Computer Engineering and Applications,2009,45(24):207-209.

Abstract: Aiming at the inadequacy of ontology and transfer network in existing NEQAS applications,a method of chaptered ontology and transfer network is proposed.A confirming of question mode based on chaptered ontology and the making answers of standardization questions based on chaptered transfer network are presented.Semantic similarity model is discussed based on knowledge of chaptered ontology to solve the making answers of non-standardization questions,and the search method based on indexed transfer subnet is also built.Experimental results show that a method of chaptered ontology and transfer network can reduce the searching time by about 38% with the same efficiency as ontology and transfer network.

Key words: intelligent question answering system;ontology;transfer network;chapter

摘 要:针对现有本体和转移网络在网络教育自动答疑系统中应用的不足,提出了一种基于章节划分的本体与转移网络方法。定义了章节本体,确定了基于章节本体的问题模式,给出了基于章节划分转移网络解决规范化问题的答案生成。基于章节本体知识给出了语义相似度模型,用于解决非规范问题的答案生成;基于章节划分转移网络建立了章节索引转移子网的搜索方法。实验结果表明,基于章节划分的本体论和转移网络方法在不降低搜索效果的情况下,比一般的本体论和转移网络方法的查询时间减少了 38%。

关键词:智能答疑系统;本体;转移网络;章节
DOI:10.3778/j.issn.1002-8331.2009.24.062 **文章编号:**1002-8331(2009)24-0207-03 **文献标识码:**A **中图分类号:**TP311

网络教育自动答疑系统(Network Education Question Answering System,NEQAS)是一个面向特定领域的中文答疑系统,语义理解是 NEQAS 主要解决的技术难点[1]。本体(Ontology)是一种用来描述概念以及概念之间关系的模型[2]。在 NEQAS 中,本体具有非常重要的地位,是解决语义层次上信息共享和交换的基础。文献[3]引入了本体技术,构建网络教育领域本体.这种本体能提供完整的关于网络教育特定领域中概念以及概念之间关系的描述,使得目前 NEQAS 中存在的语义问题得到较好解决。

研究网络教育自动答疑系统,除了要研究自然语言的语义理解外,还要研究问题和答案的快速匹配问题。也就是说,当学生提出一个问题后,系统怎样在答案库中快速地找到答案。本体理论虽然很好地解决了语义理解问题,但是由于其全局关联性形成了庞大的本体网络,不利于答案的快速查找。据此,提出了基于章节本体的语义提取方法。

1 章节本体

1.1 领域本体与全局关联

NEQAS 中,构建本体的方法是要列举出有关课程的所有概念(专业词语)、概念的详细解释、每个概念所有可能的属性和属性值以及概念之间的各种关系等。构建的领域本体中,概念与概念之间存在继承关系、整体-部分关系、同义关系以及关联关系[4]。据此,可把领域中的概念构建成一个具有等级的网络结构。

领域本体建模虽然描述了有关课程内所有的概念与关系,称之为全局关联,但是却忽略了课程本身各概念间固有的联系。比如章节关系,作为教学内容,每个学科的知识点相对固定。其知识点间的相互关系也比较固定。不同版本的教材有时虽然章节顺序不一样,但章节的内容、章节的关系都不会有很大不同。例如,在《数据库系统原理》这门课里,"代数优化"不会和"加锁协议"放在一章;"加锁协议"一定和"并发控制"在一章中,这些都是固定的。即,每门学科都有自己固有的知识体系结构。抓住这个特点,可以利用这个固有的联系对领域本体加以约束形成章节本体。

1.2 章节本体的定义

定义 1 一个构建完成的章节本体可以视为一个无环的有向网络 G:$G=G(S,V)$。其中 $S=(C_1,C_2,\cdots,C_i,\cdots,C_j,\cdots,C_{\max})$,是图中所有结点的集合,每个结点表示特定领域中的一个概念,

作者简介:王晓艳(1979-),女,硕士,研究方向:计算机网络与软件;赵政文(1956-),男,硕士生导师,教授,研究方向:机器翻译;田振刚(1980-),男,硕士,研究方向:电子信息。

收稿日期:2009-02-12 **修回日期:**2009-04-20

$V=\{(C_i->C_j)|0<i<max,0<j<max$,且 C_i 和 C_j 之间存在某种关系$\}$,是图中所有有向边的集合,表示概念之间的关系。

领域本体中的概念可用实体类加以描述:

<实体类>::=实体类{

同义词:{<synonymy_set>}

概念的解释:<description>

继承关系:<is_a>

整体-部分关系:<part_of>

章节关系:<chapter_to>

关联:<关联、与关联结点相关的概念、关联结点的说明>

属性:<属性、属性值>}

这里的章节关系是一种广义的章节关系,可根据教材目录分为章、节、小节三层。如果考虑跨学科和课程,可在章的上面再加两层:学科和课程。图 1 所示为计算机专业关于数据库课程的某些概念间的关系。由于关联关系可从其他关系导出,图中不表示这种关系。领域中的所有概念用同样的方法加以描述,就构成了一个章节本体。该本体提供了丰富的语义信息是人们共同认可的、可共享的知识。

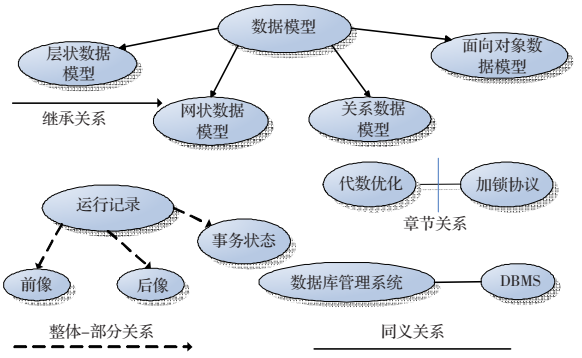

图 1　章节本体网中的各种关系

2　问题规范化

2.1　问题模式

在提问过程中，一个问题虽然可能会有多种提问方式,但是都会遵循一些固定的模式。学生一般是针对课程中的一个或多个概念而提出相关的问题，这种概念称之为该问题的主概念。系统除了要分析出问题的主概念外,还要分析它是针对主概念的哪些方面进行提问。例如,对概念进行提问、对概念的属性进行提问、对概念的一个关联关系进行提问。可以得到上述三个问题的模式[5]:

(1)主概念模式。这种模式的问题只涉及到领域中的一个概念。

(2)主概念属性模式。这种模式的问题是针对领域中概念的一个属性进行提问。

(3)关联关系(主概念 1,主概念 2)模式。这种模式的问题是针对概念 1 的关联关系进行提问,该关联关系涉及另一个概念 2。

在章节本体关系下,还需要增加一种模式——章节模式。

(4)章节关系(主概念 1,主概念 2)模式。这种模式在搜索问题相关概念与属性时起到阻止的作用,即跨章节的关联关系定义为停止关联。

2.2　问题的规范化

定义 2　从各种问题中抽取问题模式的过程称为问题的规范化。

问题的规范化主要考虑以下两个方面:(1)问题模式的确定;(2)同义词的统一。对学生问题进行分词处理后,可以得到一组词语,过滤掉其中与领域和问题不相关的词语后得到一组关键词语。

3　基于章节本体转移网络的问题匹配与推理

通过借助语义信息和问题模式库中的问题模式对这组关键词语进行分析,便可以确定问题的模式。问题模式确定后,再利用转移网络进行推理直接查找问题答案。

3.1　章节本体转移网络

转移网络是自然语言处理中常用的一种自动机[6],可用有向图来表示一个自动机的状态转移过程。用转移网络表示问题的文法，则每个转移网络由一个状态集和一个标号集组成,构造方式表示为:状态×标号→状态。含义是给定当前状态和当前状态的标号后,可以求得下一步的状态。上述四种问题模式对应的转移网络如图 2 所示。图中 Start 为初始状态;A、B、E、H 为可终结状态,用同心圆表示;C、D、F、G 为中间状态,用单圆表示;主概念、属性、关联关系以及章节关系是转移网络中的四种标号类型。其中,可终结状态表明:如果标号输入完毕,转移网络刚好到达该状态,则正常终止;如果还有标号输入,则继续向下一个状态转换;当遇到章节关系,则终止。中间状态表明:转移网络一般不能在中间状态终止;当遇到章节关系,则终止。

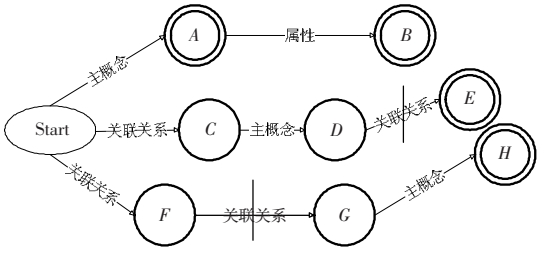

图 2　章节本体转移网络

对每一个终结状态，本体中都有一个相应的查找答案规则。当一个正常终止发生时,在找到正确答案的前提下,利用网络中丰富的语义资源,系统同时可以返回一组与用户问题相关的答案。答案中除了返回主概念的解释外,还同时返回与主概念属性相关的属性值、与主概念相关的子概念或父概念的解释等。例如,学生提问:“什么是数据库管理系统?”。此时,主概念是“数据库管理系统”,通过查找本体,系统给出正确答案:“数据库管理系统是管理数据库的软件,它实现数据库系统的各种功能”,同时给出“数据库管理系统的功能”、“什么是数据库系统”等相关答案。这里,“功能”是主概念“数据库管理系统”的属性,而“数据库系统”是“数据库管理系统”的父概念。

当一个章节终止发生时,它通常中断了一个关联关系。例如,学生提问:“什么是关系代数? ”。此时,主概念是“关系代数”,通过查找本体,系统给出正确答案,同时给出“关系代数的基本操作”、“什么是关系”等相关答案。这里,“基本操作”是主概念的属性,而“关系”是“关系代数”的父概念。同时,由于关联关系的存在,系统拟给出“关系演算”与“关系模式”相关的概念。但是从章节角度来讲“关系演算”与之不同节而“关系模式”与之不同章,属于应该提前终止的查找。可见,章节约束的引入加强了 NEQAS 的语义处理能力,避免了不必要的查找,加快了答案的查找速度。

3.2　语义相似度

利用章节转移网络对问题模式进行推理目前只能处理一些规范问题。对于一些非规范问题，NEQAS 采用基于章节本体的语义相似度匹配方法，即利用章节本体计算概念间的语义相似度[7]，然后通过概念间的语义相似度进一步计算用户问题和知识库中问题之间的语义相似度，对计算结果进行排序，取排在前面的作为问题的答案。

当章节本体网中任意两个概念 X_i、Y_j 在同一章节，则其语义相似度 $Sim(X_i,Y_j)$ 为：

$$Sim(X_i,Y_j)=2d_{\max}-Dist(c_i,c_j)=$$

$$2d_{\max}-\sum_{c\in\{path(c_i,c_j)-LSup(c_i,c_j)-C(c_i,c_j)\}} wt(c,parent(c)) \quad (1)$$

其中，$Dist(c_i,c_j)$ 为本体网中任意两个结点间的距离，通过结点 c_i、c_j 最短路径中不跨越章节分割的所有有向边的权重之和进行计算。由于既考虑了两个概念在本体网中相应结点间的距离，又考虑了概念间的章节关系，计算得到的概念间语义相似度值更具合理性。利用概念间语义相似度可进一步计算用户问题与知识库中问题之间的语义相似度。

4　章节转移子网与索引

使用章节本体转移网络构建的整个课程的转移网络图是一个带有层次边界的有向无环图，如图 3 所示就为一个一章四节的章节转移网络。其中，A、E、F、J、K、L、M、Q、B、C、G、H、N、O、T、X、D、I、P、U、V、W、R、S 分别为一节；整个图形成了一章。相对于整张的转移网络，称每一节形成的为子网。

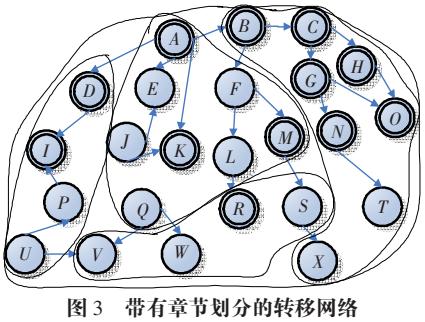

图 3　带有章节划分的转移网络

在使用转移网络[3]进行问题匹配与答案推理时，通常面临两个效率降低的障碍：搜索范围太大、结束条件太宽。结束条件太宽的问题在第 3 章已得到解决，针对搜索范围太大的问题建立了分层索引的方法。具体步骤如下：根据章节转移网络的层次关系，建立章节索引；一步步缩小包围圈，最后在最小的包围圈里填充转移子网。这时形成的索引结构如图 4 所示。

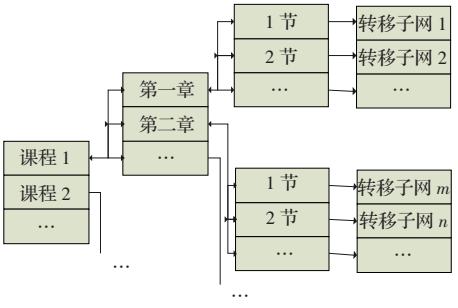

图 4　转移子网索引结构

索引层次和转移子网都是由系统预处理模块事先填好。系统运行后，首先在内存迅速生成索引结构。以后每次用户提问时，系统只需负责分析问题，其他数据可直接从内存的各个索引结构获得，查询速度很快。基于索引层次和转移子网的快速定位算法的步骤如下：

算法输入：问题向量（分词模块的输出）。

算法输出：一个或一组相关答案。

（1）分析问题向量中各维关键词语的“所属章节”属性（查章索引文件），取交集。即计算所有关键词语共同出现的章。把求得的共同章里的所有转移子网作为候选网。

（2）对每一个候选网，按照 3.1 节的方法寻找答案。如果命中则返回退出。否则根据 3.2 节的方法计算提问与概念向量的相似度。

（3）根据第二步的计算结果，取具有最大相似度候选答案的关键字。并由该关键字在候选网中获得答案的具体内容。

在带有章节划分的转移网络中，通过把问题的分类定位到章，可以缩小答案的查找范围。实际搭建系统时可以通过将章节索引结构常驻内存，使得系统能够很快地将问题与答案进行匹配。

5　性能分析

测试中，随机抽取出 30 个问题，使用了两个测试系统。第一个系统使用一般的转移网络定位算法，而第二个系统使用该文提出的带有章节划分的转移网络定位算法。反复实验 20 次，得到查询时间和查询效果的比较结果。其中：

（1）TN（Transfer Network），一般转移网络方法。

（2）CTN（Chaptered Transfer Network），带章节划分的转移网络方法。

图 5 是 20 次实验平均查询时间数据的比较折线图，横坐标是 30 个实验样本，纵坐标是查询时间。其中虚线部分为一般的转移网络结构平均查询时间；实线部分为带有章节划分的转移网络平均查询时间。

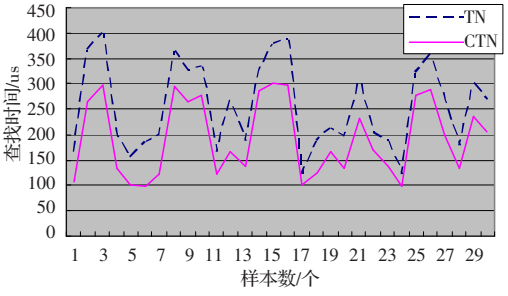

图 5　查找时间分布图

从图 5 中可以看到，使用了带有章节划分的转移网络的实验数据线集中在图的下方，使用一般的转移网络的实验数据线集中在图的上方，该文算法的有效性非常明显。经计算，较一般转移网络而言，带有章节划分的转移网络下平均查询时间少了 38%。

当然，仅仅对查询时间进行对比是不够的，因为无法保证查询效果的查询时间的减少是毫无意义的。评价检索系统的两个主要指标是查准率和查全率，定义为：

$$\text{查准率 } P=\frac{a}{b}\times 100\% \quad (2)$$

$$\text{查全率 } R=\frac{a}{c}\times 100\% \quad (3)$$

（下转 230 页）