基于支持向量机的问句分析

刘 颖,韩 杰,滕至阳
(东南大学 计算机科学与工程学院,江苏 南京 210096)

摘 要:为提高问答系统对问句理解的准确率,以概念层次网络理论结合传统计算语言学为思路,提出了适用于限定领域中问句分析模型,根据限定领域的知识特点,设计了新的问句分类方法。在此问句分类方法的基础上,构建了基于支持向量机理论的问句分类器。在以实际教学过程中所收集的真实问句为问题集和训练集的测试中,取得了较好的实践效果。

关键词:概念层次网络理论;问句分类;支持向量机;中文信息处理;问答系统
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2007)08-0001-04

Research of Question Analysis Based on Support Vector Machine

LIU Ying, HAN Jie, TENG Zhi-yang
(School of Computer Sci. and Eng., Southeast University, Nanjing 210096, China)

Abstract: A novel closed-domain oriented question analysis module based on hierarchical network of concepts and traditional computational linguistics is proposed to enhance the rate of accuracy of question interpretation of a question answering system. A new question catalog is developed on the basis of characteristics of closed-domain. A novel question classifier based on support vector machine is constructed on the grounds of this new catalog. The result of experiments tested on questions gathered during process of instruction shows better promise to this method.

Key words: hierarchical network of concepts theory; question catalog; support vector machine; Chinese information processing; question answering system

0 引 言

问答系统(Question Answering System),又称人机对话系统(Human Machine Conversation, HMC),是指系统接受用户以自然语言形式描述的提问,从大量半结构化或者非结构化的数据中,获取能回答此自然语言形式问句的准确、简洁、个性化的答案。这种答案通常是一小段正面回答用户提问的文本,而不是像目前大多数基于关键字串匹配技术的信息检索系统那样返回数以千计的文档链接。

在远程教育中,为了提高网络教学质量,限定专业领域内的智能答疑系统成为研究热点。要实现限定领域内的问答系统,本质上要解决对用户问题的理解和对领域知识文本的理解,从而使系统根据用户具体问题从知识文本中提取相应信息转化为用户需要。

笔者以概念层次网络(Hierarchical Network of Concepts, HNC)理论为指导构建了面向限定领域的问答系统 OSAnsExtr,提出在特定领域中的问句分析模型,注重对限定专业领域内中文问句在语义概念层次上进行分析,抽取出问句中的领域知识和语义信息,提高问答系统的性能。

收稿日期:2006-10-09
基金项目:国家"十五"重大科技攻关项目(2509000012)
作者简介:刘 颖(1975-),男,江苏淮阴人,硕士,研究方向为嵌入式系统、自然语言理解等;滕至阳,教授,研究方向为人工智能、ICAI等。

1 系统介绍

答疑解惑是教学过程中不可缺少的环节。文中以《现代操作系统教程》[1]课程答疑为背景,结合 HNC 理论构建系统总体模型,如图1所示,其特点是能够抽取问句文本中的领域知识和语义信息并据此回答用户提问。系统定位于限定专业领域,一定程度上减少了系统的复杂性。

系统主体部分采用目前比较流行的问答系统模型,即由问句分析、文档检索、句段检索、答案抽取四个模块组成。当用户通过系统接口提交自然语言形式的问句文本时,系统首先调用由领域知识库和 HNC 知识库支持的 HNC 句类分析器对文本预处理,提取领域信息和语义信息以备系统后续部分使用。问句分类部分根据文本预处理中提取的领域信息和语义信息对问句

进行分类,确定问句的答案类型,这部分工作由基于支持向量机理论的问句分类器处理。策略生成器则综合HNC句类分析器和问句分类器所产生的信息生成面向文本搜索的策略和面向答案选择的策略,以指导文档检索和答案抽取工作的进行。文档检索以策略生成器所产生的搜索规则对领域文档文本库中文本信息进行检索、分级,并按照领域主题对文档进行切分,进而对段落分级。句段检索模块在对分级后的段落进行扫描之后以句群为单位生成摘要。答案抽取部分则根据摘要信息和策略生成器所产生的答案选取策略查找候选答案并对其分级,最后选择最佳答案输出。

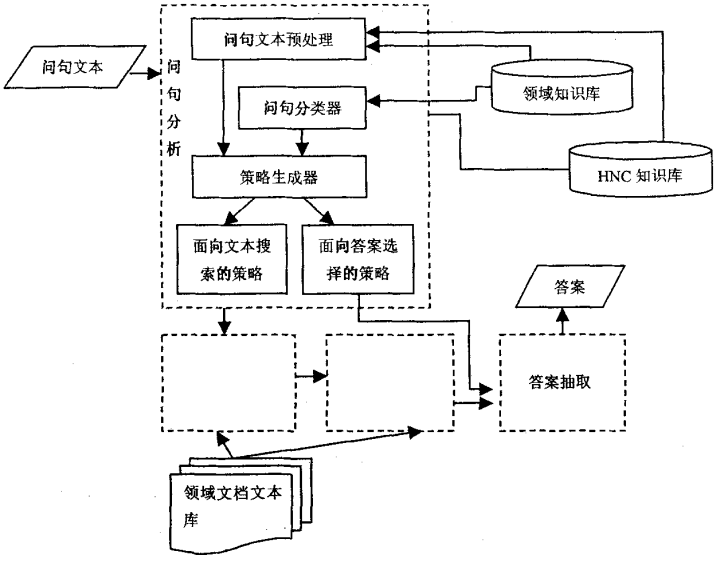

图 1 OSAnsExtr 系统模型示意图

由于篇幅所限,将着重介绍问句分析部分。

2 问句分析

问句分析作为整个问答系统的子模块,其效能直接影响智能问答系统后续处理流程的展开,具有极其重要的作用。目前以词义为基础,与句法规则结合,以句为突破单位的中文信息处理可以分为三大流派[2]。

第一个流派处理思路是以传统计算语言学为基本理论,从词素分析入手,以语料统计为基础,结合语言规则,分析研究词-短语(词组)-语段-句子。

第二个流派是由陆汝占提出基于内涵模型论的语义分析[3]。其特点是将中文信息处理的研究从单纯的语法研究转为深入到语义层面,将汉语表达式抽象成能恰当表示语义内涵和外延义的数学表达式,然后把这些语义表示在计算机内进行处理。

第三个流派是由中国科学院声学研究所黄曾阳创立的面向整个自然语言理解的概念层次网络(HNC)理论[4]。HNC理论以概念化、层次化、网络化的语义表达为基础而得名,其中心目标是建立自然语言的表述和处理模式,使计算机能够模拟人脑的语言感知功能,让计算机理解自然语言,即交互理解。HNC认为,"思维的机制绝不是语法或句法,而是概念联想网络的建立、激活、扩展、浓缩与存储",从而提出计算机对汉语的处理应该以对语言模糊的消解能力为第一标准的理论[5]。

HNC理论从一个全新的角度,开拓了对中文信息处理和汉语语言研究的新领域,找到了一种描述自然语言感知过程的适当模式,实现了信息处理向知识处理的转换。

2.1 问句文本预处理模型

问答系统的目标,就是要根据用户提交的自然语言形式的问句文本,从大量结构化或者半结构化的领域知识文本库中找到满足用户要求的信息并加以处理,最终返回与用户请求相关性最强的一小段文本。为了准确快速检索、抽取答案,必须对用户提交的文本信息进行预处理,提取文本的语义信息和领域知识。

在设计系统 OSAnsExtr 问句文本预处理模块过程中,采用了概念层次网络理论中句类分析的思想设计了限定领域的问句预处理器[6]。

问句分析模块中问句文本预处理器的主要功能是从用户提交的问句文本中获取领域知识、提取语义信息,分别生成可供问句分类器使用的概念层次网络符号和能由策略生成器处理的语义块构成信息。

2.2 问句分类

问句分类的目的是确定答案的语义类别及其搜索分析策略。通过将用户问句文本映射至不同的答案类型来确定答案搜索、答案抽取策略,从而提高系统的准确率。在答案抽取工作中,问句分类通过提供对答案的语义限制来对多个候选答案进行择优选取。

问句分类的结果为问答系统其他部分提供有效的指导信息,提高问答系统的效能,影响系统的准确性[7]。只有确定了问句语义和其对应的答案类型之后,才能应用与其类型相对应的策略分析问句、搜索答案、生成答案[8,9]。

HNC理论能够有效地减少问句分类及问句预处理过程中问句携带信息的丢失,提高准确率。无重要语义信息损失的问句分类方法能够减少问句的复杂

性,并将帮助问答系统在合理的时间内更准确地回答基于领域知识的问题。

目前流行的基于 Open Domain 的问答系统大多数基于简单的问句疑问词或者基于领域知识的关键字进行分类。大量的语义信息在问句处理中发生了丢失,致使系统无法得到语义层次的信息从而导致答案抽取精度不高的情况。而基于领域知识关键字的问句分类并未真正达到对问句的语义层次的理解[10]。

较有影响的分类法是由 Wendy Lehnert 于 1978 年提出的概念分类学[11]。但因其只能反映部分在问答框架中问句类型,从而在实际系统中应用较少。

为了避免上述问句分类的不足,笔者研究分析了在《现代操作系统教程》课程教学过程中收集的 3848 个真实问句,结合该课程中概念较多的特点,以 HNC 理论为指导,提出了自己的问句分类方法。

在构建 OSAnsExtr 系统的问句分类时,遵循了以下几点分类原则:

1)其所设计的类型应是针对特定文本的携带足够问句语义信息的答案类型。

2)其所设计的类型应该能够提供可参考的答案检索方法。

3)充分考虑问句中的隐含信息。

4)自然语言形式的问句能够依照此分类法自动地分类。

依照上述原则对《现代操作系统教程》课程教学中收集的部分问句进行了分类,提出了如下六种分类:CONCEPT(概念类型)、RELATIONSHIP(关系类型)、VERIFICATION(判断类型)、ALGORITHM(算法类型)、COMPUTER(计算类型)、UNKNOWN(未知类型)。

比如"操作系统中为何要引入进程这一概念?","什么是操作系统?"等都属于 CONCEPT(概念类型)。而将"为什么要区分系统和用户态?","原语和系统调用有什么不同?","进程挂起和入睡操作有何异同?"归为 RELATIONSHIP(关系类型)。

将这六个大类进行更加详细的分类,以 CONCEPT(概念类型)为例,将 CONCEPT(概念类型)分为如下的六小类,如图 2 所示,进而将六个大类分为十六个小类型。

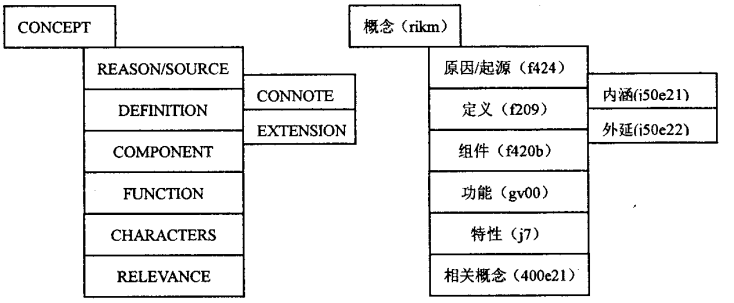

图 2　CONCEPT 概念类型

3　基于支持向量机的问句分类器

目前常用的文本分类器有最小距离分类器、K－最邻近分类器、Naïve Bayes 分类器和支持向量机分类器(Support Vector Machine)[12]。支持向量机(SVM)是 Vapnik 领导的 AT&T 实验室研究小组提出的一种新型通用的有监督的机器学习方法[13]。它建立在结构风险最小化原则基础之上[14],具有很强的学习能力和泛化性能,能够较好地解决小样本、高维数、非线性、局部极小等问题,可以有效地进行分类、回归、密度估计等[15]。由于这些优点,支持向量机已成为机器学习领域的研究热点。目前,支持向量机已经成功应用于三维物体识别、时间序列分析、文本自动分类、遥感图像分析、人脸检测、手写体数字识别,蛋白结构预测等诸多方面。

基于文中在系统 OSAnsExtr 中构建的问句分类方法,即将问句归属到五个大类(CONCEPT, RELATIONSHIP, VERIFICATION, ALGORITHM, COMPUTER, UNKNOWN),十六个小类中。在此基础上,设计了基于支持向量机的问句分类器。

支持向量机理论可以简单描述为:对于数据的二分类问题,寻找一个最优分类超平面,该平面既要保证分类精度还要最大化超平面两侧的空白区域,因为空白区域越大意味着泛化能力越强,即预测性能越好。

设 $\{(x_i, y_i)\}_{i=1}^{l} \subset R^n \times \{1, -1\}$ 是二分类问题的训练样本,根据统计学习理论,其分类超平面定义如下:

$$y_i(w \cdot x_i + b) \geqslant 1 - \xi_i \tag{1}$$

其中 x_i 是超平面上的点,w 是超平面的法向量,b 是偏置量,也称为分类阈值,ξ_i 为松弛变量,表示对错分样本的惩罚程度。

分布在两类分类超平面上的点向量称为支持向量,它们距离最优超平面最近。两分类超平面之间的距离叫做分类间隔(Margin),在两类线性可分的情况下,SVM 通过寻找最优分类超平面实现结构风险最小化,所谓最优分类超平面就是要求分类面不但能将两类正确分开(训练错误率为 0),而且使分类间隔最大。这等同于求一凸二次规划的解:

$$\min\left(\frac{1}{2}\| w \|^2 + C\sum_{i=1}^{l}\xi_i\right) \tag{2}$$

约束条件式为(1);C 是常数,用来控制对错分样本惩罚的程度,实现在错分样本数与模型复杂性之间的折衷。上面的二次规划可以通过其对偶问题解决:

$$\max\left[\sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_j k(x_i \cdot x_j)\right]$$

$$\text{s.t.} \sum_{i=1}^{l} a_i y_i = 0 \qquad (3)$$

$$0 \leqslant a_i \leqslant C,\ i = 1,2,\cdots,l$$

其中,a_i 为拉格朗日乘子(Lagrange multiplier),$k(x_i, y_j)$ 是满足 Mercer 条件的核函数,根据最优理论中的 KKT 条件,只有少量样本的 a_i 值不为零(决策函数值等于 ±1 的样本和错分样本),这些样本就是支持向量。

核形式的判决函数为:

$$f(x) = \text{sgn}\left(\sum_{i=1}^{N} a_i y_i k(x \cdot x_i) + b\right) \qquad (4)$$

$i = 1,2,\cdots,N$ 为支持向量。

常用的核函数有:① 多项式核函数;② 径向基核函数(RBF);③Sigmoid 核函数等,其中 RBF 因其优秀的局部逼近特性在 SVM 中应用最为广泛。文中采用 RBF 核函数。

当用户通过系统接口输入"操作系统中为何要引入进程这一概念?"之后,其将被问句预处理部分处理成如下形式:

操作系统中 /条件辅语义块 C_n、为何要 /特征语义块上装 QE、引入 /特征语义块 E、进程 /作用对象 B、语义块信息 /$(FJ = C_n + QE + E + B)$。这样提取出问句的语义信息之后,即可以将问句预处理文本的处理结果看作词袋模型,因为语义信息已经从问句中抽取出来,所以并没有损失有效的语义信息。

文中根据概念层次网络理论中语义相关度计算的相关理论,设计实现了基于支持向量机的问句分类器。在文本预处理之后,将其处理结果 ——C_n,QE,E,B 等属性值作为问句分类器的输入值。其后结合文献[16]中设计的算法,计算问句文本的类别并输出。

4 结 论

通过将概念层次网络理论与传统计算语言学相结合,设计了限定领域问句分类方法,从理解问句的角度来进行问句处理,构建了基于支持向量机理论的问句分类器,使得系统能更好地理解问句,对问答系统的效率提高起到很重要的作用。在以实际教学过程中所收集的真实问句为问题集和训练集的测试中,分类准确率达到 77.88%,如表 1 所示,取得了较好的实践效果。

表 1 基于 HNC 理论的支持向量机问句分类器测试

训练集数目	测试集数目	分类准确数目	准确率	召回率
2000	1848	1401	77.88%	75.81%

概念层次网络(HNC)理论创立了基于语义的自然语言表述和处理的科学模式。笔者把 HNC 应用到问答系统的问句分析模块中,主要从两个方面提高了系统的性能。第一,在对用户文本问句进行理解的问句预处理过程中,利用领域知识和 HNC 知识库,将用户文本问句转化为领域知识和语义信息,为后续模块准确理解提供了前提;第二是在问句分类的过程中,对问句预处理中所抽取的属性值,尤其是语义信息进行 HNC 语义相关度计算,确保了系统对用户需求的准确把握,提高了系统的精度。

在构造问句预处理器和问句分类器的过程中,发现还有以下问题需要解决:首先,构建基于 HNC 理论限定领域知识库的方法还需要进一步的研究,由于 HNC 概念层次符号都是人工填写的,不仅需要深厚的语言功底而且需要对领域知识中知识点之间的关系的理解;其次,在进行 HNC 语义相似度计算时,相似度权重的选择标准也需要进一步讨论。

参考文献:

[1] 滕至阳.现代操作系统教程[M]. 北京:高等教育出版社,2000.

[2] 许嘉璐.现状和设想——试论中文信息处理与现代汉语研究[J].中文信息学报,2001,15(2):3-6.

[3] 陆汝占,靳光瑾.现代汉语研究的新视角[J].语言文字应用,2004(2):13-17.

[4] 黄曾阳.语言概念空间的基本定理和数学物理表示式[M].北京:海洋出版社,2004:5-10.

[5] 苗传江.HNC(概念层次网络)理论导论[M].北京:清华大学出版社,2005:73-76.

[6] 晋耀红.HNC(概念层次网络)语言理解技术及其应用[M].北京:科学出版社,2006.

[7] Kwok C, Etzioni O, Weld D S. Scaling Question Answering to the Web[J]. ACM Transactions on Information Systems (TOIS),2001,19(3):242-262.

[8] Li Xin, Roth D. Learning Question Classifiers[C]// The 19th International Conference on Computational Linguistics. COLING'02. Taipei, Taiwan:[s.n.],2002:556-562.

[9] Hermjakob U. Parsing and Question Classification for Question[C]// Proceedings of the ACL Workshop on Question Answering. Toulouse, France:[s.n.], 1999:17-22.

[10] Day Min-Yuh, Lee Cheng-Wei, Wu Shih-Hung, et al. An Integrated Knowledge-based and Machine Learning Approach for Chinese Question Classification[C]// Proceedings

(下转第 22 页)

表 1 样本纹理特征

样本 \ 规则		R_1	R_2	R_3
		{1,7}=>{4}	{2,8}=>{5}	{3,9}=>{6}
	支持度	0.422194	0.419643	0.400510
	置信度	0.887399	0.889189	0.889518
	支持度	0.441327	0.447704	0.437500
	置信度	0.895954	0.897436	0.903790

这里对规则的含义解释如下:例如规则 R_1:{1,7}=>{4},表示在 3×3 邻域(如图 6 所示)中的 1 号、7 号位置的灰度概念为"亮",可以推断出 4 号位置的灰度概念为"亮",其中在样本的所有根像元(阴影所示部分)中满足该规则的百分比为 42.2194%,该推断的置信度为 88.7399%。

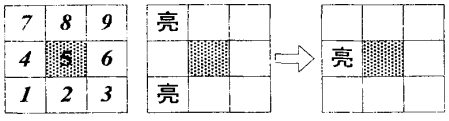

图 6 规则的解释

(2)将待分割的图像采用同样的方法进行亮度概念提升,结果如图 5 (b)所示。

(3)对于图 5(a)的每个样本大小的区域,分析其对应每条规则的置信度 S_i 与支持度 C_i,比较该结果与模式向量的距离 D,按照极大判定原则对根像元作分割标记;$D = \sqrt{(S_i - S_{ij})^2 + (C_i - C_{ij})^2}$。

(4)按标记对图像进行分割,结果如图 5(c)所示。

4 结束语

RSImageMiner 是一个基于(遥感)图像数据挖掘的空间数据挖掘工具,图像纹理特征数据挖掘模块 ITDM 是其重要模块,采用概念驱动的方式,将云模型与概念格结合起来,对图像纹理概念的粒度进行不确定性提升,从中提取有价值的语义特征,指导纹理分析与处理。实验表明,该软件原型能得到较满意的结果。但是,文中实现的模块只是一个软件原型,需要进一步地修改完善,使其能更好地融合 RSImageMiner,使 RSImageMiner 成为一个具有自主知识产权的、实用的空间数据挖掘软件。

参考文献:

[1] Fayyad U M, Weir N, Djorgovski S G. SKICAT: A Machine Learning System for Automated Cataloging of Large Scale Sky Surveys[C]//International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1993: 112 - 119.

[2] Zaïane O R. The First International Workshop on Multimedia Data Mining (MDM/KDD 2000)[EB/OL]. 2006 - 10 - 20. http://www.cs.ualberta.ca/~zaiane/mdm_kdd2000/.

[3] Rushing J A. Image segmentation using association rule features[J]. IEEE Transaction on Image Processing, 2002, 11(5):558 - 567.

[4] Zaïane O R, Han Jiawei. MultiMediaMiner: A system Prototype for MultiMedia Data Mining[C]//Proceedings of ACM SIGMOD International Conference on Management of Data. Seattle: ACM Press, 1998:581 - 583.

[5] 秦 昆.基于形式概念分析的图像数据挖掘研究[D].武汉:武汉大学,2004.

[6] Datcu M, Seidel K. An Innovative Concept for Image Information Mining[C]// MDM/KDD 2002: International Workshop on Multimedia Data Mining (with ACM SIGKDD 2002). Edmonton: University of Alberta, 2002:11 - 18.

[7] 张 钹,张 铃.问题求解理论及应用[M].北京:清华大学出版社,1990.

[8] Ganter B, Wille R. Formal Concept Analysis - Mathematical Foundations[M]. Berlin: Springer Verlag, 1999.

[9] 朱 明.数据挖掘[M].合肥:中国科技大学出版社,2002.

[10] Han Jiawei, Kamber M.数据挖掘概念与技术[M].范 明,孟小峰等译.北京:机械工业出版,2001.

[11] 李德毅,杜 鹢.不确定性人工智能[M].北京:国防工业出版社,2005.

(上接第 4 页)

of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP - KE 2005). Beijing: China Scientific Book Service Co. Ltd, 2005:632 - 640.

[11] Burger J. Issues, tasks and program structures to roadmap research in Question Answering[C]// In Proceedings of The Ninth Text Retrieval Conference (TERC 2000). TERC - 9. Gaithersburg, Maryland: NIST Special Publication, 2000: 500 - 521.

[12] Cristianini N, Shawe - Taylor J. An Introduction to Support Vector Machines and Other Kernel - based Learning Methods[M].北京:机械工业出版社,2005.

[13] Vapnik V. The Nature of Statistical Learning Theory[M]. [s.l.]:Springer - Verlag, 1995.

[14] Amari S, Wu S. Improving support machine classifier by modifying kernel Function[J]. Neural Networks, 1999, 12: 783 - 789.

[15] Burges C J C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2):121 - 167.

[16] 张运良,张 全.基于 HNC 理论的语义相关度计算方法[J].计算机工程与应用,2005(34):1 - 3.