

文章编号:1001-9081(2010)02-0415-04

## 基于本体的自动答疑系统的研究与实现

刘汉兴,林旭东,田绪红

(华南农业大学 信息学院,广州 510642)

(liuhx666@21cn.com)

**摘要:**针对现有自动答疑系统中知识表示的不足,提出了用本体构建课程领域知识库的方案。根据答疑问题的特点进行了问题分类,将用户问句意图转换为对本体中基本元素的查询,并通过 Jena 语句加以实现,最终抽取出答案,实验证明方案是可行的。

**关键词:**本体;自动答疑系统;知识表示;问题分类;答案抽取

**中图分类号:** TP18; TP391.1 **文献标志码:** A

## Research and implementation of automatic question answering system based on ontology

LIU Han-xing, LIN Xu-dong, TIAN Xu-hong

(College of Informatics, South China Agriculture University, Guangzhou Guangdong 510642, China)

**Abstract:** An approach to constructing course knowledge base with ontology was proposed to overcome the deficiency of knowledge representation in the existing automatic question answering system. Questions were classified according to their traits. The intention of a question was transformed into the query of basic elements in ontology, and was implemented via Jena statements, and the answer was extracted finally. The proposed approach is proved by experiments.

**Key words:** ontology; automatic question answering system; knowledge representation; question classification; answer extraction

### 0 引言

答疑系统作为一种 QA 系统,丰富了教师与学生之间沟通的交互形式,分为人工答疑和自动答疑两种。前者是通过电子邮件、留言板、聊天室等方式实现,需要人工参与;后者允许用户以自然语言的形式提问,从知识库中自动检索后给出准确的答案,由于其不受时间和空间限制,能实时回答问题,提高了学习的效率,是目前的研究热点<sup>[1-4]</sup>。自动答疑系统能够智能回答问题,首先必须具备对用户提问的语义理解能力,其次需要智能表示课程知识,能被计算机自动处理,同时还必须方便扩展。

自动答疑系统的知识库多采用问句-回答集(Frequently Asked Questions, FAQ)的形式,答疑时系统先计算用户问句与 FAQ 中间句相似度,列出一个或多个相似度最高的问句供用户选择,或直接以相似度最高的问句对应的答案作为回答<sup>[2]</sup>。这种自动答疑系统对已有的问题的回答准确率高,缺点是计算问句相似度时以关键词匹配为主,对问句缺乏语义理解,如果 FAQ 中没有与用户提问相匹配的问句,系统不具备自动回答能力,另外需要事先准备大量问句-答案的集合。

知识库可以采用本体(Ontology)<sup>[3-6]</sup>形式,本体在计算机领域中定义为“本体是概念模型的形式化规范说明”,目标是获取、描述和表示相关领域的知识。文献[3]使用了本体来表示领域知识,但只是用来计算问句的语义相似度;文献[4]设计了基于本体的答疑系统,但关键的问题分类及答案抽取没有说明;文献[5]的银行领域问答系统和文献[6]的旅游信息问答

系统,都展示了本体的知识表示及计算机自动处理能力。

本文提出了一个智能答疑系统实现方案,把课程知识表示在本体中,对用户的问题进行语义分析,将问题模式转换为对本体的检索,并从查询结果中抽取答案来自动回答问题。

### 1 本体知识库

#### 1.1 知识表示

本体是指概念化对象的明确表示和描述。领域本体  $O$  定义为三元组:  $O = (D, W, P)$ , 其中  $D$  是一个领域,  $W$  是领域内所有相关概念的集合,  $P$  是领域空间  $\langle D, W \rangle$  上概念之间关系的集合。概念间语义关系复杂多样,包括上下位关系、从属关系、部分整体关系等,它们之间组成的是一个网状结构。在构建本体时,最重要的是确定概念之间的上下层关系,并以此为基础添加概念之间的横向关系,把独立的各个树联系起来,形成一个连通的语义关系图。

**定义 1** 采用类来对领域本体中的概念、概念的定义、属性等对象进行描述。 $\langle \text{概念} \rangle :: = \text{实体类} \mid \text{概念定义}; \langle \text{Description} \rangle$ ; 属性:  $\langle \text{Attribute\_of, Value} \rangle$ ; 同义关系:  $\{ \langle \text{Synonymy\_of} \rangle \}$ ; 部分整体关系:  $\langle \text{Part\_of} \rangle$ ; 上下位关系:  $\langle \text{Kind\_of} \rangle$ ; 相关关系:  $\langle \text{Relevant\_of} \rangle$ ; 实例:  $\langle \text{Instance\_of} \rangle$

**定义 2** 本体  $O = (C, Relations)$ , 其中  $C$  是  $O$  上的概念集,  $Relations = \{ F, B, I, A, R \}$  表示为概念之间的关系集合。其中: 符号  $F$  代表概念之间的上下位关系,  $F(c_1, c_2)$  表示概念  $c_2$  是概念  $c_1$  的上位概念,  $FF1(c)$  函数返回概念  $c$  的上位概念,  $FF2(c)$  函数返回概念  $c$  的下位概念的集合; 兄弟关系  $B(c_1,$

收稿日期:2009-08-21;修回日期:2009-10-10。

基金项目:广东省科技计划项目(2007B020706006);华南农业大学校长基金资助项目(5600-K08010)。

作者简介:刘汉兴(1971-),男,湖北鄂州人,讲师,硕士,CCF 会员,主要研究方向:语义网、自然语言处理;林旭东(1973-),男,湖南吉首人,讲师,博士,主要研究方向:语义网、Web 挖掘;田绪红(1966-),男,湖北汉川人,教授,博士,主要研究方向:图形图像处理、人工智能。

$c_2$ ),  $BF(c)$  函数返回概念  $c$  的兄弟概念; 实例关系  $I(c_1, c_2)$ ,  $IF(c)$  函数返回概念  $c$  的所有实例; 属性关系  $A(c_1, c_2)$ ,  $AF(c)$  函数返回概念  $c$  的所有属性名称; 相关关系  $R(c_1, c_2)$ ,  $RF(c_1, c_2)$  函数返回概念  $c_1$  与  $c_2$  之间存在的关系: 父子关系(直接相连)、祖先关系(非直接相连的上下位关系)、兄弟关系、实例关系, 或者返回用户定义在  $c_1$  与  $c_2$  之间存在的关系名称等。

## 1.2 课程知识的本体表示

课程知识本体的构建实质就是研究课程知识对象的属性特征和各知识点之间的相互关系, 将它们形式化表示并存储于本体中。以《数据结构》<sup>[7]</sup> 课程为例, 数据结构是指同一数据元素类中各数据元素之间存在的关系, 分为逻辑结构、存储结构(物理结构)和数据的运算。数据的逻辑结构(简称数据结构)是对数据之间关系的描述, 有四类基本结构: 集合、线性结构、树形结构、图状结构。数据结构在计算机中的表示(映像)称为数据的物理(存储)结构, 分为顺序存储和链式存储。数据的运算是在数据的逻辑结构上定义的操作算法, 如检索、插入、删除、更新的排序等, 常见算法有查找算法和排序算法。以算法部分知识点为例, 图1为其中部分内容的层次及相互关系效果图。

用本体表示课程知识点时, 主要是定义知识点所对应的概念及其层次关系。定义概念(类)时, 除概念本身的描述和属性外, 还根据答疑的需要添加其他属性, 例如回答算法类问题时, 用户既需要算法的文字描述, 还需要有算法源代码及演示过程。在构建本体时, 对于算法类概念, 如“冒泡排序”, 除文字定义外, 还添加“源代码”和“示例”属性, 其类型为 string 类型的 `DataTypeProperty`, 属性的值分别为算法源代码及排序过程演示说明。

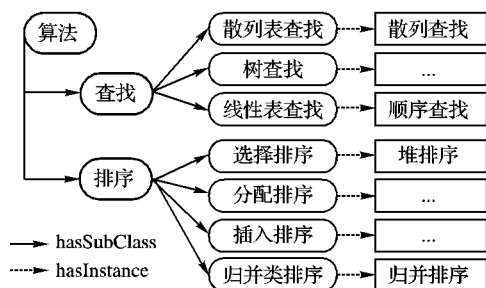


图1 算法知识点表示示意图

在数据结构课程中, 存在大量数学函数关系, 如二叉树有节点和树高的属性, 其数学表述为: 二叉树第  $i$  层上的节点数目最多为  $2^{i-1}$  ( $i \geq 1$ ); 深度为  $k$  的二叉树至多有  $2^k - 1$  个节点 ( $k \geq 1$ )。在回答二叉树相关问题时, 经常需要根据树的深度(或层高)计算节点数。这种数学函数关系在本体中难以表达或者不方便进行运算, 本文将它们定义到一个公共数学函数数据库中, 在回答数值推理问句时使用。如问句“深度为4的满二叉树有多少个节点?”的回答过程就需要进行数值计算, 回答为15。

另外, 课程的知识点还包括一些公理, 如以下推断成立: 满二叉树是完全二叉树, 完全二叉树不一定是满二叉树。这种知识点在本体中可以直接用规则表示。

## 2 系统设计模型

如图2所示, 将课程知识表示成本体形式, 为答疑系统提供知识库功能。如果需要答疑其他课程, 只需要简单地替换课程本体文件, 或者扩充知识库内容构建学科知识库, 包含所有核心课程知识点。事实上答疑的内容通常会包括其他课程

内容, 知识库创建时应考虑整个学科的知识节点。

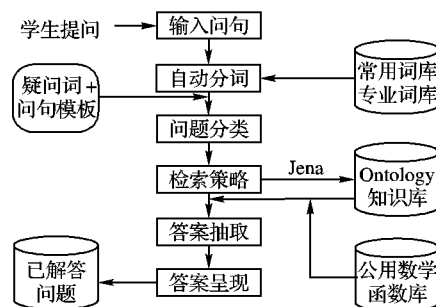


图2 答疑系统模型

学生提问时可用自然语言问句输入问题, 对于较常见的概念解释类问题如“什么是栈”则直接输入概念名称“栈”, 答疑系统会自动识别。如果输入问句, 系统先进行自动分词处理, 分词的词库采用“常用词库+专业词库(本体知识库中的术语)”来提高分词准确率。然后进行基本句法分析, 根据问句模板提取疑问词, 确定问句的中心词及其修饰词, 并进行问题分类。下一步围绕中心词检索知识库, 即将用户的询问意图转化为对本体的 Jena 语句查询, 最终抽取并输出结果。每个已解答问题自动存入“已解答问题库”, 可供学生浏览学习。例如:

问题1 什么是线性表? 疑问词“什么”, 中心词“线性表”, 根据问句模板确定为概念解释类问题, 回答知识库中的定义。

问题2 线性表有哪几种? 疑问词“哪几种”, 中心词“线性表”, 根据问句模板确定为列举类问题, 查询知识库, 回答概念(类)的实例或子类(分类)名称。

问题3 完全二叉树是满二叉树吗? 中心词: “完全二叉树、满二叉树”, 根据问句模板确定为简单回答(是否/有无)类问题, 从中心词之间的关系来回答。

## 3 问题分类

问题分类是指预先定义一个问题类型集合, 找到用户随机提出的问题集合中对应的类型。目前常见的分类为人物、时间、地点、数量、代码、对象6大类, 另外一种 TREC QA<sup>[8]</sup> 的分类方法, 则将问答类型分为不同的任务: 1) Factoid 任务是对基于事实、有简短答案的提问的处理能力; 2) List 任务列出满足条件的几个答案; 3) Definition 任务给出某个概念, 术语或现象的定义、解释; 4) Context 任务表示对相关问题的系列提问的处理能力; 5) Passage 任务只要给出包含答案的一个字符序列; 6) Other 任务, 不包括以上类型的其他任务。

表1 问题分类统计

类别	TREC 分类	数量	百分比/%
概念解释	Definition	201	25.00
简单回答	Factoid	163	20.27
算法	Definition	140	17.41
列举	List	98	12.19
对比分析	Other	87	10.82
数值计算	Factoid	65	8.08
其他	Other	50	6.22

答疑系统主要涉及课程知识点, 学生提问的内容与课程或学科内容有关, 以数据结构课程为例, 通常不会涉及人物时间地点等内容。本文按照 TREC 分类的思想, 对收集到关于

数据结构课程的 804 条问句进行了类别统计,提出了一种答疑系统的问题分类方法,如表 1 所示。其中用户提问较多的三类问题:概念解释、简单回答、算法类,简单回答表示回答时需要根据知识点做出简单判断,如是否/有无等,而较多的算法类问题则反应了数据结构课程的特点。另外对比分析是指两个概念的关系或比较异同问题,简单数值表示需要进行数学计算的问题,其他类表示除了以上分类外的其他问题。

以上问题分类方法初步划分了学生提问的问句类型,但对于用户输入的自然语言问句,还必须明确问题的模板。如常见的概念解释类问题,用户可以直接输入概念名称外,同一个意思可以输入不同的问句:什么是线性表?/线性表是什么?表 2 列出了各问题类型的典型问题模板。

表 2 问题模板

疑问词类	句型模式	问题类型
什么	np + 是 + ~, ~ + 是 + np	概念解释
是否/有无	np + 是不是/有没有 + ~	简单回答
...	np(排序/查找) + ~	算法
哪/哪些	np + 有 + ~, ~ + 是 + np	列举
有何	np + 与/和 + np + ~ + 异同	对比分析
多/多少	np + 是/为 + ~	数值计算

#### 4 检索策略和答案抽取

经过问题分类后,下一步的工作首先是针对本体进行信息检索,即将问题转化为对本体中的基本元素(概念、属性、关系、实例)的查询,本文采用了 SPARQL 语言和 Jena<sup>[9]</sup>对本体进行信息检索,进而抽取答案。设定用户提出的问题为  $Q$ ,系统回答为  $\text{Answer}(Q)$ ,不同类型的问题其检索策略和答案抽取方法不同。

1) 概念解释类。  $\text{Answer}(Q) = \text{Des}(c) + [\text{A}(c)] + [\text{FF2}(c)] + [\text{I}(c)]$ ,其中  $c$  为概念,  $\text{Des}(c)$  表示概念在本体中的定义描述,回答时从问句中中心词的定义进行回答。中心词可能是类、实例、关系、属性中的一种,不同类型其回答不同,如果中心词是类的名称,除了类的定义外,还应回答类所具备的属性、子类、实例。

类名称解释:列出概念定义时给出的描述;列出概念的直接上位概念和下位(子)概念;列出该概念具备的一些属性特征;列出与概念直接相关关系;列举概念的一个实例举例说明。

属性名称解释:在本体中查找概念的属性定义,如无修饰词则列举各自对该属性的定义:属性类型(属性类型),属性描述,属性限制(如类型为整数,取值范围为多少)。

实例名称解释:说明实例所属的概念名称;列举实例所有的属性特征及对应值。

关系名称解释:说明关系的域(Domain)及范围(Range)。

例如问句:什么是数据结构?或数据结构是什么?其回答分为:①定义,数据结构是指同一数据元素类中各数据元素之间存在的关系;②分类,逻辑结构、存储结构(物理结构)和算法。

2) 简单回答类。  $\text{Answer}(Q) = \text{R}(c1, c2)$

对于判断概念  $x$  是否属于(或父子、实例)概念  $y$  等问题,可以通过函数  $\text{RF}(x, y)$  来进行回答。

3) 算法类。  $\text{Answer}(Q) = \text{Des}(c)$ ,从算法的定义、源代码、示例等角度进行回答。

4) 列举问题类。  $\text{Answer}(Q) = [\text{FF2}(c)] + [\text{AF}(c)]$ 。询问概念的分类(子类)和实例等问题,称为列举问题,可以

直接输出函数的检索结果,如  $\text{FF2}(x)$  函数返回概念  $x$  的子类集合,  $\text{AF}(x)$  函数返回实例集合。

例如以下语句可以检索概念名称为  $\text{className}$  的所有子类、属性和实例:

```
OntModel model = ModelFactory.createOntologyModel
    (OntModelSpec.OWL_MEM, null);
OntClass c = model.getOntClass(className);
//下列分别返回所有子类、属性、实例
Iterator is = c.listSubClasses(true);
Iterator ip = c.listDeclaredProperties(true);
Iterator ii = c.listInstances(true);
```

5) 对比分析类。  $\text{Answer}(Q) = \text{R}(c1, c2)$ 。对本体中的两个对象(概念、实例或属性)等进行对比分析,分为一般性对比和特殊对比两种。一般性对比是指分析两者之间存在什么关系或联系,主要是指本体中的基本关系(part-of, kind-of, instance-of, subclass-of等)或层次结构关系(如兄弟、祖先关系等),以及推理关系(利用函数或公理推导出来的关系);而特殊对比是比较两个对象之间有何异同/区别/联系之类的问题。

如图 3 所示,概念  $C_1$  与  $C_2$  为兄弟关系,直接共同父类为  $C_0$ ;概念  $C_0$  与  $C_3$  为祖先关系;概念  $C_1$  与  $C_5$  为共同祖先关系。

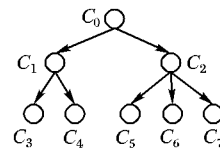


图 3 本体中概念片段示意图

一般性对比问题答案:  $\text{Answer}(Q) = \text{RF}(x, y)$ ,列举对应的定义;判断并列举在本体中直接定义的关系;列举相互间的层次结构关系;列举相互间的推理关系。

特殊对比问题答案:列举各自的定义;列举共同点,即双方都有的属性特征;列举异同点,即一方有而另一方无的属性特征。

如果比较图 3 中概念  $C_4$  与  $C_6$  的异同点,由于双方存在共同祖先  $C_0$ ,则回答时主要从双方的属性特征进行比较说明。

6) 数值计算类。  $\text{Answer}(Q) = \text{Math}(x, y)$ ,是指通过数学函数运算可以得出答案的问题,如二叉树的深度与节点数的二元关系问题。回答此类问题的前提是先制定常见数值计算类问题的问句模板,以及在公共数学函数库中定义函数体,系统工作时首先从问句中提取出用户条件,再调用函数计算结果。

7) 其他类。是指以上六种分类之外的问题。虽然上述六种类型问题覆盖了大部分用户提出的问题,但由于用户问句的复杂度及表达随意性等因素,仍有部分问题难以准确分类和回答,系统回答时主要从概念解释的角度进行回答,或者要求用户重新提问。

#### 5 实验结果

根据上述的设计方案,系统采用 Eclipse + Protégé + Jena 实现,设计了一个针对数据结构课程的本体知识库问答系统,本体使用 OWL 描述语言,用 Protege<sup>[10]</sup> 进行编辑,查询和推理通过 Jena 来实现,图 4 显示了课程本体的部分效果图。

系统测试是面向用户现场测试,由用户随机按自然语言方式提出问句。系统回答时先提示问题的类别,然后输出答案,由用户来判断答案是否正确。如果系统要求重新提问或分类错误,则视为回答不正确。

另外定义性能指标:

- 1) 召回率, 正确分类的问题数占总问题的比率;  
2) 准确率, 正确回答的问题数占总问题的比率;

- 3) 抽取率, 正确回答的问题数占正确分类的问题数的比率。  
实验结果如表3所示。

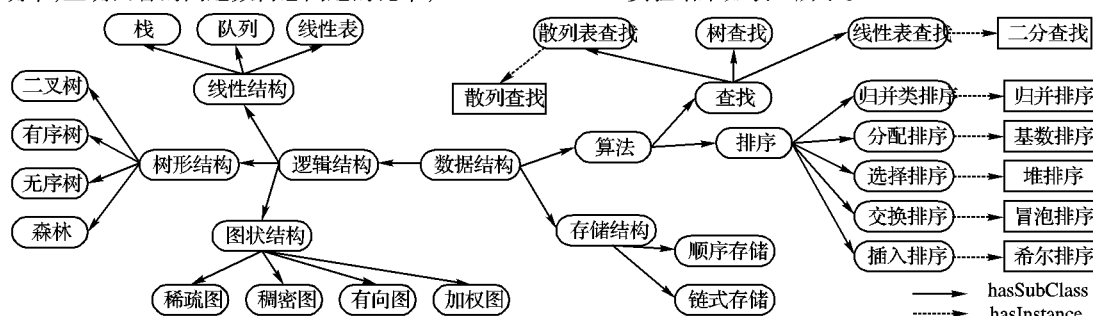


图4 课程本体的部分效果图

表3 问答系统实验结果

问题类型	问题数	召回率/%	准确率/%	抽取率/%
概念解释	115	95.7	87.8	91.8
简单回答	85	91.8	85.9	93.6
算法	45	93.3	84.4	90.5
列举	80	93.8	88.8	94.7
对比分析	65	92.3	86.2	93.3
数值计算	40	92.5	75.0	81.1
其他	20	90.0	60.0	66.7

误差分析如下: 1) 本体知识库模型的完善, 如与课程相关的领域知识的覆盖范围; 2) 问题模板的完善, 对于复杂的问句的识别能力有待提高; 3) 对于表达不规范、没有使用专业术语的问句的处理能力有待提高。其中, 能正确回答的问题主要是概念解释、列举、概念对比分析和简单推理问题。而问题回答错误的主要原因包括: 用词不当、问句依存关系复杂、需要复杂的推理规则等。实验结果表明, 该问答系统能自动进行问题分类和答案抽取, 是因为本体知识库定义了概念之间的层次结构和语义关系, 通过制定相关推理规则和问句模板, 使计算机自动识别问句和抽取出答案成为可能。

## 6 结语

本文设计的基于本体知识库的自动答疑系统, 利用本体的语义特征和层次结构来表达课程的知识点, 结合课程的特点进行问题分类, 并制定了不同的检索策略来生成答案, 能识别并

自动回答问题, 基本满足答疑系统的要求, 但性能仍有待提高。由于本体的构建是一项复杂的系统工程, 除了需要熟悉课程及相关领域知识外, 还要求掌握合理的构建方法, 因此进一步的工作包括: 完善课程领域知识库、完善问题模板增加复杂问句的处理能力、提高问题分类准确性、优化抽取策略等。

## 参考文献:

- [1] 高光来, 王玉峰. 基于智能技术的远程教育答疑系统研究[J]. 中文信息学报, 2003, 17(6): 53-59.
- [2] 郭晓燕, 张博锋, 方爱国, 等. 智能答疑中问题相关度算法研究及系统实现[J]. 计算机应用, 2005, 25(2): 449-452.
- [3] 宗裕朋. 基于本体的中文智能答疑系统研究与实现[D]. 上海: 上海交通大学, 2007.
- [4] 张爱军. 基于本体的智能答疑系统的研究与实现[J]. 计算机应用与软件, 2006, 23(5): 69-71.
- [5] 骆正华, 樊孝忠, 刘林. 本体论在自动问答系统中的应用[J]. 计算机工程与应用, 2005, 41(32): 229-232.
- [6] 李茹, 王文晶, 梁吉业, 等. 基于汉语框架网的旅游信息问答系统设计[J]. 中文信息学报, 2009, 23(2): 34-40.
- [7] 严蔚敏, 吴伟民. 数据结构[M]. 北京: 清华大学出版社, 2007.
- [8] 吴友政, 赵军, 段湘煜, 等. 问答式检索技术及评测研究综述[J]. 中文信息学报, 2005, 19(3): 1-11.
- [9] Jena Semantic Web Framework [EB/OL]. [2009-08-16]. <http://jena.sourceforge.net/>.
- [10] The Protégé Ontology Editor and Knowledge Acquisition System [EB/OL]. [2009-08-16]. <http://protege.stanford.edu/>.

## 中国计算机学会2010年度部分活动计划

4月10—11日: 第三届 Agent 理论与应用学术会议 (Agent2010)

主办: CCF 人工智能与模式识别专委会

承办: 国防科学技术大学计算机学院

联系人: 周会平 E-mail: icent@21cn.com

5月21—23日: 2010 第四届中国可信计算与信息安全学术会议

(CTCIS 2010) 北京

主办: CCF 容错专委会

承办: 北京工业大学和国家信息中心

联系人: 张兴 E-mail: TCS@bjut.edu.cn

6月4—5日: 第五届 IEEE 面向服务的系统工程国际研讨会

(SOSE2010) 南京

主办: IEEE 和 CCF 电子政务专委会

联系人: 白晓颖 E-mail: cengyi@njnu.edu.cn

8月20—22日: 全国第五届 Web 与本体论学术研讨会

(SWON2010) 呼和浩特

主办: CCF 电子政务与办公自动化专委会

联系人: 周敏奇 E-mail: mqzhou@sei.ecnu.edu.cn

8月20—22日: 第七届全国 Web 信息系统及其应用学术会议

(WISA2010) 呼和浩特

主办: CCF 电子政务与办公自动化专委会

联系人: 周敏奇 E-mail: mqzhou@sei.ecnu.edu.cn

8月20日: 全国第三次电子政务技术及应用学术研讨会

(EGTA2010) 呼和浩特

主办: CCF 电子政务与办公自动化专委会

联系人: 邢春晓 E-mail: xingcx@tsinghua.edu.cn

9月16—20日: 第十七届全国网络与数据通信学术会议

(NDCC2010) 北戴河

主办: CCF 网络与数据通信专委会

联系人: 王翠荣 E-mail: ndcc2010@mail.neuq.edu.cn

10月28—30日: 第四届中国传感器网络学术会议

(CWSN2010) 长沙

主办: CCF 传感器网络专委会

联系人: 罗娟 电话: 0731-8882 1907