文章编号：1003-0077（2011）00-0000-00

基于答案辅助的半监督问题分类方法 *

张栋 李寿山 周国栋

苏州大学自然语言处理实验室 苏州 215006

E-mail: d.zhang.suda@gmail.com, {lishoushan, gdzhou}@suda.edu.cn

摘　要：问题分类旨在对问题的类型进行自动分类，该任务是问答系统研究的一项基本任务。本文提出了一种基于答案辅助的半监督问题分类方法。首先，将答案特征结合问题特征一起实现样本表示；然后，利用标签传播方法对已标注问题训练分类器，自动标注未标注问题的类别；最后，将初始标注的问题和自动标注的问题合并作为训练样本，利用最大熵模型对问题的测试文本进行分类。实验结果表明，本文提出的基于答案辅助的半监督分类方法能够充分利用未标注样本提升性能，明显优于其他的基准方法。

关键词：问答系统；问题分类；答案辅助；半监督分类；标签传播

中图分类号：TP391　　**文献标识码：**A

Semi-supervised Question Classification with Answers

Dong Zhang, Shoushan Li, Guodong Zhou

Natural Language Processing Lab of Soochow University, Suzhou, 215006

E-mail: d.zhang.suda@gmail.com, {lishoushan, gdzhou}@suda.edu.cn

Abstract：Question classification aims at classifying the types of questions automatically, and this task is a basic task of question answering system. This paper proposes a method of semi-supervised question classification with answers. Firstly, answer features bind question features to realize samples said; Secondly, a question classifier will be trained on labeled questions using label propagation algorithm to annotate the category of unlabeled questions automatically. Finally, the questions of initial annotation and automatic annotation are merged with each other as training samples, using maximum entropy model to classify the testing samples. The experimental result demonstrates that the method of semi-supervised question classification with answers in this paper can make full use of the unlabeled samples to improve the performance, and is better than other benchmark methods.

Keywords: Question answering system; Question classification; Answer aiding; Semi-supervised classification; label propagation

1 引言

问答系统能够为用户提出的自然语言问题提供一个简明、准确的答案，越来越受到人们的关注。现有的问答系统主要包括三个模块：问题分析、信息检索和答案抽取。问答系统为了能够正确回答用户所提出的问题，首先需要对问题进行分析，知道用户想要寻找什么信息。此时，问题分类作为问题分析最基础的任务，提供了重要支持[1]。

问题分类就是把给定的某个问题映射到多个类型中的某一个或者几个类别中，以确定问题的类型。问题分类的第一个作用是有效的减小答案的候选空间。如：“怎么学习电脑维修呢？”经过问题分类，该问题是一个“电脑”类的问题，问答系统就可以把这个问题的候选答案限制在“电脑”类的相关答案集合中。这样就非常有效地减少了候选答案集合，充分提高了检索效率。

问题分类的第二个作用是能够决定答案的抽取策略，根据问题的不同类别采用不同的答案选择策略和知识库。如：“水瓶座男对天蝎座女表白说什么话最好？”经过问题分类，该问题是“感情”类问题，检索这类问题的答案就需要利用情感分析技术。

问题分类可以看作一种特殊的文本分类，然而，问题分类与传统的文本分类存在一定差别。一方面，在传统文本分类中词频信息对于区分文本中每个词汇的贡献程度很大，但在问题分类中词频信息不具明显区分作用，因为问题通常比较短，问题中每个词汇的词频普遍为1[2]。单单利用问题进行分类，往往由于信息量少而分错；另一方面，已标注的问题资源比

* **收稿日期：**2015-06-15　　**定稿日期：**2015-08-10

基金项目：国家自然科学基金重点项目（61331011）；国家自然科学基金（61375073 和 61273320）

较匮乏，标注语料又需要大量的时间、人力和物力[3]。因此，这就需要我们加入更多的辅助特征扩充问题信息，同时充分利用大量的未标注样本信息，才能获得较高的分类精度。

此外，传统的问题分类普遍都是基于全监督的分类方法，并且仅仅从问题中抽取特征进行分类。与以往研究不同的是，本文提出的基于答案辅助的半监督问题分类方法，一方面，该方法能够充分利用问题已有的答案来扩充分类信息，解决上面提到的问题包含的词汇信息量少的难点。如表 1 所示，在未利用答案特征的情况下，直接利用问题特征进行分类，“360 问答开放平台是做什么的?”被误分为“非电脑”类问题。其原因可能是该问题中没有包含“电脑”、“网络”等关键词；然而加入了答案特征后再进行分类，该问题则被准确识别为“电脑”类问题。因为答案特征中包含“互联网”这类明显地与“电脑”类相关的关键词。

表 1 利用答案信息辅助分类实例

Table 1 Examples of question classification with answers

	正确类别——电脑	未利用答案特征分类为	利用答案特征分类为
问题	360 问答开放平台是做什么的?	非电脑类	电脑类
答案	360 问答企业平台是开放、简单、高效、精准的互联网企业服务平台……		

另一方面，该方法是一种半监督学习方法，能够充分利用未标注问题的信息提升分类性能，解决标注语料匮乏的问题[4]。此外，本文首次在问题分类研究中引入基于标签传播的半监督学习方法，该方法既可以使问题之间的标签互相传播，也可以使答案之间的标签互相传播，可以有效地提升标签预测准确率。

具体而言，本文的方法先将答案特征加入到问题中，利用标签传播方法预测未标注问题的类别；再将已确定类别的问题作为训练样本，利用最大熵模型进行问题分类。实验结果表明，答案的信息有助于大幅提升问题分类准确率。

本文其他部分组织如下：第二节介绍问题分类的相关工作；第三节描述问答语料的收集和构成；第四节介绍本文提出的基于半监督的问题分类方法；第五节给出实验设置与结果分析；第六节简述结论及下一步工作展望

2 相关工作

目前，问题分类研究主要集中在基于统计的机器学习方法上面。Ray[5]等人充分利用WordNet强大的语义特征和维基百科存储的大量相关知识来扩充问题所蕴含的信息，从而提升问题分类性能；Hui[6]等人考虑了问题文本中词序和词间距对问题分类的影响，提出一种扩展类顺序规则模型；Mishra[7]等人从问题文本中抽取出词特征、句法特征、语义特征，融合这些特征训练三种分类器：最近邻、朴素贝叶斯、支持向量机，进行问题分类；Yadav[8]等人使用了一元、二元、三元词特征以及词性特征，采用朴素贝叶斯分类方法进行问题分类的研究；LIMSI-CNRS[9]等将一部分英语问题语料库翻译为法语问题语料库，问题采用传统的6个大类别以及细分的50个小类别，并使用LibSVM分类器分类这些问题。

田卫东[10]等根据对中文问题的分析，得出问题中的疑问词和中心词等关键词对问题所属类型起着决定性的作用。提出利用自学习方法建立疑问词-类别和疑问词+中心词-类别两种规则，并结合改进贝叶斯模型的问题分类方法。该方法充分利用了关键词对分类的贡献。

刘小明[11]等先对问题进行浅层语义分析，再根据预定义的问题焦点结构和焦点抽取规则，获取问题焦点语义特征，然后标示问题的类别为问题焦点中疑问对象在领域本体中的标识，最后根据焦点不同则问题不同的事实，将焦点相同的问题归为一类。

张巍[12]等针对中文问题分类方法中布尔模型提取特征信息损失较大的问题，提出了一种新的特征权重计算方法。在提取问题特征时，通过把信息熵算法和医院本体概念模型结合在一起，进行问题的特征模型计算，在此基础上使用支持向量机方法进行中文问题分类。

Liu[13]等人认为标准核函数的SVM方法忽视了中文问题的结构信息，因而提出一种问题

文本属性核函数的SMO方法，该方法还同时使用了句法依赖关系和词性特征。

多年来，传统的问题分类研究仅仅着眼于使用各种全监督技术只针对问题本身进行操作。与之不同的是，本文采用半监督学习方法，一方面减小了人工标注的工作量，另一方面通过将答案特征加入问题训练集中来扩充问题的信息，协助问题进行分类，以此提升问题分类的准确率。

3 语料收集与描述

本文语料来自好搜问答社区 [1]，其大类别总共有 15 个。本文为了实验方便直观以及下一步工作的需要，抓取了其中六个类别的问答数据，分别是：电脑/网络、文化/艺术、健康/医疗、生活、感情/家庭、体育/运动。每个类别包含 2000 条问答（每个问题对应一个答案），共 12000 条问答，问答实例如表 2 所示。

表 2 各类别问题与答案实例

Table 2 Question and answer examples of every category

类别	问题	答案
电脑/网络	没有 dns 网关 可以做桥接吗？	不可以的，需要提供 dns 网关。
文化/艺术	《当祖国召唤的时候》写于什么年代？	上世纪八十年代中期. 战争年代
健康/医疗	有治头发早白的偏方吗？	吃何首乌啊~
生活	亮艾补水面膜贴完之后要洗脸吗?	一定要洗脸。
感情/家庭	爱情是什么？爱情的定义？	两情相悦，仅此而已
体育/运动	冬奥会是不是指冬天的奥运会？	正规点来说是 冬季的奥运会

4 基于答案辅助的半监督问题分类方法

如图 1 所示，是本文所提出的问题分类方法的完整架构图。首先将答案特征叠加到问题特征中；其次根据改进后的标签传播方法，利用已标注问题预测未标注问题的类别；然后剔除这些问题中的答案特征，确定问题文本训练集，使用最大熵模型训练问题文本分类器；最终利用问题文本测试集测试问题文本分类器的性能。

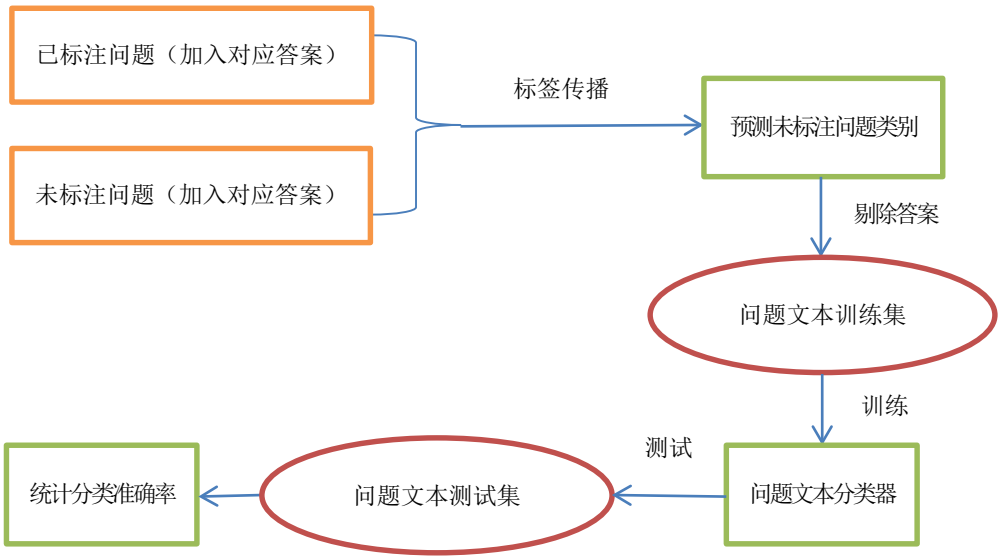


图 1 基于答案辅助的半监督问题分类方法架构图

Figure 1 Semi-supervised question classification with answers

[1] http://wenda.haosou.com/

4.1 特征介绍

对于问题和答案文本，我们均采用一元词特征。此外在预测未标注问题时，我们将答案特征加入到问题特征中来辅助问题分类。具体描述如表 2 所示。

表 2 问题与答案一元词特征及特征叠加实例

Table 2 Examples about Unigram of question and answer and features of superposition

	未分词	词特征
问题	爱情是什么？爱情的定义？	爱情 是 什么 ？ 爱情 的 定义 ？
答案	两情相悦，仅此而已	两 情 相悦 ， 仅 此 而 已
问答特征叠加		爱情 是 什么 ？ 爱情 的 定义 ？两_* 情_* 相悦 _* ，_* 仅_* 此_* 而_* 已_*

4.2 基于答案辅助的标签传播方法

根据标签传播（Label Propagation，LP）算法基本理论，每个节点的标签按相似度传播给相邻节点。在节点传播的每一步，每个节点根据相邻节点的标签来更新自己的标签；与该节点相似度越大，其相邻节点对其标注的影响权值越大；相似节点的标签越趋于一致，其标签就越容易传播。在标签传播过程中，保持已标注数据的标签不变，使其像一个源头把标签传向未标注数据。最终，当迭代过程结束时，相似节点的概率分布也趋于相似，可以划分到同一个类别中，从而完成标签传播过程[14][15]。

算法流程：

输入：

初始已标注问题样本集合 L_q，对应的答案样本集合 L_a，分别包含 n^+ 个正类样本和 n^- 个负类样本；

初始未标注问题样本集合 U_q，对应的答案样本集合 U_a；

初始已标注问题与答案叠加样本集合 $L = L_q + L_a$，包含 n^+ 个正类样本和 n^- 个负类样本；

初始未标注问题与答案叠加样本集合 $U = U_q + U_a$；

输出：

更新后的标注问题样本集合 L_q；

程序：

（1）初始化；

P： $n \times r$ 标注矩阵，同时 P_{ij} 标识文档 $i(i = 0 \cdots n)$ 属于类别 $j(j = 1 \cdots r)$ 的概率

P_L： P^0 的前 m 行对应的 m 个标注实例 L

P_U： P^0 的后 $n-m$ 行对应的 $n-m$ 个未标注实例 U

$\overline{T}$： $n \times n$ 矩阵，每一项 $\overline{t_{ij}}$ 表示从文档 i 到文档 j 的转移概率

a）设置迭代标记 $t = 0$，根据标注样本设定 P_L^0 的值；

b）初始化 P_U^0；

（2）循环迭代 N 次直到收敛；

a）传播实例的标注信息到相邻的实例依据公式 $P^{t+1} = \overline{T} P^t$；

b）还原标注实例的标注信息，即用 P_L^0 替代 P_L^{t+1}；

（3）对于每个未标注实例，根据 $\arg\max_j P_{ij}(j = 0 \cdots r)$ 得到它的正负标签，并添加到 L 中，从 U 中删除。

（4）从 L 中剔除答案样本集合 L_a，得到最终的标注问题样本集合 L_q

图 2 基于答案辅助的标签传播方法

Figure 2 Label propagation (LP) with answers

在许多问题分类相关研究中，文档通常用词袋（Bag-of-words）模型化并用向量形式描述。在这些设置中，单词与文档间的关联是不清晰的。为了更好地捕捉单词和文档之间的关系，本文采用基于文档-词的二部图表述文档与单词的关系。文档-词的二部图的连接关系由文档和词的连接矩阵表示，即 $n \times V$ 矩阵 X ； n 为文档数目， V 是词的数目。文档-词的二部图仅存在文档到词及词到文档的连接关系。具体来讲，文档到词及词到文档的转移概率计算如下[16]：

如果文档 d_i 包含词 w_k ，其权重为 x_{ik} ，则文档 d_i 到单词 w_k 的转移概率为 $\frac{x_{ik}}{\sum_k x_{ik}}$ ；同理，单词 w_k 到文档 d_j 的转移概率为 $\frac{x_{jk}}{\sum_k x_{jk}}$ 。文档 d_i 到文档 d_j 的转移概率是由文档 d_i 通过该文档里面的所有词到达文档 d_j 的概率之和，即 $t_{ij} = \sum_k \frac{x_{ik}}{\sum_k x_{ik}} \cdot \frac{x_{jk}}{\sum_j x_{jk}}$ 。得到文档间的转移概率之后，可以通过标签传播算法计算未标注样本的标签。本文所提出的方法在每个文档中均加入了答案特征，辅助问题分类。图 2 为本文提出的基于答案辅助的标签传播方法的算法流程。

5 实验

5.1 实验设置

实验使用六个主题的问答语料，每个主题设计为一个二元分类问题，即“该主题”与“非该主题”分类。例如：“电脑”类 2000 条问答对（一个问题对应一个答案），“非电脑”类是从其他五个主题中分别随机选取 400 条问答对，构成 2000 条问答语料。因此，实验中一共包括 6 个二元分类问题。语料分词采用复旦大学自然语言处理实验室开发的分词软件 FudanNLP[2]。分类算法采用 MALLET 机器学习工具包中的最大熵分类器 [3]，所有参数都设置为默认值。分类特征选取词的一元特征（Unigram），使用准确率作为结果的评价标准。根据初始标注样本规模的大小，我们给出两组不同的实验设置：

（1）第一组六个二元主题分类任务的实验，分别随机选取每个主题 5%的问答作为已标注问题样本，75%作为未标注问题样本，20%作为测试样本。

（2）第二组六个二元主题分类任务的实验，分别随机选取每个主题 10%的问答作为已标注样本，70%作为未标注问题样本，20%作为测试样本。

5.2 实验结果与分析

实验比较的方法详细描述如下：

- **Baseline：** 只利用初始标注样本训练问题分类器（没有利用任何非标注样本）
- **Self-training：** 利用整个特征空间构建分类器，并用它迭代加入置信度最高的样本扩充标注样本集合
- **LP（问题）：** 利用已标注样本通过标签传播方法预测未标注样本的类别，将这些已确定类别的问题全部作为训练样本，训练问题分类器
- **LP（问题+答案）：** 将对应的答案特征加入到已标注和未标注问题中，利用已标注样本通过 LP 预测未标注样本的类别，再剔除答案特征，将这些已确定类别的问题全部作为训练样本，训练问题分类器

图 3 显示当初始标注样本 5%时，四种方法的分类性能比较。从图中结果可以看出，我们的方法获得的分类效果明显优于其他方法，分类准确率比使用 Self-training 和 LP 方法分别平均提高了 11.1%和 3.9%。图 4 显示当初始标注样本 10%时，四种方法的分类性能比较。从图中结果可以看出，我们的方法同样获得最佳的分类效果，分类准确率比使用 Self-training 和 LP 方法分别平均提高了 10.7%和 3.6%。具体比较结果如下：

（1） Self-training 方法性能比 Baseline 还差，可能原因是，少量的问题标注样本刚开始预测的准确率很低，再一步步的迭代错误类别的问题样本，带来更大的错误。

（2） 利用 LP 方法预测未标注样本再进行问题分类，其分类准确率要明显高于 Baseline 和 Self-training 方法，这是因为 LP 方法可以在标注和未标注的问题样

[2] https://code.google.com/p/fudannlp/
[3] http://mallet.cs.umass.edu/

本中互相传播标签，有效地提升标签预测准确率。该实验结果说明利用 LP 方法在该任务中能够发挥较好的优越性。

（3） 本文提出的利用答案信息的 LP 方法，在 12 组实验中分类准确率均远远高于其他三种方法。该结果表明答案信息确实可以扩充问题分类信息，有效地提升了问题分类准确率。

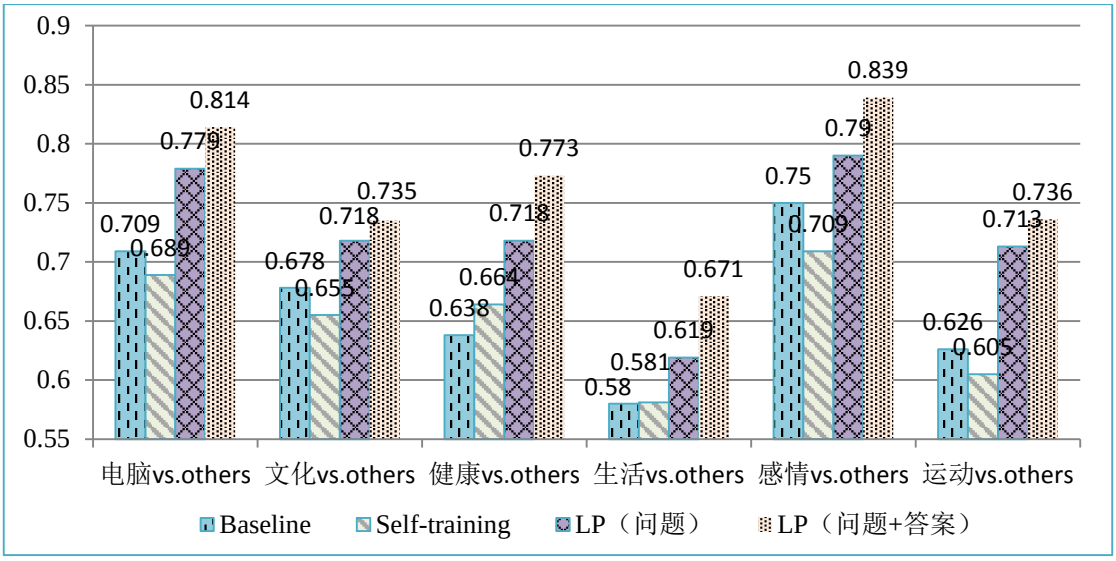


图 3 初始标注样本 5%时不同半监督分类方法性能比较

Figure 3 Comparison of different methods based on 5% of initial labeled samples

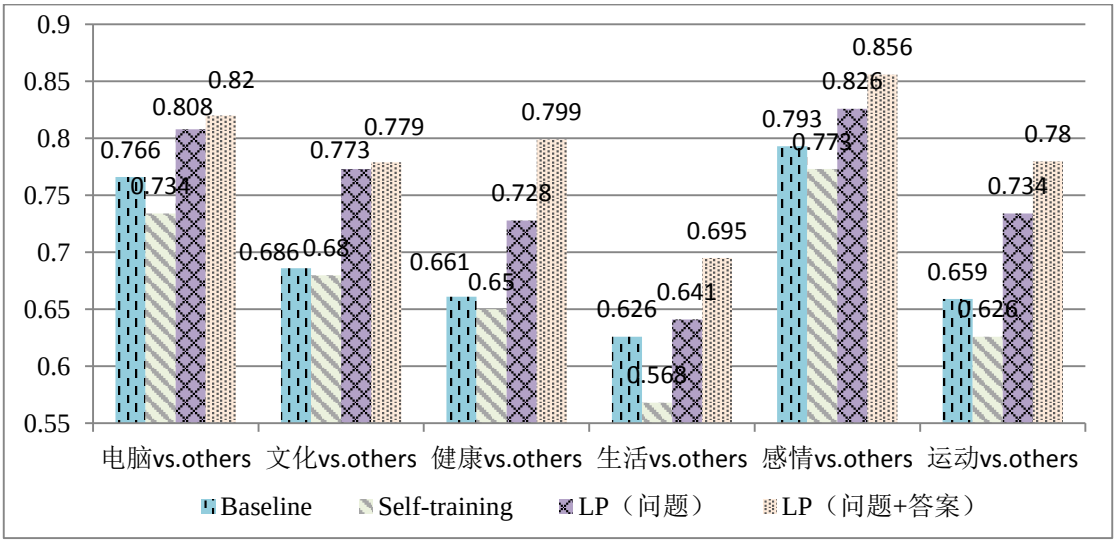


图 4 初始标注样本 10%时不同半监督分类方法性能比较

Figure 4 Comparison of different methods based on 10% of initial labeled samples

6 总结

本文针对问题分类任务，提出了一种基于答案辅助的半监督问题分类方法。该方法的特色在于充分利用已有答案的分类信息并能够结合 LP 方法进行半监督分类，从而减少了大量的人工标注工作。实验结果表明，该方法的性能在不同主题的任务中表现的都非常优秀，分类准确率明显高于传统的半监督学习方法，进一步地提高了半监督问题分类的准确率。

下一步工作中，我们将考虑使用更多的分类方法（如矩阵分解模型）进一步提高半监督

问题分类性能。我们也将考虑利用更多的特征（如：语义、句法），考察这些特征是否可以提高问题分类的准确性。

参 考 文 献

[1] 李鑫，黄萱菁，吴立德. 基于错误驱动算法组合分类器及其在问题分类中的应用[J]. 计算机研究与发展, 2008, 45(3):535-541.

[2] 高超. 中文问题分类中特征选择研究[D]. 安徽：安徽工业大学, 2011.

[3] Li S, Huang C R, Zhou G, et al. Employing personal/impersonal views in supervised and semi-supervised sentiment classification[C]. Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, 2010: 414-423.

[4] Li S, Huang L, Wang J, et al. Semi-Stacking for Semi-supervised Sentiment Classification[J]. Volume 2: Short Papers, 27.

[5] Ray S K, Singh S, Joshi B P. A Semantic Approach for Question Classification Using WordNet and Wikipedia[J]. Pattern Recognition Letters, 2010, 31(13):1935–1943.

[6] Hui Z, Liu J, Ouyang L. Question Classification Based on an Extended Class Sequential Rule Model[C]. IJCNLP. 2011: 938-946.

[7] Mishra M, Kumar Mishra V, Sharma H R. Question Classification Using Semantic, Syntactic and Lexical features[J]. International Journal of Web & Semantic Technology, 2013, 4(3).

[8] Yadav R, Mishra M, Bhilai S. Question Classification Using Naïve Bayes Machine Learning Approach[J]. International Journal of Engineering and Innovative Technology (IJEIT), 2013, 2(8).

[9] LIMSI-CNRS, ENSIIE. Question Classification Transfer[J]. In Proceedings of ACL. 2013:429-433.

[10] 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[J]. 计算机应用研究，2010, 27(8):2869-2871.

[11] 刘小明，樊孝忠，李方方. 一种结合本体和焦点的问题分类方法[J]. 北京理工大学学报，2012, 32(5):498-502.

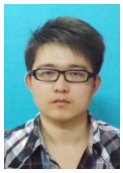
[12] 张巍，陈俊杰. 信息熵方法及在中文问题分类中的应用[J]. Computer Engineering and Applications, 2013, 49(10).

[13] Liu L, Yu Z, Guo J, et al. Chinese Question Classification Based on Question Property Kernel[J]. International Journal of Machine Learning & Cybernetics, 2014, 5(5):713-720.

[14] 张俊丽，常艳丽，师文. 标签传播算法理论及其应用研究综述[J]. 计算机应用研究, 2013, 30(1):21-25.

[15] Li S, Xue Y, Wang Z, et al. Active learning for cross-domain sentiment classification[C]. Proceedings of the Twenty-Third international joint conference on Artificial Intelligence. AAAI Press, 2013: 2127-2133.

[16] 高伟，王中卿，李寿山. 基于集成学习的半监督情感分类方法研究[J]. 中文信息学报，2013, 27(3):120-126.



张栋（1991-），男，通讯作者，硕士研究生，主要研究领域为自然语言处理。
E-mail: d.zhang.suda@gmail.com

李寿山（1980-），男，教授，硕士生导师，主要研究领域为自然语言处理。
E-mail: lishoushan@suda.edu.cn

周国栋（1967-），男，教授，博士生导师，主要研究领域为自然语言处理
E-mail: gdzhou@suda.edu.cn