DOI:10.3969/j.issn.1673-4785.201310014
网络出版地址:http://www.cnki.net/kcms/doi/10.3969/j.issn.16734785.201310014.html

基于无标记 Web 数据的层次式文本分类

何力,谭霜,贾焰,韩伟红
(国防科学技术大学 计算机学院,湖南 长沙 410073)

摘 要:传统的文本分类方法需要标注好的语料来训练分类器,然而人工标记语料代价高昂并且耗时。对此,通过无类别标记的 Web 数据来训练文本分类器,提出一种基于无标记 Web 数据的层次式文本分类方法,该方法结合类别知识和主题层次信息来构造 Web 查询,从多种 Web 数据中搜索相关文档并抽取学习样本,为监督学习找到分类依据,并结合层次式支持向量机进行分类器的学习。实验结果表明,该方法能够利用无标记 Web 数据学习分类器,并取得了较好的分类效果,其性能接近于有标记训练样本的监督分类方法。

关键词:层次式文本分类;主题层次;无标记数据分类;支持向量机
中图分类号: TP181 **文献标志码**:A **文章编号**:1673-4785(2014)03-0330-06

中文引用格式:何力,谭霜,贾焰,等. 基于无标记 Web 数据的层次式文本分类[J]. 智能系统学报, 2014, 9(3): 330-335.
英文引用格式:HE Li, TAN Shuang, JIA Yan, et al. Hierarchical text classification with non-labeled web data[J]. CAAI Transactions on Intelligent Systems, 2014, 9(3): 330-335.

Hierarchical text classification with non-labeled web data

HE Li, TAN Shuang, JIA Yan, HAN Weihong
(School of Computer, National University of Defense Technology, Changsha 410073, China)

Abstract: Traditional text classification methods require a labeled corpus to train classifiers, however, it is costly and time-consuming to label corpus manually. This paper proposes a hierarchical text classification method, which trains the text classifier with web data that does not require any classification labels. This method constructs web inquiry by combining classification knowledge and topic hierarchical information, searches relevant documents and extracts the learning sample from many kinds of web data, finds a classification basis to monitor the learning, and combines a hierarchical support vector machine to train classifiers. The experimental results show that this method is able to train classifiers through non-labeled web data, and gains a better result of classification with a performance that is at a level close to the supervised classification method with labeled training samples.

Keywords: hierarchical text classification; topic hierarchy; classification without labeled data; support vector machine

为了实现对互联网上信息的有效管理和访问,人们一般按照一个概念或主题类别层次对网络信息进行标记和组织,以更好地搜索和访问网络资源。这些主题类别一般被组织为树形结构,例如雅虎目录(Yahoo! directory)和 ODP(open directory project)。对于 Web 文本分类这个问题,传统的有监督方法需要标注好的语料来训练分类器。在实际分类问题中,对于一个由专家编制的类别层次,通常并没

收稿日期:2014-03-25. **网络出版日期**:2014-06-14.
基金项目:国家“863”计划资助项目(2010AA012505, 2011AA010702, 2012AA01A401, 2012AA01A402);国家重点基础研究发展计划资助项目(2013CB329601, 2013CB329602);国家自然科学基金资助项目(60933005, 91124002);国家科技支撑计划资助项目(2012BAH38B04);国家 242 信息安全计划资助项目(2011A010).
通信作者:何力. E-mail:lihe@nudt.edu.cn.

有标注好的语料,而 Web 分类目录的规模往往比较大,通常包含数百甚至数千个类别,此时通过人工标记文档类别来构建语料库将是一项非常繁重的工作。因为需要为每个类别人工标记足够多的训练样本,这项工作需要耗费巨大的人力成本来完成。对此,本文试图实现一个不需要有标记训练样本的层次式文本分类方法。

针对文本主题分类缺少训练样本的问题,已有工作利用外部数据源来丰富类别的特征信息[1-9],这些方法利用类别的特征关键词以及类别层次的上下文信息,到 Web 中获取更多的相关数据,为分类学习产生训练样本,增加类别的分类依据。因为这一类方法在分类学习过程中不需要人工标记训练样本,称为无标记数据分类方法。无标记数据分类利用 Web 搜索引擎和开放数据库来获取训练样本。对于 Web 搜索引擎,如谷歌,可以利用类别名称以及类别的上下文信息搜索相关页面,那么搜索结果应该和该类别具有一定相关性,体现了该主题类别的特征。对于开放数据库,如维基百科、ODP 等,可以利用主题类别在知识库中搜索相关文档,将这些文档作为该类别的样本。

无标记数据分类方法借助外部数据源学习分类模型,但是通过 Web 获得的学习样本可能会包含噪声数据,从而影响分类学习效果[2,5],这是其面临的一个主要挑战。本文针对 Web 搜索结果中含有噪声数据的问题,采用以下 3 个手段来提高分类学习效果。

1)利用类别知识和类别层次信息构造准确的 Web 查询,采用节点的标签路径来产生查询关键词;

2)利用多数据源产生样本,同时从谷歌搜索引擎、维基百科这 2 个数据源搜索相关页面和文档,以获取更加全面的样本数据;

3)结合类别层次对样本数据分组,为每个类别获得更加完整的特征源,根据搜索到的样本数据,利用主题类别层次学习分类模型,减小噪声数据的影响。

相比已有的无标记数据分类方法,本文提出的方法通过这些手段可以获取更加有效的样本数据。在得到样本数据之后,采用支持向量机分类算法训练层次式分类模型,最后在 ODP 数据集上对提出的无标记数据分类方法进行实验验证。

1　无标记数据分类相关工作

对于有监督学习缺少训练样本的问题,Weiss[10]可以将已有的解决策略概括为 3 类:过采样技术[11-13]、利用外部数据生成新样例[2-9]、采购新样例。过采样技术通过简单复制已有样本来增加训练集,学习过程中会产生过拟合问题,而直接购买数据则要承担非常高昂的代价。利用外部数据源生成新样例则可以通过技术手段获取比较有价值的新样例。

在生成新样例方面,已有工作主要是利用 Web 数据为主题层次中的类别产生新的样本文档,来丰富类别的特征信息[2-9]。Ha-Thuc 等[2]根据类别层次为每个类别构造一个查询,查询词包括类别名称、类别描述以及父、子节点名称,利用 Web 搜索引擎获取相关文档,然后采用产生式模型对文档建模,为每个类别建立一个语言模型,最后采用自上而下的分类方法对文档分类。Wetzker 等[3]利用类别主题名以及父类别主题名构造查询,采用雅虎搜索获取每个类别的 top-k 相关文档,然后为每个类别学习一个中心向量,构造了一个支持多标签分类的层次式分类器。Zhang 等[4]利用 ODP 目录的数据集学习主题分类模型,并对知识监督学习中的预测风险进行优化。Huang 等[5]利用那些在一个类别和它祖先类别中同时出现的单词构造查询,以谷歌搜索结果作为训练样本,然后采用 KNN 算法对文档进行分类。这些方法利用主题目录的类别层次信息和类别知识来构造 Web 查询。类别知识包括一个类别的主题名称、关键词、描述信息等。除了类别自身知识之外,还可以利用主题层次的结构特征,例如类别在主题层次中的父类别、子类别、邻居类别等信息。Wang 等[6]利用维基百科知识库构造通用分类器,该方法首先人工为每个类别确定一组关键词,然后根据这些类别关键词到维基百科中获取相关概念与文档,最后利用这些概念与文档训练分类器。Hung 等[7-8]提出了一种 Web 语料获取方法,该方法首先为每个类别搜索少量相关度较高的 Web 文档,然后从这些文档中抽取出类别的关键词,然后利用这些关键词搜索更多的相关 Web 文档。刘丽珍等[9]提出一种模糊划分聚类方法,该方法对无标记样本进行模糊划分聚类,通过一定的相似度度量,将相关文本归并,得到少量标记文本,从而为监督学习找到了分类依据。

另外,Chen 等[1]试图利用外部数据源学习类别和词汇之间的关系,即在每个类别中不同单词的概率权重,从自然语言处理角度考虑,就是为每个类别建立一个语言模型,从而实现对微博短文本的主题分类。Ko 等[14]采用自举法进行机器标注样本,根据无标记文档集合和类别的标题词来自动生成标记

文档,然后针对机器标注过程中的噪声数据,采用特征投影技术训练分类器。Veeramachaneni 等[15]提出了一个层次式狄利克雷产生式模型,对类别层次中的语料文档进行主题建模,通过学习每个类别中不同单词的概率权重,实现了一种无监督分类方法。

2 无标记数据的 HSVM 分类模型

首先利用类别知识和主题层次信息从 Web 数据获取每个类别的相关文档,然后根据这些相关文档为主题类别层次学习分类模型。

2.1 获取 Web 样本

本文采用多种技术手段来提高样本文档的质量,首先根据类别标签路径构造 Web 查询,然后融合多个 Web 数据源的搜索结果产生相关文档,最后利用类别层次结构对相关文档进行数据分组,具体过程如图 1 所示。

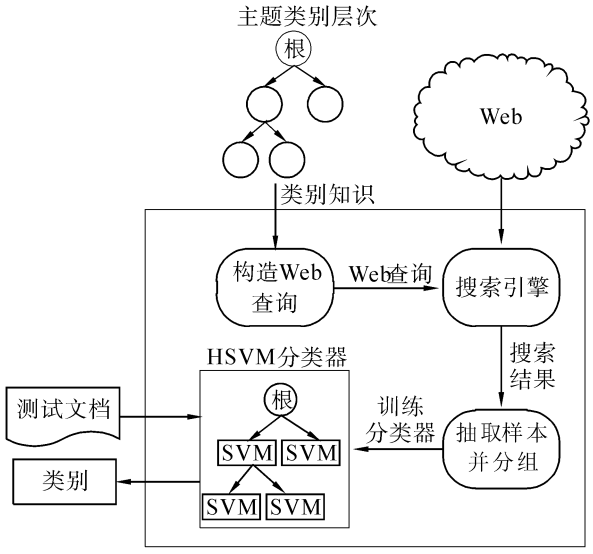


图 1 无标记文档分类方法示意图

Fig.1 The classification method with no labeled data

1)构造 Web 查询

已有方法利用类别的本体知识和类别层次的上下文信息为类别构造查询[2-3,5]。本文利用类别在类别层次树中的标签路径来构造查询,即以从根节点到该节点路径上的所有类别的名称作为查询词。例如,对于 ODP 中的类别“英语”,其标签路径为“科学_社会科学_语言学_语言_英语”,那么以“科学”、“社会科学”、“语言学”、“语言”、“英语”这些词汇作为该类别的查询词。相比已有的这些方法,利用类别树为每个类别生成查询词,更能代表一个类别的完整语义。

2)搜索相关文档

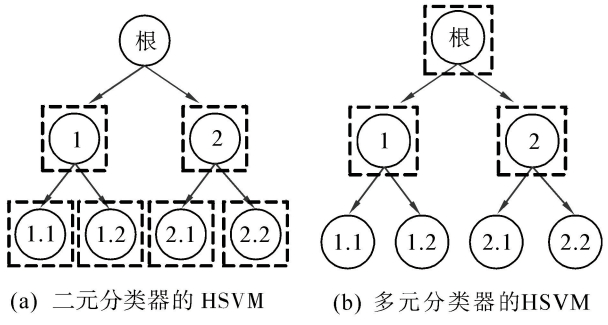
本文同时从 Web 搜索引擎和开源分类目录来获取样本。对于 Web 搜索引擎,采用谷歌从互联网上搜索相关页面。对于开源目录,由于本文采用 ODP 目录进行实验测试,因此采用了维基百科来搜索相关文档。通过利用多种 Web 数据,本文能够获取更多的相关文档数据,减少噪声数据的影响。

3)样本抽取

在从 Web 数据中搜索到这些相关页面之后,需要将其转化为训练样本。对于搜索到的相关页面,本文按照标准的文本处理过程抽取网页中的文本,删除停用词和低频词,将文档转换为 TFIDF 特征向量。另外采用 top-down 数据分组方式,对于一个类别,首先找到类别树中以该类别为根节点的子树,然后将该子树中所有节点的相关文档作为这个类别的训练样本。这样结合类别层次对数据分组,可以为每个类别获得更加完整的特征源。

2.2 学习分类器

在从 Web 获取样本数据之后,接下来结合主题类别层次学习分类模型。本文采用层次式支持向量机(hierarchical SVMs,HSVM)学习分类模型,基于搜索到的 Web 样本训练 HSVM 分类器。HSVM 是一个基于支持向量机的层次式分类模型,已被验证是一个有效的层次式文本分类方法[16]。本文实现并比较了 2 种 HSVM 方法,分别是二元分类器的 HSVM 和多元分类器的 HSVM。二元分类器的 HSVM 为类别层次树中除根节点以外的每个节点训练一个二元 SVM 分类器,对文档进行自上而下的分类。二元分类器的 HSVM 如图 2(a)所示,每个虚线框表示一个二元分类器,对于一个文档,自上而下进行分类预测,由每个节点上的本地分类器判断文档是否属于当前类别。多元分类器的 HSVM 如图 2(b)所示,根据类别层次树逐层为具有相同父节点的所有类别建立一个多类 SVM 分类器,即在类别层次树中所有中间节点上分别训练一个多类分类器,对文档进行自上而下的分类。这 2 种 HSVM 均是对测试文档进行自上而下的分类预测。



(a) 二元分类器的 HSVM (b) 多元分类器的HSVM

图 2 HSVM 分类模型

Fig.2 The classification models of HSVM

Liblinear[17]是台湾大学林智仁教授开发的一

个 SVM 分类器,根据林智仁小组的研究结果,Liblinear 适用于具有高维特征的 Web 文档分类,因此本文采用 LibLinear 来实现 HSVM。

3　实验结果与分析

3.1　实验准备

本文采用 ODP 简体中文网站目录作为实验对象,ODP 简体中文网站目录是一个深度为 6 层的类别层次树,包括参考、商业、休闲、体育、健康、计算机、新闻、家庭、社会、游戏、艺术、购物、科学等 13 个大类,1 763 个类别,整个目录包括 24 570 个网站。根据 ODP 中的网站 URL 爬取页面,然后对采集到的网页进行解析、分词和停用词过滤,最终将每个网站表示为一个文档。ODP 数据的类别分布和文档分布如图 3 所示。

在 ODP 样本集中,有 1 048 个类别的样本个数不足 10 个,由于这些稀有类别的实例非常少,采用现有的机器学习方法很难对这些类别的网页进行有效地自动分类。为了使有监督分类算法能够同本文提出的方法进行公平比较,采用父类别模型对稀有类别进行分类预测,即将文档分到稀有类别的父类别后就不再继续往下细分,以避免这些稀有类别对有监督分类方法的性能影响。

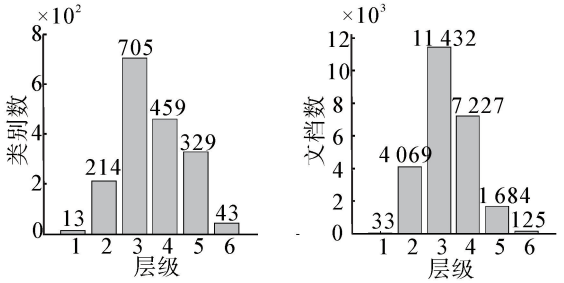


图 3　数据的层次分布

Fig.3　Data distribution on different level

网页文档是一种高维数据,因此需要进行特征降维以解决文本特征向量高维问题,本文采用基于词频逆文档频率值的特征词子集选择方法进行特征降维。对于有监督分类方法,先将 ODP 数据集随机分为 10 份,其中 1 份为测试集,其余作为训练集,然后训练分类器并计算各评价指标,如此反复 10 次,以这 10 次的平均值作为最终结果。对于无标记数据分类方法,本文采用 Web 样本训练分类器,然后对 ODP 数据集进行测试并计算各评价指标。

为了获取更加广泛的 Web 数据,同时从谷歌和维基百科搜索相关文档。对于一个主题类别,首先利用谷歌搜索引擎搜索相关页面,并从中抽取出相关文档,然后同样在维基百科中搜索该主题类别的相关文档,补充到该类别的训练样本中去。最后结合所有从谷歌和维基百科获取到的样本训练分类器,并将其记为 GW-HSVM (Google Wikipedia based HSVM)。具体在实验中,取谷歌搜索结果的 top-50 作为相关文档,取维基百科搜索结果的 top-10 作为相关文档。

对于标注样本的有监督分类方法,文中采用有标记的 ODP 数据集训练 HSVM 分类器,记为 S-HSVM (Supervised-HSVM)。显然,GW-HSVM 是基于 Web 样本的无标记数据分类方法,S-HSVM 是有监督分类方法。

对于文本分类问题,通常采用精度 precision、召回率 recall、F_1 评价分类算法的好坏,同时根据这些指标的宏平均值和微平均值来衡量算法在所有类别上的性能。微平均评价指标体现了大类别对结果的影响,宏平均评价指标给每个类别以相等权重,更能体现算法在小类别上的性能表现。

本文实验中的数据为单标签文档,此时 precision、recall 和 F_1 的微平均值均相等,等于分类的准确率 accuracy。因此,采用 Macro-Precision, Macro-Recall, Macro-F_1 和 accuracy 作为分类算法的评价标准。另外,层次式分类方法在自上而下的分类过程中会产生错误传播问题,对此分析了算法在不同层级上的性能表现。在类别层次中,随着深度增加,会出现大量的小类别,对此采用宏平均指标评价算法在各层级上的性能。具体在计算第 n 级的宏平均指标时,只考虑第 n 级上所有类别精度、召回率和 F_1 的宏平均值。

3.2　实验结果

在实验中可以发现,二元分类器的 HSVM 和多元分类器的 HSVM 在分类准确率上性能接近,但是多元分类器的 HSVM 需要的训练和预测时间要更少,这是因为多元分类器方法不需要在叶子节点上训练分类器,如图 2(b)所示。因此,本文在实验中采用多元分类器实现的 HSVM。

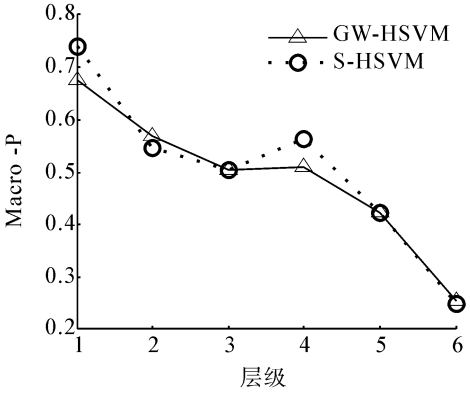
GW-HSVM 和 S-HSVM 对 ODP 中文目录所有类别的分类性能如表 1 所示,包括精度、召回率、F_1 的宏平均值以及准确率。可以看到,GW-HSVM 的分类准确率稍低于有监督分类方法 S-HSVM,但是在宏平均指标上,GW-HSVM 的性能接近 S-HSVM,这说明 GW-HSVM 能够对小类别进行更为有效的分类,这是因为 GW-HSVM 为每个类别采集了足够多的 Web 训练文档,而 S-HSVM 所采用的 ODP 数据集中则包含有大量的小类别。

表 1 整体分类性能比较

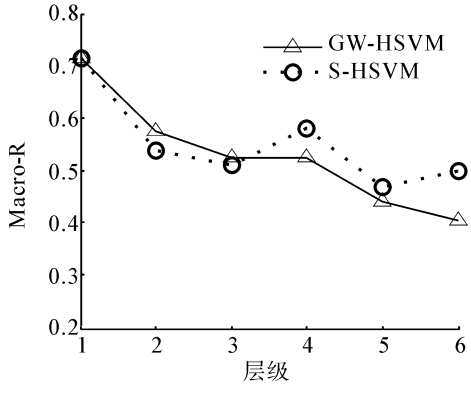
Table 1 Overall classification performance comparison

模型	Macro-P	Macro-R	Macro-F_1	准确率
GW-HSVM	0.5196	0.5367	0.5280	0.4775
S-HSVM	0.5285	0.5379	0.5332	0.5482

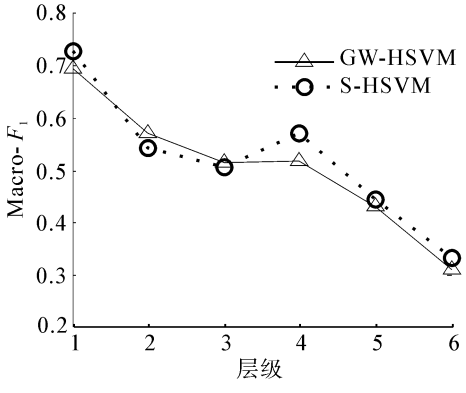
本文还比较了 S-HSVM 和 GW-HSVM 在类别树中不同层级上的分类性能,包括 Macro-P、Macro-R 和 Macro-F_1,如图 4 所示。



(a) Macro-P



(b) Macro-R



(c) Macro-F_1

图 4 不同层级上的分类性能

Fig.4 Performance on different level

可以看到,GW-HSVM 在第 1 级和第 4 级上的性能差于 S-HSVM,这是因为 ODP 中文目录中这两层上的类别包含较多的实例。对于目录中其他几层,由于这些层级中包含有大量稀有类别,这时 GW-HSVM 的分类性能接近甚至优于 S-HSVM。结合表 1 和图 4 的实验结果可以发现,本文提出的无标记数据分类方法取得了较好的分类效果,其性能接近于有标记训练样本的监督分类方法。

4 结束语

本文针对主题分类目录缺少训练样本的问题,提出了一种无标记数据的层次式文本分类方法,该方法利用搜索引擎从 Web 数据中获取训练样本,通过有效的 Web 查询和样本抽取方法降低了噪声数据的影响,获得了较好的分类效果,其分类性能接近于有标注训练样本的监督分类方法。

参考文献:

[1] CHEN Y, LI Z, NIE L, et al. A semi-supervised bayesian network model for microblog topic classification[C]//Proceedings of the 24th International Conference on Computational Linguistics. Mumbai, India, 2012: 561-576.

[2] HA-THUC V, RENDERS J M. Large-scale hierarchical text classification without labelled data[C]//Proceedings of the fourth ACM International Conference on Web Search and Data Mining. Hong Kong, China, 2011: 685-694.

[3] WETZKER R, ALPCAN T, BAUCKHAGE C, et al. An unsupervised hierarchical approach to document categorization[C]//Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. Silicon Valley, USA, 2007: 482-486.

[4] ZHANG C, XUE G R, YU Y. Knowledge supervised text classification with no labeled documents[C]//Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence. Hanoi, Vietnam, 2008: 509-520.

[5] HUANG C C, CHUANG S L, CHIEN L F. Liveclassifier: creating hierarchical text classifiers through Web corpora [C]//Proceedings of the 13th International Conference on World Wide Web. New York, USA, 2004: 184-192.

[6] WANG P, DOMENICONI C. Towards a universal text classifier: transfer learning using encyclopedic knowledge[C]// Proceedings of the Ninth IEEE International Conference on Data Mining Workshops. Miami, USA, 2009: 435-440.

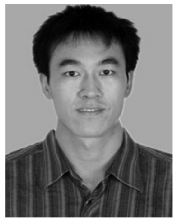
[7] HUNG C M, CHIEN L F. Web-based text classification in the absence of manually labeled training documents[J]. Journal of the American Society for Information Science and Technology, 2007, 58(1): 88-96.

[8] HUNG C M, CHIEN L F. Text classification using Web corpora and em algorithms[C]//Proceedings of the Asia Information Retrieval Symposium. Beijing, China, 2005: 12-23.

[9] 刘丽珍, 宋瀚涛, 陆玉昌. 无标记训练样本的 Web 文本

分类方法[J]. 计算机科学, 2006, 33(3): 200-201.
LIU Lizhen, SONG Hantao, LU Yuchang. The method of Web text classification of using non-labeled training sample [J]. Computer Science, 2006, 33(3): 200-201.
[10]WEISS G M. Mining with rarity: a unifying framework[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 7-19.
[11]CHEN S, HE H, GARCIA E A. Ramoboost: ranked minority oversampling in boosting[J]. Neural Networks, IEEE Transactions on. 2010, 21 (10): 1624-1642.
[12]NGUYEN H M, COOPER E W, KAMEI K. Borderline over-sampling for imbalanced data classification[J]. International Journal of Knowledge Engineering and Soft Data Paradigms, 2011, 3(1): 4-21.
[13]GAO M, HONG X, CHEN S, et al. A combined smote and pso based rbf classifier for two-class imbalanced problems[J]. Neurocomputing, 2011, 74(17): 3456-3466.
[14]KO Y, SEO J. Learning with unlabeled data for text categorization using bootstrapping and feature projection techniques[C]//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Barcelona, Spain, 2004: 255-262.
[15]VEERAMACHANENI S, SONA D, AVESANI P. Hierarchical dirichlet model for document classification[C]// Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany, 2005: 928-935.
[16]CAI L, HOFMANN T. Hierarchical document categorization with support vector machines[C]//Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management. Washington, DC, USA, 2004: 78-87.
[17]FAN R E, CHANG K W, HSIEH C J, et al. Liblinear: a library for large linear classification[J]. Journal of Machine Learning Research, 2008, 9: 1871-1874.

作者简介:



何力,男,1984 年生,博士研究生,主要研究方向为网络与信息安全、数据库与数据挖掘,发表学术论文 6 篇。

谭霜,男,1984 年生,博士研究生,主要研究方向为网络与信息安全、云计算,发表学术论文 5 篇。

贾焰,女,1960 年生,教授,博士生导师,主要研究方向为网络与信息安全、数据库与数据挖掘、社会网络,发表的学术论文被 SCI 和 EI 检索 200 余篇。

2014 年第 4 届中国智能产业高峰论坛

中国拥有全球规模最大、增长速度最快的智能产业市场,我国未来经济的持续发展,寄希望于科学技术的不断发展创新;而智能科技产业领域的创新与合作,必将成为中国未来科技整体发展与创新的原动力。中国智能产业高峰论坛由中国人工智能学会发起主办,获得了国内外众多学术团体和智能产业相关企业的鼎力支持和广泛合作,并得到了中国科学技术协会、中国工程院、中国科学院等单位的指导。高峰论坛将邀请国内外高水平智能产业领域专家和知名学者发表前沿学术研究报告,同时也将邀请国内外企业界高管做产业发展报告。另外,高峰论坛也会针对学术界和企业界最新成果进行融合讨论,促进智能技术产学研上的一条龙式发展。

高峰论坛的探讨话题主要涵盖:

1)智慧城市 ;2)智慧医疗 ;3)智能交通 ;4)智能电网 ;5)智能控制 ;6)智能安全 ;7)云计算 ;8)物联网 。

联 系 人:邹老师

联系电话:010-62281360

会议网站:http://conference.bupt.edu.cn/summit2014/