容错粗糙模型的事件检测研究

毋 非[1],封化民[1,2],申晓晔[1]
(1. 西安电子科技大学 通信工程学院,陕西 西安 710071; 2. 北京电子科技学院 多媒体智能处理实验室,北京 100070)

摘 要:对网站发布的 Web新闻内容进行必要的、合理的监督管理,是保障网络信息内容安全的重要研究内容.将现有的文本表示模型应用于 Web新闻会导致文本表示的稀疏性问题和话题跟踪过程中的主题词漂移问题,一种基于容错粗糙集的文本表示模型解决了这些问题.在理论分析和实验验证的基础上,结合向量空间模型 (VSM),利用特征项在文档集中协同出现,构造了特征项的容错粗糙集.然后用特征项容错粗糙集生成文档的容错粗糙模型,来扩充原先的文档表示模型.最后用特征项容错类描述文档之间的相似性关系,实现事件检测过程.实验结果证明,容错粗糙模型能够改进事件检测系统的性能.

关键词:事件检测;粗糙集;容错粗糙模型
中图分类号: TP391 **文献标识码**: A **文章编号**: 1673-4785 (2009) 02-0112-06

Research on event detection based on the tolerance rough set model

WU Fei[1], FENG Hua-min[1,2], SHEN Xiao-ye[1]
(1. School of Telecommunication Engineering, Xidian University, Xi'an 710071, China; 2. Multimedia Intelligent Information Processing Laberatory, Beijing Electronic Science and Technology Institution, Beijing 100070, China)

Abstract: Proper monitoring of the content of web news is crucial to the maintenance of network content security. Current text representational models are not suitable for web news because of the sparseness of text representation and the drifting of key words in event tracking processes. To solve these problems, a modeling method for text representation based on tolerance rough sets was used to extend text representation. Following theoretical analysis and experimental verification, we constructed a tolerance rough set for feature terms by considering the vector space model (VSM) and the co-occurrences of feature terms in test sets. Then the tolerance rough set model of tests was generated using the tolerance rough set for feature terms, which extended the original text representation model. Finally, the similarities of texts were described by the feature term's tolerance classes. Experimental results showed that the tolerance rough set model improved the performance of event detection systems.

Keywords: event detection; rough set; tolerance rough set model

随着网络技术的迅速发展,越来越多的人选择通过网络渠道来表达自己的想法.互联网逐渐成为舆情产生和传播的重要场所,网络舆情在当前的社会生活中扮演着重要角色.对网络舆情的监控、分析和管理成为各级政府部门亟待解决的问题.话题检测与跟踪 (topic detection and tracking, TDT)技术作为舆情分析的重要技术手段,已成为近几年信息检索领域的热点研究课题.其主要任务是在以新闻专线和广播新闻等为来源的数据流中自动发现话题,并把话题相关的内容联系在一起[1].事件检测和事件跟踪是 TDT的 2个重要子任务.从本质上看,两者都是将新闻报道流进行聚类.

现有的系统在事件检测方面大都采用了以下步骤: 1)建立报道和事件的文本表示模型; 2)采用某种算法计算报道与事件,或者是报道与报道之间的相似度,确定与当前报道最相似的事件; 3)若报道被归入某事件,则调整该事件的表示模型,若报道没有归到现有的任何事件,则认为它是新检测到的事件; 4)输出检测到的事件中权重最高的几个特征

收稿日期: 2008-12-16.
通信作者:毋 非. E-mail: wuf@besti.cn.

词、或者具有代表性的标题作为事件描述.本文主要讨论事件的文本表示模型.

目前应用到话题检测中的文本模型主要有 2种,一种是基于向量空间模型(vector space model, VSM)的方法[2],一种是基于概率模型(probability model, PM)的方法[3],两者各有优缺点.其中向量空间模型一直是应用的主流,因为它易于将文本转化为向量,使得文本之间的相互计算成为可能.但是,向量空间模型局限于文本之间相互独立的假设[4],使得文本在向量空间转换的过程中丢失了关联信息;而概率模型有扎实的理论基础,发展潜力较大,但是由于 TDT中的新闻报道通常都比较短小,使概率模型原本就存在的稀疏问题更加严重.

文本表示模型建立和使用方法的优劣,会在很大程度上影响整个系统的性能.本文在向量空间模型的基础上,在文本表示模型中引入了粗糙集,利用容错粗糙集表示文档之间的关联信息.通过在中等规模的 Web文档数据集上的实验,可以证明,使用特征项容错粗糙集建立的文本表示模型,可以有效地改进系统性能.

1　容错粗糙集模型

1.1　粗糙集和容错粗糙集的概念

粗糙集理论(rough set theory)是 1982年由波兰数学家 Z. Pawlak提出的,它提供了一种特殊的处理不确定性的方法.

粗糙集理论的中心观点就是集合的近似表示[5-6].设非空有限对象集合 U为论域,则若要在 U上定义一个概念,那么这个概念可以由 U的子集 X表示,即任何在 U上的集合概念,都能用它的 X近似集合和上近似集合表示.

设 R是集合 U上的二元关系,如果它是自反的、对称的和传递的(即具有:自反性, xRx;对称性, $xRy\rightarrow yRx$;传递性, $xRy\wedge yRz\rightarrow xRz$, $\forall x, y\in U$),则它是 U上的等价关系.对于 $x, y\in U$,如果 xRy,那么称 x、y是相互不可分辨的.关系 R可将 U完全划分成等价类 $[x]_R$, $x\in U$,即 $U=\cup_{x\in U}[x]_R$成立, $[x]_R$表示在等价关系 R下,与 x相互不可分辨的等价类对象.

定义对于近似空间 $A=(U, R)$, $x\subseteq U$, X的上近似和下近似集合如下:

$$U_R(X)=\{x\in U\mid [X]_R\cap X\neq\Phi\},\quad (1)$$

$$L_R(X)=\{x\in U\mid [x]_R\subseteq X\}.\quad (2)$$

直观上看, X的近似集合所包含的对象肯定属于 X,而上近似包含的对象则可能属于 X,也可能不属于 X. (L_R, U_R)表示了概念 X的一种粗糙近似关系,即粗糙集[7].

在早期的信息检索领域,粗糙集的应用都是基于上面这种"等价粗糙模型"(equivalence rough set model, ERSM).该模型基于以下假设:特征项集合 T可以根据等价关系 R划分为等价类.但是,在信息检索领域,当处理的对象是词语等文本信息时,等价关系的 3个属性中,传递性并不是总能保证的.因为词语的含义是分离的,但是又可能相互重叠,而且它们的含义不符合传递性.

这种相互重叠的类可以由容错关系(tolerance relation)生成,容错关系只要求自反性和对称性.文献[8]使用容错关系定义了容错空间来表示这种叠交的类,即容错类(tolerance class).容错关系可以用一个四元组表示 $R=(U, I, v, P)$, U为对象的集合, $I: U\rightarrow 2^U$是不确定函数(uncertainty function), $v: 2^U\times 2^U\rightarrow[0, 1]$是模糊包含(vague inclusion), $P: I(U)\rightarrow\{0, 1\}$是结构函数(structurality function).

其中,通过不确定函数 $I: U\rightarrow 2^U$可以找到所有对象中与 x具有相似含义的对象,即 x的容错类,用 $I(x)$来表示.不确定函数的定义为:对任意 $x, y\in U$,有 $x\in I(x)\wedge y\in I(x)\Leftrightarrow x\in I(y)$.该函数符合以下关系: $\Gamma\subseteq U\times U$, $x\Gamma y\Leftrightarrow I(x)$. Γ是一个容错关系,满足自反性和对称性.

模糊包含 $v: 2^U\times 2^U\rightarrow[0, 1]$用来度量集合的包含程度.特别地,它用来度量一个对象 $x\in U$的容错类 $I(x)$是否被包含在一个集合 X中.对于 v,只对第 2个参数有单调性要求,即 $\forall X, Y, Z\subseteq U\wedge Y\subseteq Z$,有 $v(X, Y)\leq v(X, Z)$.

在构造下、上近似集的过程中,只考虑结构化的粗糙集元素.现定义 $P: I(U)\rightarrow\{0, 1\}$将所有 $x\in U$的 $I(x)$分为 2类:结构化子集($P(I(x))=1$)和无结构子集($P(I(x))=0$).

则对于任意 $x\in U$在容错空间 R上的下近似集和上近似集定义如下:

$$L_R(X)=\{x\in U\mid P(I(x))=1\&v(I(x), X)=1\},\quad (3)$$

$$U_R(X)=\{x\in U\mid P(I(x))=1\&v(I(x), X)>0\}.\quad (4)$$

这样,只要定义恰当的 I、v和 P,就可以在具体应用中使用容错空间.

1.2　容错粗糙模型的建立

下面讨论如何在事件检测的应用中确定 I、v和 P.

设容错空间为 R,选取文档集的所有特征项作 $T=\{t_1, t_2, t_3, \cdots\}$为论域.

采用文档集中特征项协同出现来确定特征项 t_i 的容错类 $I(t_i)$. 因为它较好地解释了上下文的语义从属关系,而且相对简单,计算上也是可行的. 设 $f_D(t_i, t_j)$表示特征项 t_i和 t_j在文档集 D中协同出现的次数. 则以 θ为阈值的不确定函数 I定义为

$$I_\theta: I_\theta(t_i) = \{t_j \mid f_D(t_i, t_j) \geq \theta\} \cup \{t_i\}. \quad (5)$$

显然,函数 $I_\theta(t_i)$满足以下条件:若 $t_i, t_j \in T$, $t_i \in I_\theta(t_j)$,则有 $t_i \in I_\theta(t_i)$和 $t_j \in I_\theta(t_i)$成立,即 I_θ是自反的和对称的. 这个函数符合容错关系 $\Gamma \subseteq T \times T$,其中 $t_i \Gamma t_j \Leftrightarrow t_j \in I_\theta(t_i)$. 通过改变阈值 θ的大小,可以控制容错类中特征项的相互关系程度,即可以改变容错类的精度[16].

模糊包含定义为

$$v(X, Y) = \frac{|X \cap Y|}{|X|}. \quad (6)$$

显然,该函数的第 2个参数是单调的. 基于这个函数,对于 $t_i \in T, X \subseteq T$的隶属于函数 μ可定义为

$$\mu(t_i, X) = v(I_\theta(t_i), X) = \frac{|I_\theta(t_i) \cap X|}{|I_\theta(t_i)|}. \quad (7)$$

假设特征项集 T在整个处理过程中是封闭的,在这个假设条件下,可以把所有特征项的容错类看作是结构化的子集,即对于任意 $t_i \in T$, $P(I_\theta(t_i)) = 1$.

从以上定义,可以得到在容错空间 $R = (T, I, v, P)$上,文档 $d_i \in D$的下近似集和上近似集分别为:

$$L_R(d_i) = \{t_j \in T \mid v(I_\theta(t_j), d_i) = 1\}, \quad (8)$$

$$U_R(d_i) = \{t_j \in T \mid v(I_\theta(t_j), d_i) > 0\}. \quad (9)$$

这样,文档 d_i就可以用它的近似集来表示,其中,下近似集 $L_R(d_i)$表示 d_i的“核心”,上近似集 $U_R(d_i)$表示与 d_i的特征项有交叉语义的特征项的集合. 于是,可以使用 $U_R(d_i)$来表示特征项的容错类,建立文档的容错粗糙模型.

1.3 特征项容错类生成算法

如前所述,特征项容错类就是协同出现的相关特征项的集合,用特征项协同关系矩阵 TOL = $[tol_{x,y}]_{M \times M}$表示,具体算法如下:

算法 1 产生特征项容错类的算法

输入:文档特征项频率矩阵 TF,协同阈值 θ;

输出:特征项容错类的二值矩阵 TOL;

1)对文档特征项频率矩阵 TF进行二值化,生成特征项的二值矩阵 OC:

$$OC = [oc_{i,j}]_{N \times M},$$

$$oc_{i,} = \begin{cases} 1, & tf_{i,j} > 0; \\ 0, & \text{其他}. \end{cases} \quad (10)$$

即在特征项的二值矩阵 OC中,每行表示特征项在一个文档中是否出现,若出现则该列置 1,否则置 0.

2)建立特征项协同出现矩阵 COC:

$$COC = [coc_{x,y}]_{M \times M},$$

$$coc_{x,y} = \text{card}(OC^x \ AND \ OC^y). \quad (11)$$

式中:OC^x、OC^y表示 OC矩阵中特征项 x、y的列向量; card表示向量的基; $coc_{x,y}$表示特征项 x、y的协同发生频率,即在整个文档集中,特征项 x、y协同出现的次数.

3)给定协同出现阈值 θ,在 COC矩阵中过滤数值小于 θ的特征项,就得到特征项容错二值矩阵:

$$TOL_{x,y} = [tol_{x,y}]_{M \times M},$$

$$tol_{x,y} = \begin{cases} 1, & coc_{x,y} \geq \theta; \\ 0, & \text{其他}. \end{cases} \quad (12)$$

矩阵的每行给出了一个特征项的容错类, $tol_{x,y}$置 1表示特征项 x、y存在容错关系.

以下实例较好地说明了容错类的含义. 选取 4篇文档 d_1、d_2、d_3、d_4,每篇文档用 10个特征项 t_i来表示,例如, t_1表示“日本”, t_2表示“地震”,等等. 则当阈值 $\theta = 2$时,可得到以下特征项的容错类: $I_2(t_1) = I_2(t_2) = I_2(t_{18}) = \{t_1, t_2, t_{18}\}$,于是得到各个文档的上近似集如表 1所示.

表 1 4篇文档的上近似集表示

Table 1 Upper approximations of 4 documents

文 档	特征项集	文档的上近似集
d_1	$t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}$	$t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}, t_{18}$
d_2	$t_1, t_{11}, t_{12}, t_2, t_{13}, t_{14}, t_{15}, t_{16}, t_{17}, t_{18}$	$t_1, t_{11}, t_{12}, t_2, t_{13}, t_{14}, t_{15}, t_{16}, t_{17}, t_{18}$
d_3	$t_1, t_{19}, t_{20}, t_{21}, t_2, t_{22}, t_{18}, t_{14}, t_{23}, t_{24}$	$t_1, t_{19}, t_{20}, t_{21}, t_2, t_{22}, t_{18}, t_{14}, t_{23}, t_{24}, t_{18}$
d_4	$t_1, t_{25}, t_{26}, t_{27}, t_{28}, t_{29}, t_{30}, t_{31}, t_{32}, t_{33}$	$t_1, t_{25}, t_{26}, t_{27}, t_{28}, t_{29}, t_{30}, t_{31}, t_{32}, t_{33}, t_2, t_{18}$

1.4 特征项权重计算

计算特征项权重的基本方法涉及到两方面的修改. 一方面,由于事件检测的文档集是动态增加的,因此要使用增量 TF-IDF模型[9]. 即每经过一个时间

窗口更新一次模型,在一个时间窗内的更新方式如下:

$$f_{D'}(t_i) = f_D(t_i) + f_{D_k}(t_i). \tag{13}$$

式中:D_k 表示窗口内的文档集,$f_{D_k}(t_i)$表示在窗口内特征项 t_i的文档频率,$f_{D'}(t_i)$为更新之后的文档频率.另一方面,由于在文档表示中使用了特征项的上近似集 $U_R(d_i)$,因此在计算特征项权重时,还需要考虑那些出现在文档上近似集中,但不出现在文档中的特征项.对于这些特征项,由于它本身并没有出现在文档中,因此需要使其权值小于任意一个出现在 d_i中的特征项的权值.于是,使用以下扩展权值计算公式替换一般的 TF-DF公式[10-11].

$$w_{ij} = \begin{cases} (1 + \ln(f_{d_i}(t_j))) \times \ln\dfrac{N}{f_D(t_j)}, & t_j \in d_i; \\ \min_{t_k \in d_i} w_{ik} \times \dfrac{\ln\frac{N}{f_D(t_j)}}{1 + \ln\frac{N}{f_D(t_j)}}, & t_j \in U_R(d_i)/d_i; \\ 0, & \text{其他}. \end{cases} \tag{14}$$

式中:w_{ij}是 t_j在文档 d_i中的权值.此计算公式用 d_i中特征项的权值的最小值,乘以一个小于 1的数,确保了出现在 d_i的上近似集中但没有出现在 d_i中的特征项权值,小于任何 d_i中的特征项权值.

文档向量的权值 w_{ij}的归一化方法为

$$w_{ij} = \frac{w_{ij}}{\sqrt{\sum_{t_k \in U_R(d_i)} (w_{ij})^2}}. \tag{15}$$

1.5 相似度计算

选取相似度计算函数的一个重要标准就是该函数能否区分描述相同事件和描述不同事件的新闻报道对.基于向量的相似度计算方法有多种,如余弦相似度、Hellinger相似度等,文献[12]指出,余弦相似度性能最好,也最稳定.因此采用余弦相似度作为向量相似度计算函数.

$$\mathrm{Sim}(d_i, d_j) = \frac{\sum_{k=1}^{M} W_{ik} \times W_{jk}}{\sqrt{(\sum_{k=1}^{M} W_{ik}^2) \times (\sum_{k=1}^{M} W_{jk}^2)}}. \tag{16}$$

2 实验及分析

2.1 评测标准

本文依据 TDT评测标准,采用漏报率 (Miss_i)、误报率 (FA_i)以及归一化开销 $(C_{\mathrm{Det}})_{\mathrm{Norm}}$[13]来评价该检测方法的性能,话题 i的漏报率和误报率定义为

$$\mathrm{Miss}_i = \frac{\text{未检没到的与话题 } i\text{相关的报道数}}{\text{与话题 } i\text{相关的报道总数}}, \tag{17}$$

$$\mathrm{FA}_i = \frac{\text{检测到的与话题 } i\text{不相关的报道数}}{\text{与话题 } i\text{不相关的报道总数}}. \tag{18}$$

则系统的平均漏报率 P_{Miss}、平均误报率 P_{FA}和归一化检测开销 $(C_{\mathrm{Det}})_{\mathrm{Norm}}$如下所示:

$$P_{\mathrm{Miss}} = \sum_i \mathrm{Miss}_i / t_n, \tag{19}$$

式中:t_n为话题个数;

$$P_{\mathrm{FA}} = \sum_i \mathrm{FA}_i / t_n, \tag{20}$$

式中:t_n为话题个数,

$$(C_{\mathrm{Det}})_{\mathrm{Norm}} = \frac{C_{\mathrm{Miss}} P_{\mathrm{Miss}} P_{\mathrm{target}} + C_{\mathrm{FA}} P_{\mathrm{FA}} P_{\neg\,\mathrm{target}}}{\min(C_{\mathrm{Miss}} P_{\mathrm{target}}, C_{\mathrm{FA}}, P_{\mathrm{target}})}. \tag{21}$$

式中:$(C_{\mathrm{Det}})_{\mathrm{Norm}}$越小,表明系统性能越好,理想情况下,$(C_{\mathrm{Det}})_{\mathrm{Norm}} = 0$;$C_{\mathrm{Miss}}$为漏报一个新话题的代价;$C_{\mathrm{FA}}$为误报一次的代价;$P_{\mathrm{target}}$是目标话题的先验概率;$P_{\neg\,\mathrm{target}} = 1 - P_{\mathrm{target}}$;$C_{\mathrm{Miss}}$、$C_{\mathrm{FA}}$和 P_{target}都是预设值,不同的评测中取值不一样,本文中它们的取值分别为 1.0、0.1、0.02.

2.2 文档预处理

预处理的内容主要包括按照一定规则生成特征项.传统的做法是根据文档的词语频率,去掉一定阈值内的高频词和低频词,去掉停用词之后形成的集合作为特征项集合.但是由于在事件检测过程中,需要使用增量 TF-DF方法动态更新文档频率,这样,使用传统的特征项选择方法在计算上会比较复杂.

本文提出了使用词频词典来选择特征项的方法.对于每篇新闻报道,使用哈工大信息检索研究室语言技术平台共享包[14],进行分词和词性标注.将标注好的文档中的名词提取出来,作为特征候选集合,然后根据词频词典,生成相应的向量.若词典中的词在文档中出现,则在向量中标记该词出现的次

数,若没有出现在文档中,则标记为 0. 这样,每篇文档都表示成为统一维数的向量,整个文档集被表示为特征项矩阵 TF. 在增量计算的过程中,可以方便地计算词语在所有文档中出现的次数.

词频词典是搜狐研发中心[15]提供的,该词典为 2006年 10月统计的互联网词库,涉及语料规模在 1亿页面以上. 本文选取了词频在 100万以上的高频名词作为实验使用的词频词典,共 7 338个词.

2.3 实验语料集

实验语料是搜狐研发中心 2006年 11月提供的中文互联网语料[15],随机选取其中的 1 500篇,其中 1 000篇作为容错粗糙集的训练集, 500篇作为测试集. 人工标注话题 15个,将实验所得的话题聚类结果与人工标注话题相比较,得到最后的实验结果.

2.4 实验结果及分析

实验过程中使用了迭代 Single-Pass算法进行事件聚类. 为了同一般的 VSM作比较,还实现了 VSM下的实验结果. 为了探讨容错类精度 θ对检测结果的影响,实验中取了 4个 θ的值进行对比,结果如表 2所示.

表 2 实验结果表

Table 2 Experiment results

θ	P_{Miss}	P_{FA}	$(C_{Det})_{Norm}$
3	0.352 9	0.037 54	0.502 9
8	0.208 3	0.008 97	0.244 2
15	0.218 8	0.004 17	0.235 5
25	0.229 2	0.004 81	0.248 4
VSM	0.273 8	0.023 73	0.368 7

从上表可以看出,当容错类精度较小时 (例如 θ=3时),由于生成的容错类中特征项的数量很大,因此漏检率和错检率都很高,甚至比向量空间模型高出很多. 而随着 θ取值的增大,特征项容错类中词语的个数在减少,可以找到一个比较恰当的范围 (如 $\theta \in [15, 25]$). 在这个范围内,漏检率、错检率和识别代价都达到一个稳定的水平,且总体效果优于使用 VSM的识别结果,归一化识别代价最多降低了 31.2%.

但是,容错粗糙模型对漏检率的提高十分有限,经过分析,认为原因是容错类建立的标准有待改进. 本文在生成容错类时,仅考虑了文本中的名词,而新闻报道中的实体词 (地点、时间)和动词在事件框架的构成中占有重要地位,过滤掉这些词后,文档的特征项对文档的描述准确性会降低. 同时,新闻报道本身也存在一定的模糊性,有一些报道无论是人工还是机器都很难判断它们是否属于同一个话题.

另外,使用词频词典作为特征项选择方法虽然减少了增量 TF-DF的计算量;但是也增加了特征项矩阵的稀疏性,这在一定程度上削弱了容错粗糙模型的优势. 但仍给出一个启示:词典可以经过进一步的降维处理,同时,可以将词典扩充为带权值的词典,对于某些具有特殊意义的词可以提高权值. 在内容安全的具体应用中,可以根据不同的需求,生成不同的领域词典,是实现热点新闻追踪的一个有效手段.

实验结果还表明,容错精度较大时 (θ>3时),容错粗糙模型有效地降低了系统的错检率.

3 结束语

事件检测是话题检测与跟踪的核心任务,目前围绕该任务的检测方法有很多种. 本文采用了基于特征项协同出现的容错粗糙模型来建立事件检测框架,并对比了该框架使用向量空间模型的结果. 实验结果表明,应用该模型降低了检测代价,改进了系统性能. 但是本文的方法仅仅考虑了名词,在未来工作中还应该加入更加丰富的文档表示方法. 另外,本文尝试使用词频词典来选择特征项的方法,也有待进一步改进.

参考文献:

[1] ALLEN J, CARBONELL J, DODDINGTON G, YAMRON J, YANG Y. Topic detection and tracking pilot study: final report[C]//Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. Virginia: Lansdowne, 1998: 194-218.

[2] CONNELL M, FENG A, KUMARAN G, et al. UMass at TDT 2004[C]//The 7th Topic Detection and Tracking Conference. Gaithersbury, USA, 2004: 35-41.

[3] NALLAPATI R. Semantic language models for topic detection and tracking[C]//Proceedings of HLT-NAACL 2003 Student Research Workshop. Edmonton, CA, 2003: 1-6.

[4] 苏新宁. 信息检索理论与技术[M]. 北京:科学技术文献出版社, 2004: 33-35.

[5] PAWLAK Z. Rough sets: theoretical aspects of reasoning about data[M]. Dordrecht: Kluwer Academic Publishers, 1991: 9-27.

[6] KOMOROWSKI J, POLKOWSKI L, ANDRZEJ S. Rough sets: a tutorial, a new trend in decision-making[M]. Singapore: Springer, Singapore Pte Ltd, 1998: 2-5.

[7] 刘 清. Rough集及 Rough推理[M]. 北京:科学出版社, 2003: 11-13.

[8] SKOWRON A, STEPANIUK J. Generalized approximation spaces[C]// 3rd International Workshop on Rough Sets

and Soft Computing[s l], 1994: 156-163.
[9] YANG Y, PIERCE T, CARBONELL J. A study on retrospective and on-line event detection[C]// Proc of the SIGIR '98. Melbourne, 1998: 28-36.
[10] BAO HO T, BINH NGUYEN N. Nonhierarchical document clustering based on a tolerance rough set model[J]. International Journal of Intelligent Systems, 2002, 17 (2): 199-212.
[11] LANG N C. A tolerance rough set approach to clustering web search results[D]. Warsaw: Warsaw University, 2003.
[12] YANG Y, CARBONELL J, JIN C. Topic-conditioned novelty detection[C]//Proceeding of the 8th ACM SIGKDD. New York: ACM Press, 2002: 688-693.
[13] The 2003 topic detection and tracking (TDT2003) task definition and evaluation plan [EB/OL]. [2003-04-21]. http://www.nist.gov/speech/tests/tdt/tdt2003/evalplan.htm.
[14] 哈工大信息检索研究室. 语言技术平台共享包 [EB/OL]. [2008-06-12]. http://ir.hit.edu.cn/.
[15] 搜狗实验室. 互联网语料库 2006版 [EB/OL]. [2008-06-12]. http://www.sogou.com/labs/.
[16] 易高翔, 胡和平. 一种基于容错粗糙集的 Web搜索结果聚类方法 [J]. 计算机研究与发展, 2006, 43(2): 275-280.
YI Gaoxiang, HU Heping. A web search result clustering based on tolerance rough set [J]. Journal of Computer Research and Development, 2006, 43(2): 275-280.

作者简介:

毋　非,女, 1984年生,硕士研究生. 主要研究方向为 Web新闻内容安全、信息检索.

封化民,男, 1963年生,教授,硕士生导师. 主要研究方向为多媒体智能信息处理、网络安全.

申晓晔,女, 1984年生,硕士研究生. 主要研究方向为 Web新闻内容安全、舆情倾向性分析.

第 48届 IEEE决定与控制大会和第 28届中国控制大会
48th IEEE Conference on Decision and Control and 28th Chinese Control Conference

The combined 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference will be held during the third week of December, 2009 at a location in China. This will be the third time that CDC has been held outside the United States in the Asia-Pacific region, and it is very fitting that it is being held in China, where numbers of new IEEE members are increasing so rapidly. China today is one of the most dynamic and exciting countries in the world. With a thriving economy, huge recent investment in education and research, together with an extraordinarily rich history of culture, it is the ideal location for CDC.

The annual IEEE Conference on Decision and Control (CDC) is internationally recognized as the premier scientific and engineering conference dedicated to the advancement of the theory and practice of systems and control. The CDC brings together an international community of researchers and practitioners to discuss new research results, perspectives on future developments, and innovative applications relevant to decision making, automatic control, and related areas.

Papers are invited in the form of regular manuscripts (allotted 6 Proceedings pages). Note that short manuscripts are not considered. Papers must be submitted through the conference submission website (PaperPlaza) and must conform to the submission policy requiring that all manuscripts be in 2-column format and meet strict page limits. For the purpose of review only, manuscripts may be up to eight (8) pages long. However, normal length for the final manuscript is limited to six (6) pages. Papers exceeding the normal length may be submitted upon payment of overlength page charges of USD 175.00 for each page in excess of six. A maximum of two extra pages above normal six are permitted for regular papers and invited session papers.

Web site: http://www.ieeecss.org/CAB/conferences/cdc2009/index.php?page=timetable