

面向审计领域的短文本分类技术研究

伍 洋¹, 钟 鸣¹, 姜 艳², 李石君¹

(1 武汉大学 计算机学院, 湖北 武汉 430072; 2 湖北省审计厅计算机中心, 湖北 武汉 430072)

摘 要: 针对审计问题这种短文本所具有的特征稀疏、问题类别界限模糊问题, 提出了一种改进的面向审计领域的短文本分类方法. 该方法首先为审计问题构造了专门的特征集, 以审计领域的同义词词集和法规库为基础, 并结合特定规则来调整特征权重, 然后以修改的 SVM 决策树作为多类分类器进行短文本分类. 实验结果表明, 该方法在对审计问题分类的应用上, 具有较为满意的正确率, 能满足实际的分类需求.

关键词: 审计问题分类; 审计领域; 信息增益; SVM 决策树; 短文本分类; 审计报告

中图分类号: TP391.1

文献标识码: A

文章编号: 1000-7180(2015)01-0005-06

Study on Short Text Categorization Technology Oriented Towards Field of Auditing

WU Yang¹, ZHONG Ming¹, JIANG Yan², LI Shi-jun¹

(1 School of Computer, Wuhan University, Wuhan 430072, China;

2 Computer Center of Hubei Audit, Wuhan 430072, China)

Abstract: To deal with the problems of feature sparseness and fuzzy boundaries of categorization exists in classification of audit problems, an improved short text categorization method oriented towards field of auditing is put forward. Firstly, a specialized feature set is builded for audit problems, the primary calculation method is designed based on synonym word set oriented towards field of auditing as well as designated rules, law library. Then the feature weight of those words with highly similarity to target words is adjusted. Finally, the SVM decision tree is used as multi-class classifier for short text classification. Experimental results show that a satisfy result can be got with this method from problem categorization of audit reports and it can be used in practical needs.

Key words: classification of audit problem; field of auditing; information gain; SVM decision tree; short text categorization; audit report

1 引言

目前,在许多领域,由于存在大量的非结构化文本数据,需要通过文本分类技术来对其进行处理,以便于开展进一步的分析工作^[1-3].

本文主要关注审计领域的一个典型的文本分类问题,即审计报告中审计问题的分类. 审计问题是一段短文本,其在概念描述上信息较为贫乏,文本特征不明显^[4],此外,审计问题的分类比较细,粒度较高,分类之间界限较为模糊. 这两个主要特征决定了审

计问题分类的困难性.

本文提出了一种面向审计领域的短文本分类技术,能够有效地改善审计问题分类的效果. 以审计出的问题金额为例,通过将审计问题进行短文本分类,可间接将其中所涉及的问题金额按性质划分. 综合全省的大量数据,可得到问题金额分布图、历年问题金额趋势,以及挖掘问题金额之间的关联关系等一系列具有综合分析意义的信息,从而可以指导审计的工作重点方向,这对于推动审计信息化进程、提升处理非结构化数据能力具有重要意义.

收稿日期: 2014-03-20; 修回日期: 2014-05-13

基金项目: 国家自然科学基金(61272109, 61202036)

2 审计问题短文本分类

2.1 短文本表示

为了便于计算机处理非机构化的文本数据,我们将文本以向量空间模型(Vector Space Model)的方式来表示,一个文本便可对应于一个向量 $V = (v_1, v_2, \dots, v_n)$, 其中 v_i 为选取的第 i 个特征的权值. 基于 VSM 的理论基础,文本分类的过程就转换为将空间中的向量以一定的规则划分为不同集合的过程.

2.2 短文本分类流程

目前针对分类方法的研究工作比较少,长期以来是沿用着长文本的分类方法^[5],包括最近邻算法(kNN)、朴素贝叶斯(Bayes)、支持向量机(SVM)^[6]等基于单一分类器的方法,以及以 Boosting 为代表的基于集成学习的方法,这些依据词频特征相似性来进行分类的方法都取得了较为满意的分类结果,但相较于文本分类,短文本分类却面临以下几个难点.

(1)特征词稀疏:短文本对于问题主题、细节的描述较少,使得能够刻画问题特征的词汇较少(一般不超过一百个),从而难以利用特征词之间的关联性.

(2)样本分布不均衡:短文中充斥着无用信息,对于分类影响较大,需要在文本预处理阶段进行文本消噪.

(3)分类界限模糊:对于部分目标分类,由于对问题间的定义本来就存在交集,分类之间的界面就变得较为模糊,易于混淆,使得训练集中便存在部分错分类.

(4)文本相似性:部分由长文本切分而成的短文本,在一定程度上,文本结构、用词用语较为相似,但是其分类却不同,使得可用区别二者的特征较少.

本文结合审计问题的短文本分类实际需求,提出了改进的短文本分类过程,该分类流程如图 1 所示. 其中预处理阶段是文本分类过程中的重要过程. 对于审计报告的特殊文档结构,需要通过信息抽取的方式对文档进行特殊处理. 首先利用 POI(操作 OFFICE 文档的开源项目)提供的 API 函数来获取 DOC 文档内容,同时需要进一步纠正由于 WORD 的自动标题编号功能,以及转换过程中的误差所造成的标题无法正确提取的问题;然后以抽取的标题来构建文档目录,剔除审计基本情况、审计建议等内容,裁剪与分类工作不相关文本段落,从而将一个长

文本切分为多个独立的短文本.

对于每个独立的短文本,还需要采用中科院最新的 NLPIR 工具进行中文分词与词性标注,以及利用停用词表等来剔除一些与分类工作不相关,且具有冗余性的虚词、副词等. 实现由段细化到句,再细化至词,截取与文本分类相关的重要信息的过程^[7-9].

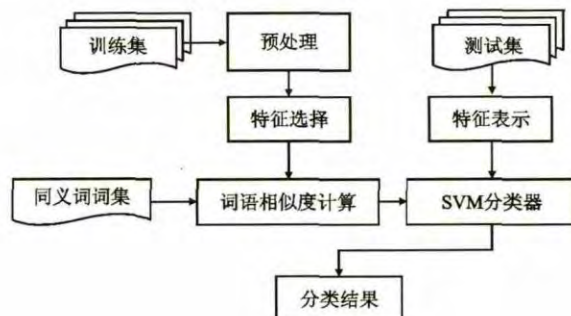


图 1 短文本分类过程

2.3 特征选择过程

特征选择阶段就是采用一定的算法或规则来选取与分类相关的特征. 在文本分类中主要使用的特征是对文本的类别有识别作用的词. 目前较为常用的方法有文档频率(Document Frequency)、互信息(Mutual Information)、信息增益(Information Gain)、卡方检验(CHI)等. 文献 1 表明,CHI 和 IG 方法的效果最佳,本文采用信息增益作为特征选择算法,其算法公式如式(1)所示:

$$IG(T) = H(C) - H(C | T)$$

$$\begin{aligned}
 &= - \sum_{i=1}^n P(C_i) \log_2 P(C_i) + \\
 &P(t) \sum_{i=1}^n P(C_i | t) \log_2 P(C_i | t) + \\
 &P(t) \sum_{i=1}^n P(C_i | \bar{t}) \log_2 P(C_i | \bar{t}) \quad (1)
 \end{aligned}$$

式中, $P(C_i)$ 表示类别 C_i 出现的概率, $P(C_i | t)$ 表示出现特征 T 的时候,类别 C_i 出现的概率. 依上式所述,其中特征 T 所产生的信息增益 $IG(T)$ 即为系统原本的熵与将特征 T 固定后的条件熵之间的差值, $IG(T)$ 越大,表示被选取为特征词的可能性越大. IG 综合考虑了特征出现与不出现对系统的影响,同 CHI 一样,是一种比较全面的算法^[1]. 经过特征选择所保留的特征词并不能极大地反映该词对文本分类结果的影响,此时需要根据实际的应用需求,来调整部分关键特征词的影响因子(权重),在特征词较少的情况下,使该词成为决定分类结果的因素.

此外为了利用审计问题中的问题定性,法律依

据等与分类结果具有较强相关性的规则(如违反《财政总预算会计制度》第二十三条的问题被划分为管理不规范问题,而违反第五十八条的问题则被划分为非金额计量问题),此时则还需对审计问题做进一步信息抽取处理,将所涉及的法律法规、定性等按一定规则进行编号,计算出对应权值并映射在 VSM 的指定维度上,增加的重要影响因子也提高了分类的精确度,同时也避免了这些规则中的同义词在同义词扩展后对系统产生不利影响。

2.4 特征权重调整

本文所采用的面向审计领域的同义词词集是在审计词条集的基础之上结合同义词词林,并效仿同义词词林的编排方式编成,同时以同义词词林的词汇相似度计算方法来计算两个词语之间的相似度。

审计领域的同义词词集的词条以树形结构形式组织在一起,共分 6 个大类共 29 个小类,涉及专项资金、经济责任、财政审计等,在每个小类之中,以词群(段落)为基础概率将词义相关性较高或者词义相近的词汇聚在一起,又将每个词群进一步划分为多行,每行中的词语在词义层次上,要么相同或相近,要么具有很强的相关性,图 2 所示为构建的同义词词集的 5 层结构图。

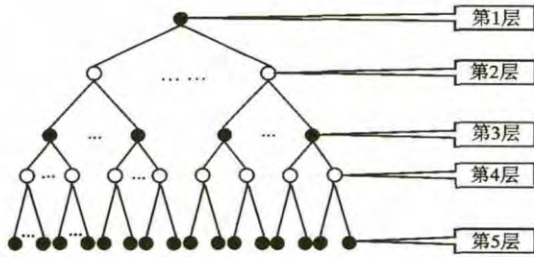


图 2 同义词词集的 5 层结构

根据田久乐等人^[7]以同义词词林的层次化结构所提出词汇相似度计算方法,词语 w_1 与词语 w_2 的相似度计算公式如式(2)所示:

$$\text{sim}(w_1, w_2) = \begin{cases} f(u) \times \cos(n \times \frac{\pi}{180}) (\frac{n-k+1}{n}), & w_1, w_2 \text{ 同树} \\ f_1, w_1, w_2, \text{不同树} \\ e, \text{其他} \end{cases} \quad (2)$$

式中, $\cos(n \times \frac{\pi}{180})$ 作为调节参数,用于将 $\text{sim}(w_1, w_2)$ 的相似度值调节为 $[0, 1]$ 区间内,控制参数 $(n-k+1)/n$ 在一定角度反映了两个词汇之间的相关性(n 为分支层节点总数, k 为两个分支间的距离),

$f(u), f, e$ 为通过大量实验所设定的参数值,当 w_1, w_2 不在同一颗树下时,其相似度统一设为 f ,并且小于置信阈值 η ($\eta \in (0, 1)$),当 w_1, w_2 位于同一颗树的树下时, $f(u)$ 根据两个词汇的义项在五层结构中的不同层次取不同值,文献 7 表明 $f(u)$ 取下值时精确度较高。

$$f(u) = \begin{cases} 0.6 & u = 2 \\ 0.8 & u = 3 \\ 0.9 & u = 4 \\ 0.96 & u = 5 \end{cases} \quad (3)$$

随着层次的加深,语义刻画越详细,即层次 u 越大,相似度就越高, $f(u)$ 应当递增。通过该算法,可以定量地获得任意两个词汇之间在语义层面上的相似度,从而可以筛选得到指定词汇的高于阈值 η 的同义词集合。

通过同义词计算,能明确短文中与分类目标相关度较高的词汇集合,此时应当加强这些词汇的权重,以使得特征更为明显,提高分类的准确度。本文根据审计报告的特点,设计加权公式如下:

$$\Phi(T_i) = f(TF, V(T_i), W(T_i)) \quad (4)$$

词频 TF 为特征词 T_i 在文档中出现的次数,由于审计报告属于公文性文档,用词重复性较多,词频越高,则表明该词的特征越明显。

$W(T_i)$ 为采用专家评定法决定的标题与正文出现词汇的权重比值,对于标题,正文中出现的目标词汇的同义词,通过专家评定法,区分不同部分出现词汇的权重,其中 $W(T_i)$ 表示如下:

$$W(T_i) = \begin{cases} \mu, & \mu > 1, T_i \text{ 在标题中出现} \\ 1, & T_i \text{ 在正文中出现} \end{cases} \quad (5)$$

在审计业务人员的参与下,确定 μ 的取值为 4.3,经实验表明,该值确为有效参数。

$V(T_i)$ 为与特征词出现的位置有关的函数,由于一般综合报告中,重要信息具有集中在前文的特征,即越重要的信息越与段首接近,对于分词后长度为 Len_i 的文档 D_i ,其第 i 个特征词 T_i 的重要性可简略表示为 $\frac{(\text{Len}_i - \text{Loc}_i + 1) \times W(T_i)}{\text{Len}_i}$,其中 Loc_i

表示该特征词出现的位置,则此时 $\Phi(T_i)$ 表示如下:

$$\Phi(T_i) = \begin{cases} \frac{\sum_{i=1}^{TF} \cos(\frac{(\text{Loc}_i - 1) \times W(T_i)}{\text{Len}_i} \times \frac{\pi}{3})}{TF} + 1, & T \text{ 是特征词} \\ 1, & T \text{ 不是特征词} \end{cases} \quad (6)$$

经调整后的特征词 T_i 对于目标分类 T' 权值为 $W(T_i) = IG(T_i) \times \text{sim}(T_i, T') \times \Phi(T_i)$ (7)

通过该公式,加强了特征更为明显的特征权值,有助于利用有限信息来突出分类特性,抑制次要特征信息.

3 SVM 决策树分类器的构造

SVM (支持向量机) 是一种典型的两类分类器,最先由 Vapnik 和 Cortes 于 1995 年提出,它在解决具有非线性、小样本性质以及在高维模式识别的应用中具有明显的优势^[4]. 目前, SVM 多类划分的研究工作主要集中在提出通过构建的 SVM 决策树作为多类分类器来逐层将输入样本点进行类别划分,其中每一次分类都将当前的输入模式划分为两个子类,这样,多类别分类问题就转成了两类分类的问题了. 对于 4 类分类,其中一特例的分类过程可如图 3 所示.

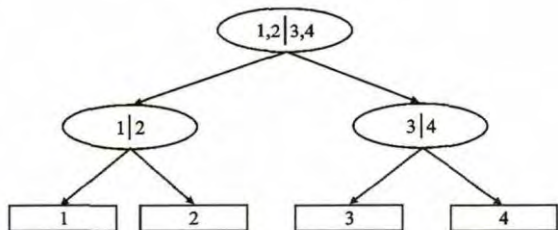


图 3 二叉决策分类树

由图 3 可知,对于 N 类问题该方法相较于传统的基于“one-against-one”和“one-against-the rest”算法的 SVM 多类别分类器在构建代价和分类效率上具有明显的提升. 三者之间的对比如表 1 所示.

表 1 三类 SVM 多分类方法的比较

	SVM 决策树	一对一	一对其余
分类器个数	$N-1$	$N(N-1)/2$	$N(N-1)/2$
评价分类次数	$\log_2 N$	N	N

对于两两可分的类,二者的类间距越大,表明此二者的可分性越好,因而可用类与类之间的间距来作为衡量两类之间可分性的一个标准^[6],即在所有的划分方法中,全部类间距最大的方案则是最佳方案. 本文在传统的基于类中心之间距离来进行度量类和类间可分性的基础上,根据优先考虑同义词群与规则的原则,对该划分算法进行如下修改,以此构建优化的 SVM 决策树分类器. 由于通过加权后,同义词词群与非同义词词群的权值所在值域不同,而同义词词群与指定维度上所存储的重要规则才是特征更为明显的关键所在,如果只考虑类中心的话,则会忽略同义词和关键规则对分类的影响,此时应该

优先考虑同义词词群与规则所在的类中心,设置规则所在的维度上的权值,使其值域与同义词词域相同,可以将二者合并.

假设分别用 S_1, S_2 来表示结点类划分中的正、反例类集合,用 N_{S1}, N_{S2} 表示对应类别集合的数目,则有该结点的总类别数 $C = N_{S1} + N_{S2}$, n_j 表示分类 j 的样本中含有同义词或规则的样本个数, $\overline{n_j}$ 表示分类 j 的样本中不含有任何同义词和规则的样本个数, x_j 表示分类 j 的含有同义词或规则的样本, $\overline{x_j}$ 表示分类 j 的不含有同义词和规则的样本. 那么修改的分类过程如下所示.

(1) 计算每一个类的含同义词或规则的样本中心与不含同义词的样本中心并取两者中心:

$$\begin{cases} C_j = \frac{1}{n_j} \sum_{x_i \in X_j} x_i \\ \overline{C_j} = \frac{1}{\overline{n_j}} \sum_{\overline{x_i} \in \overline{X_j}} \overline{x_i} \\ C_j = \frac{1}{2} (C_j + \overline{C_j}) \end{cases} \quad (8)$$

(2) 计算各类中心之间的距离:

$$d_{ij} = \|c_i - c_j\| \quad (9)$$

(3) 对于类划分方案 k , 分别计算出所有正反例的划分方案下,正例类集合 S_1 里各分类到反例集合 S_2 各分类的平均类中心距离 D_{S1S2} , 以及正反例集合内部各分类的之间的平均中心间距 D_{S1}, D_{S2} .

$$\begin{cases} D_{S1S2}^k = \frac{1}{N_{S1} \times N_{S2}} \sum_{i \in S1, j \in S2} d_{ij} \\ D_{S1}^k = \frac{2}{N_{S1} \times (N_{S1} - 1)} \sum_{i \in S1, j \in S1, i \neq j} d_{ij} \\ D_{S2}^k = \frac{2}{N_{S2} \times (N_{S2} - 1)} \sum_{i \in S2, j \in S2, i \neq j} d_{ij} \end{cases} \quad (10)$$

(4) 根据全部类间距最大的方案则是最佳方案的原则来计算出下面判别式中的最大值:

$$D^k = D_{S1S2}^k + \frac{D_{S1}^k \times N_{S1} + D_{S2}^k \times N_{S2}}{N_{S1} + N_{S2}} \quad (11)$$

通过以上的划分算法,自顶向下对各个结点类的划分方案进行确定,直到构成一棵完整的决策树,再分别对各个节点上的 SVM 分类器实行分类训练,最终便可得到一个完整的基于 SVM 决策树的多类分类器.

4 实验

4.1 实验数据

本文所采用的实验数据为某省审计厅 2010 至

2013 年度正式发文的审计报告 850 篇(不含下属审计部门数据),共计 13 357 条审计问题记录,同时以审计署所推广的“审计统计管理软件”中所保存的台账数据作为训练集。台账与审计报告具有紧密的联系,每份台账以手工的形式保存了审计报告中的重要信息,对于审计问题,前者以标题编号的形式与审计报告中的每个问题一一对应,并进行了人工分类。

审计问题共分为 4 大类,训练集中分类相对不均匀,其中,管理不规范问题为 5 797 条,占 43.4%,违纪违规问题由于其性质恶劣仅有 1 107 条记录,占 8.28%。问题记录列表分布情况如表 2 所示。

表 2 问题记录列表分布情况

类别名	记录条数	类别名	记录条数
违纪违规	1 107	管理不规范	5 797
损失浪费	3 501	非金额计量问题	2 952

4.2 分类性能评估指标

在本文中,选取常用的准确率(Precision),召回率(Recall)和 F 值作为实验评价指标,假设 a 为真正属于类 C_i 而被分类为类 C_i 的文档数, b 为属于类 C_i 而被分类为非类 C_i 的文档数, c 为不属于类 C_i 而被分类为类 C_i 的文档数, d 为不属于类 C_i 而被分类为非类 C_i 的文档数,那么则有 p (准确率)和 r (召回率)的定义如下式所示:

$$r = \frac{a}{a+c} \times 100\%; p = \frac{a}{a+b} \times 100\% \quad (12)$$

F 值为关于 r 和 p 的函数,是二者的调和平均值,具体如式(13)表示:

$$F(r, p) = \frac{2rp}{r+p} \times 100\% \quad (13)$$

4.3 实验结果

为了验证改进的分类算法的实用性,以及探究改进算法的提升空间,进行以下实验。

实验 1 改进算法的实用性

对所使用的实验数据分别按改进的算法以及不使用同义词的分类算法做比对,将实验数据以 7:3 的比率划分为训练集和测试集,实验结果如表 3 和表 4 所示。

表 3 基于不使用同义词的分类效果 %

	召回率	准确率	F 值
违纪违规	64.345	79.434	71.098
管理不规范	96.552	72.848	83.048
损失浪费	71.232	86.358	78.069
非金额计量问题	64.407	92.683	76.000

表 4 基于改进算法的分类效果 %

	召回率	准确率	F 值
违纪违规	78.493	86.434	82.272
管理不规范	95.353	82.441	88.428
损失浪费	80.362	87.846	83.937
非金额计量问题	75.617	93.272	83.522

如表 4 所示,改进的分类结果(表 4)与不使用同义词的分类结果(表 3)在召回率和准确率上均有小幅度提升,均达到较为满意的准确率,初步具有实用性。其中“违纪违规”、“非金额计量问题”的召回率,以及“管理不规范”的准确率改善较为明显,其他项改善不大,这是由于在本次具体的分类中,“管理不规范”分类样本较多,其关键字较多,随着样本的不断增多,可罗列的同义词随之缓慢增加,从而提高了准确度;而召回率增幅较小且存在下降的情况,部分原因是由于正例样本存储错分类样本,这些错分类样本在同义词扩展后对分类结果产生了负面影响。

实验 2 改进算法的有效性

由实验 1 可知,改进的方法对文本的分类结果存在一定影响,但是并不能确定改进的方法对分类的影响是否存在偶发性,通过选择不同样本数量的训练数据重复实验 1,来确定改进的方法对分类结果的影响的有效性。以平均 F 值(多个分类的 F 值均值)做为对比数据,测试数据均为实验数据集中特定的 20% 数据。图 4 所示为不同样本数量(实验数据集不同占比)下的实验结果。

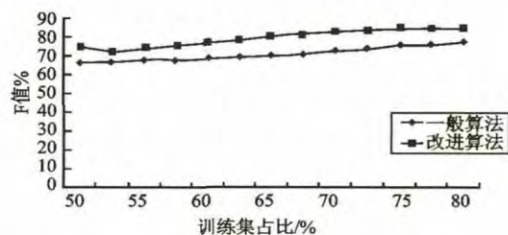


图 4 不同样本数量下两种方法下的 F 值趋势

由图 4 可知,改进算法的 F 值随样本容量的增加成缓慢上升趋势,并处于 75%~82% 区间段,部分下降点原因在于样本较少时,某一分类效果不理想对均值的影响,从总体趋势而言,改进算法相较于一般算法,确实存在提升效果,表明该算法是有效性的。

综合实验 1 和实验 2 结果,将审计领域的同义词进行特征扩展并结合部分规则的方式在审计问题的分类的具体应用上确实存在一定的提升,然而,从

实验 2 来看, F 值的平均值未达到 90% 以上的较高水平, 也表明该算法存在提升的空间. 一是由于本文仅使用了少有的几条规则, 如将问题定性、法律法规等作为关键维度, 并没有考虑从语义和最大匹配度的角度将与已分文档相似的文件直接分类, 同时忽略了文本相似度对分类的影响; 二是由于部分样本中因为分类界限模糊而存在错误样本, 这些错误样本对分类的准确度、召回率产生了破坏性影响; 三是在分类过程中强调规则和同义词对分类的影响, 并未制衡关键维度与非关键维度的影响程度. 以上几点将在以后的工作中进行优化, 使分类效果得到进一步的提升.

5 结束语

本文将信息抽取, 文本分类技术应用于处理审计报告之中, 提出的面向审计领域的新方法, 有效克服了审计报告中的出现的样本高度不均衡、短文本特征稀疏和问题分类界限模糊对分类性能的影响, 实现了将审计问题按性质进行分类的目的, 获得了较好的准确率和召回率^[10-12].

该方法在处理审计报告的问题分类上仍具有改进的空间, 下一步工作将考虑结合相似度算法与基于规则方法来计算短文本中是否含有指定的规则, 或含有与规则相似度较高的短句, 加强规则对分类结果的影响, 进一步提高分类效果.

参考文献:

- [1] 苏金树, 张博锋, 徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报, 2006(17): 1848-1859.
- [2] 李惠娟, 高峰, 管晓宏, 等. 基于贝叶斯神经网络的邮件过滤方法[J]. 微电子学与计算机, 2005, 22(4): 107-111.
- [3] 吕震宇, 林永民, 赵爽, 等. 基于同义词词林的文本特征选择与加权研究[J]. 情报杂志, 2008(5): 130-132.
- [4] 宁亚辉, 樊兴华, 吴渝. 基于领域本体的短文本分类[J]. 计算机科学, 2009, 36(3): 142-145.
- [5] 王细薇, 樊兴华, 赵军. 一种基于特征扩展的中文短文本分类方法[J]. 计算机应用, 2009, 29(3): 843-845.
- [6] 朱远平, 戴汝为. 基于 SVM 决策树的文本分类器[J]. 模式识别与人工智能, 2005, 18(4): 412-416.
- [7] 田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报, 2010, 28(6): 602-608.
- [8] 陈沈焰, 吴军华. 基于本体的概念语义相似度计算及其应用[J]. 微电子学与计算机, 2008, 25(12): 96-99.
- [9] 李生琦, 天巧燕, 汤承. 基于《知网》词汇语义相关度计算的消歧方法[J]. 情报学报, 2009, 28(5): 706-711.
- [10] James G Shanahan, Norbert Roma. Boosting support vector machines for text classification through parameter-free threshold relaxation[C]// CIKM'03. New Orleans, 2003.
- [11] Joachim T. Text categorization with support vector machines; Learning with many relevant features [C] // Proceedings of the 10th European Conference on Machine Learning, Spain; Barcelona, 1998.
- [12] Takahashi F, Abe S. Decision-tree-based multiclass support vector machines[C]// Proceeding of the 9th International Conference on Neural Information Processing, Singapore, 2002.

作者简介:

伍 洋 男, (1989-), 硕士研究生. 研究方向为信息抽取、数据挖掘. E-mail: angenimy@sina.com.

钟 鸣 男, (1982-), 博士, 讲师. 研究方向为数据库理论与系统、图形数据库、大数据、信息检索.

姜 艳 女, (1983-), 硕士研究生. 研究方向为审计信息化.

李石君 男, (1964-), 博士, 教授. 研究方向为 Web 搜索与挖掘、数据挖掘、大数据.

(上接第 4 页)

- [11] Lamech C, Rad R, Tehranipoor M, et al. An experimental analysis of power and delay signal-to-noise requirements for detecting trojans and methods for achieving the required detection sensitivities [J]. IEEE Transactions of Information Forensics and Security, 2011, 6(3): 142-155.

- [12] 刘长龙, 赵毅强. 基于侧信道分析的硬件木马建模与优化[J]. 华中科技大学学报, 2013, 41(2): 53-57.

作者简介:

苏 静 女, (1979-), 博士研究生, 副教授. 研究方向为信息安全、机器学习. E-mail: sjnicelife@163.com.