支持向量机算法在 MOOC 课程答疑系统中的研究①

岳群琴, 景 红

(西南交通大学 信息科学与技术学院, 成都 610031)

摘 要: 随着互联网技术和近期 MOOC 课程的发展, 智能答疑系统也受到了更多的关注, 应用它能够及时给学生提供学生疑惑的问题答案. 智能答疑系统通常包括问句理解、信息检索、答案抽取和选择三个主要部分, 且问句分类是问句理解的关键, 因为它的准确性将直接影响到最后答案的准确性. 以高校计算机基础课程为实际背景, 在已有基于支持向量机算法基础上, 对该方法进行了改进, 并通过训练集和测试集进行了验证. 从实验结果看, 该方法在高校计算机基础智能答疑系统中有比较好的应用效果.

关键词: MOOC; 智能答疑; 问句理解; 问句分类; 支持向量机

Study on MOOC Intelligent Answering Using Support Vector Machine

YUE Qun-Qin, JING Hong

(School of Information Science & Technology, Southwest Jiaotong University, Chengdu 610031, China)

Abstract: With the development of Internet technology and the MOOC Courses, the intelligent answering also has drown more attention, because it can solve user unsure questions timely. Intelligent answering system typically includes comprehension questions, information retrieval and answer extraction and selection. Question classification is a part of comprehension questions, which directly affects the accuracy of the final answers. It is verified by the training and test sets using the improving method which is based on support vector machine. The results show this method get high classification accuracy on intelligent answering.

Key words: MOOC; intelligent answering; comprehension questions; question classification; support vector machine

简单地讲, 智能答疑系统就是对自然语言进行处理, 然后选取满足用户要求的答案返回给用户. 它是一个典型的问答系统, 包含问句理解、信息检索、答案抽取和选择这三部分. 其中, 问句分析包括分词和词性标注、关键词特征提取、词性扩展、和问句类别等, 如图 1 所示; 在信息检索过程中, 一般对问题库中的问题采用基于关键词建立索引, 然后根据关键字进行问题相似度计算, 返回给用户候选答案; 而本系统中采用基于问句类型建立索引方式, 根据对用户输入问句的分析, 在其索引中进行问题的相似度计算, 将候选答案返回给用户. 通过从问句类型中搜索答案, 能够极大提高问题检索效率, 能够加快搜索结果的返回速度.

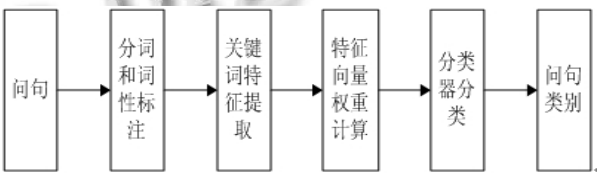


图 1 问句分析

目前, 基于中文问句分类算法一般使用基于机器学习的方法, 主要通过提取问句的特征, 然后采用相应的算法进行问句分类计算, 根据一定的规则来判断问句的类型. 本文以高校计算机基础课程为实际背景, 主要根据课程中常见问题集合的类别属性, 采用哈工大提供的句法依存分析器来实现对问句的理解过程,

① 基金项目:中央高校基本科研基金(A092050205130425)
收稿时间:2013-12-24;收到修改稿时间:2014-02-21

并根据特征提取算法对依存分析结果提取出问句的特征向量, 最后采用基于向量机的方法对问句实现分类.

1 问题分类

汉语问句的类型有很多, 通常多为根据问句一些特征进行分类. 本文针对计算机基础课程中常见的问题, 并结合汉语分类的基本方法将问句类型分为表 1 所示的基本类型[4]. 其中课程中主要涉及的问题类型为描述类型, 其他类型问题涉及较少.

表 1　问句类型

一级	二级
人物	特定人物, 团体机构, 人物描述, 人物举例
时间	日期, 时间, 时间范围
描述	简写, 定义, 意义, 方法, 原因, 描述, 判断, 描述其他
地点	城市, 国家, 地区, 地址等
其他	其他未包含的类别

问句分类的目的主要是使问答系统能够有效的减少搜索范围和搜索时间成本, 从而提高系统返回答案的准确率. 例如, 当用户输入“计算机病毒的定义是什么?”时, 这个问题属于描述大类中的定义类, 在检索文本时能把目标锁定在定义类别中进行搜索, 减少搜索空间.

2 SVM问句分类

一般对于中文文本而言, 因其含有上下文关联等大量的信息, 在进行相似度计算时, 需要提取所有的关键字进行计算; 而对于一般问句, 所含有的信息较少, 一般为十几个字. 所以在进行问句分类时, 可以将问句中的所有词作为关键字, 然后采用相应的算法进行问句分类训练[5]. 但是, 问句中含有一些无关关键字和噪音字, 其中的噪音词会影响问句分类的准确率. 如何提取其中的关键字作为分类向量, 从而提高分类的准确率成为本文研究的重点. 本人采用哈工大的语句依存工具分析的结果来提取关键字作为分类向量来进行问句分类.

2.1 语句依存分析

句法分析是指在给定文法下分析自然语言的层次结构, 它是自然语言处理中的中心问题之一, 并且在很多领域中有重要的应用. 依存语言通过分析语言单位内成分之间的依存关系揭示其句法结构, 主张句子中动词是支配其他成分的中心成分, 而它本身却不受其他任何成分的支配, 所有受支配成分都可以以某种依存关系从属于支配者[3]. 依存文法的句法结构的主要元素之间依存关系, 即句中词对的二元关系, 其中一个记为核心词(head), 另一个记为依存词(dependent). 依存关系反应了核心词和依存词之间的依赖关系. 如果两个词之间有弧相连, 表示两者之间存在依存关系, 弧发起的词(依存词)依存于弧指向的词(核心词).

本文主要采用哈工大的依存工具进行句法分析, 通过该工具来提取句中的核心词和依存关系, 例如, “计算机病毒的定义是什么?”通过该工具分析的结果如图 2 所示, 分析结果如图所示, 核心词 hed 为“是”, 主语 SBV 为“定义”, 宾语 VOB 为“什么”, 主宾和核心词之间都存在依存关系, 通过核心词就能很快的把句子的主干查找出来.

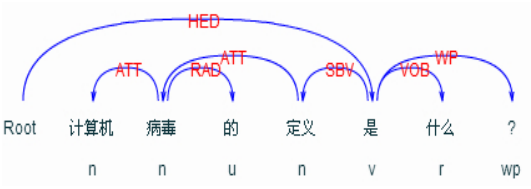


图 2　句法分析结果

2.2 特征词提取算法

一般对于一个问句而言, 简单的十几个字, 通过句法分析器很快就能将句子的主干提取出来. 在问句分类中, 一般通过一个简短的疑问词就能判断出该问句属于哪一类别, 比如 “为什么”, 很快就能判断出该问题输入描述类中的原因类[2]. 但是, 并不是所有的问句通过疑问词就能判断其类别的, 比如 “什么时间”, 通过疑问词 “什么”并不能很快的判断其类别, 必须词“时间”搭配才能判断该问题属于时间类, 而 “时间”这个词就属于疑问附属词. 本文主要通过该工具查找出句中的主干, 疑问词和疑问附属词这五元组, 标记为 Q={hed, sbv, vob, qwd, rel_qwd}, 其中 qwd 为疑问词, rel_qwd 为疑问附属词. 然后通过问句的 Q 向量来查找问句的类型. 对于问句中的疑问词而言, 可以作为句中的主语、宾语, 而且并不是所有的问句都有疑问附属词, 所以五元组中的疑问附属词和疑问词可选.

单纯使用依存分析工具提取句子的主干时, 由于该工具分词有些效果不好, 比如“称为什么”, 在使用该工具进行分词时分为“为什么”, 很容易理解为描述类中的原因类问题, 而实际分词结果分为“称为”,

“什么”, 这样会造成在问句分类中有比较大的误差. 所以本文首先使用中科院的 ICTCLAS2013 分词工具和计算机基础专业术语词库进行分词, 在分词的过程中, 将问句中含有引号的部分进行处理, 便于后续主干提取. 因为在句法依存分析的过程中, 如果句中含有引号部分词汇时, 引号会同核心词存在 SBV 或 VOB 的依存关系, 这样在提取主干过程中直接提取 SBV 或 VOB 会出错. 分词完成后, 将分词的结果和依存分析工具结合进行句法分析, 然后在提取出句子的五元组, 其算法过程如图 3 所示, 其中 COO 表示并列关系. 同时, 问句中可能存在多个疑问词, 但是每个问句中只可能有一个疑问词, 其他词可能只是作为修饰部分, 所以这时需要根据句法分析找出词性为“r”的词作为疑问词, 便于后续疑问附属词提取.

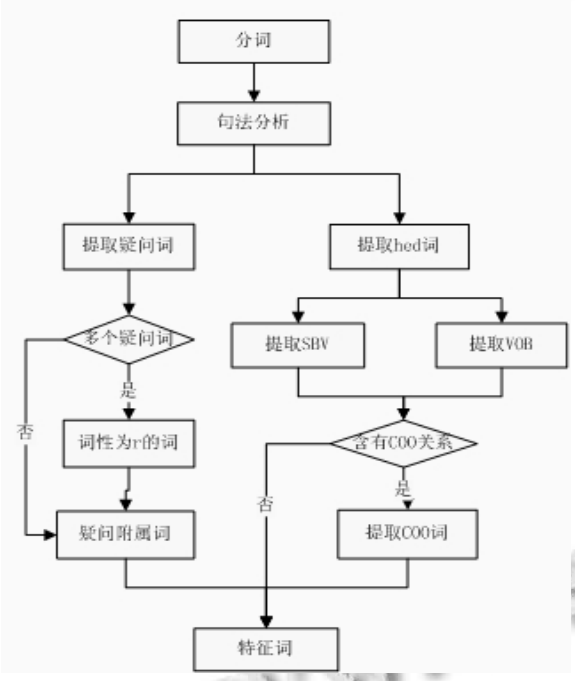


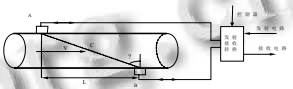
图 3 特征词提取流程

2.3 支持向量机算法

SVM[1]是基于统计学习理论和结构风险最小化原则的一种机器学习方法, 它可以针对线性可分情况进行分析, 对线性不可分的情况, 通过基于结构风险最小化原理之上在特征空间中构建最优分割超平面, 使得学习器得到全局最优化, 并且在整个样本空间的期望风险以某个概率满足一定上界. 支持向量机主要是通过构造一个目标函数将两类模式尽可能分开. SVM 可以解决小样本情况下的机器学习问题, 而对于本文的中文问题分类, 由于样本数目有限, 属于小样本问题, 通过合理的选择特征向量, 使用 SVM 算法能够很好的解决.

设给定的训练集为 Q, 对于给定的函数,y=wx+b 其中 x、w 为向量, 且 $x_i \in X=Rn$、$y_i \in Y=\{+1,-1\}$,且该函数可以被一个超进行线性分割, 记该超平面为: Q={(x1,y1),(x2,y2),…,(xn,yn)},wx+b=0

在两类问题分类中, 求解最优超平面问题等价于求解如下的二次规划问题:



$$\text{s.t.,}\ \ y_i(wx_i+b)\geq 1-\varepsilon_i,\varepsilon_i\geq 0$$

其中 C 为惩罚因子, 它决定了离群点带来的损失. ||w|| 是 W 的二范数. 为了解决凸规划问题, 使用了 Karush Kuhn Tucker(KKT)条件的理论, 这里直接给出了拉格朗日目标函数:

$$P(w,b,\varepsilon,\alpha,\mu)=\sum_{i=1}^{n}\mu_i\varepsilon_i+\sum_{i=1}^{n}\alpha_i\left[y_i\left(w^Tx_i+b\right)-1+\varepsilon_i\right]$$

$$L(w,b,\varepsilon,\alpha,\mu)=\frac{1}{2}\|w\|+C\sum_{i=1}^{n}\varepsilon_i-P(w,b,\varepsilon,\alpha,\mu)$$

求解该表达式首先让 L 关于 w、b 最小化, 分别令 L 关于 w、b、ε 的偏导数为 0, 得到关于问题的表达式:

$$\frac{\partial L}{\partial w}=0\Rightarrow w=\sum_{i=1}^{n}\alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b}=0\Rightarrow \sum_{i=1}^{n}\alpha_i y_i=0$$

$$\frac{\partial L}{\partial \varepsilon_i}=0\Rightarrow C-\mu_i-\alpha_i=0$$

将表达式进行转换后, 得到如下的函数:

$$y=wx+b=\sum_{i=1}^{n}\alpha_i y_i x_i^T x+b$$

对于 SVM 是训练样本集和核函数完全描述. 因此采用不同的核函数就可以构造实现输入空间中不同类型的非线性决策面的学习机, 导致不同的支持向量算法. 通常将线性的空间映射到高维的空间核函数有多重, 将表达式中进行部分改造, 令

$$x_i^T x_j=K(x_i,x_j)$$

式中即为核函数, 在实际问题中, 比较典型的核函数有(1),(2). 在实际的问题中, 通常是直接给出核函数, 选择高斯核函数(2)为支持向量机的核函数.

$$K(x_i, x_j) = (x_i x_j + 1)^d \quad (1)$$

$$K(x_i, x_j) = \exp(-\frac{(x_i - x_j)^2}{2\sigma^2}) \quad (2)$$

上面所谈到的分类都是 2 分类的情况, 对于问题分类而言属于 N 分类的情况, 主要有两种方式[6]:

一种是方式需要训练 N 个分类器, 第 i 个分类器是看看是属于分类 i 还是属于分类 i 的补集(除去 i 的 N-1 个分类);

另一种方式需要训练N*(N-1)/2个分类器, 分类器(i,j)能够判定某个点属于i还是属于j. 本文主要使用后一种方法, 对于 N 个分类的训练集进行两两区分.

3 实验

本文中对于 SVM 算法训练和测试中的问题分类并没有采用层次分类的方法(即先分大类, 再分大类下的小类), 而是直接把问题在小类中进行分类. 采用这种分类方法主要是因为层次分类的效果并不是特别理想, 同时本训练集主要是针对计算机基础课程中常见的问题集, 该问题集中主要涉及是描述类, 时间类和人物类相关问题, 对于其他如地点类相关问题比较少.

对于问句分类的评价, 一般采用准确性进行评价. 本文也主要采用分类的准确性对分类结果进行评测. 其公式为:

本文首先采用本文的特征提取方法对训练集进行训练, 然后使用同样的方法对测试集进行测试, 其中训练集的数目为 925 个, 测试集的数目为 325 个, 其中每个类别的具体分布如表 2 所示:

表 2　问题训练和测试集表

	简称	定义	判断	描述	原因	特点	时间	人物	意义	方法	其他
训练	90	90	85	90	75	75	80	80	80	80	100
测试	34	35	30	32	25	25	25	26	25	26	42

通过设置不同的惩罚因子C的参数, 使用SVM方法得到的准确率如图 4 所示, 当 C=16 时, 得到的准确率为 89.8438%, 而测试集采用传统的以词作为特征的方法其准确率为 83.3259%, 比其高出了 6 个百分点.

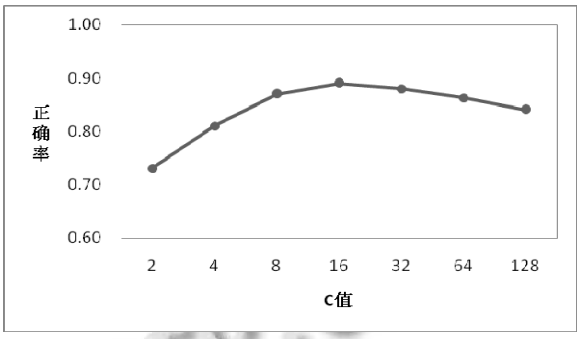


图 4　不同惩罚因子 C 下正确率

4 结语

本文研究了句法依存以及使用句法依存提取特征向量的算法, 并使用 SVM 分类算法进行分类, 与直接使用SVM分类相比提高了分类效率, 同时减少了特征向量的维数. 但是在分类过程中, 由于分词器在对于一些词的分词有些误差, 同时句法依存对于一些词的句义分析误差, 都会造成最后的分类的准确率. 因此, 在以后的研究中, 需要进一步解决这些问题带来的误差, 才能进一步提高其准确率.

参考文献

1 Zhang D, Lee WS. Question classification using support vector machines. The 26th ACM SIGIR. 2003.

2 Li X, Roth D. The role of semantic information in learning question classifiers. First International Joint Conference on Natural Language Processing. 2004. 451–458.

3 文勖,张宇,刘挺,马金山.基于句法结构分析的中文问题分类.中文信息学报,2006,20(2):33–39.

4 牛彦清,陈俊杰,段利国,张巍.中文问句分类特征的研究.计算机应用与软件,2012,29(3):108–111.

5 孙景广.基于知网的中文问题自动分类.中文信息学报,2007,21(1):90–95.

6 田卫东,高艳影,祖永亮.基于自学习规则和改进贝叶斯结合的问题分类.计算机应用研究,2010,27(8):2869–2872.