基于 CRFs 模型的敏感话题识别研究*

翟东海[1,2], 聂洪玉[1], 崔静静[1], 杜 佳[1]
(1. 西南交通大学 信息科学与技术学院, 成都 610031; 2. 西藏大学 工学院, 拉萨 850000)

摘 要: 条件随机场(CRFs)是一种判别式概率无向图学习模型,将其引入敏感话题识别中,提出了基于 CRFs 模型的敏感话题识别方法。将随机挑选出的一篇待检测文本 s 和剩余的待检测文本分别作为 CRFs 模型的观察序列和状态序列来计算文本 s 和其余待检测文本间的相关性概率值;然后将相关性最高的那篇文本和文本 s 合并表征一个类别;同时,将相关性最低的那篇文本作为另一个类别,将这两个类别作为 CRFs 模型新的状态序列,剩余的待检测文本作为新的观察序列进行迭代,据此实现敏感话题的识别。在数据集上进行的实验中,该方法的耗费函数的值为 0.01943,宏平均 F 度量的值为 0.8235,都取得了很好的效果。

关键词: 条件随机场; 敏感话题识别; 相关性概率值
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2014)04-0993-04
doi:10.3969/j.issn.1001-3695.2014.04.008

Sensitive topic detection based on conditional random fields

ZHAI Dong-hai[1,2], NIE Hong-yu[1], CUI Jing-jing[1], DU Jia[1]
(1. *School of Information Science & Technology, Southwest Jiaotong University, Chengdu* 610031, *China*; 2. *Engineering School, Tibet University, Lhasa* 850000, *China*)

Abstract: This paper introduced the conditional random field(CRFs) to construct the sensitive topic detection method based on CRFs. Firstly, the method used a text s selected randomly as observing sequence of CRFs, and used the rest of detected text as its state sequence to calculate the correlative probabilities between the text s and the rest of detected text. Secondly, it combined the text with the highest correlative probability and text s to represent a category, meanwhile, the text with lowest correlative probability was represented other category. Therefore, these tow categories were as observing sequence of CRFs, and the rest of detected text were used as its state sequence. Thirdly, it carried out the iteration as abovementioned to detect sensitive topic. The experimental results show that the cost function and macro average F of the proposed method can achieve high effect.

Key words: conditional random field (CRFs); sensitive topic detection; correlative probabilities

0 引言

近年来,我国互联网发展迅速,越来越多的网民通过互联网表达自己的观点和想法,互联网已经成为反映民情、社会动态的一个重要平台[1]。互联网作为现代社会重要的交流渠道,其存储和传输的信息,尤其是一些敏感话题,对社会民众舆论的形成和传播有着至关重要的作用,一般情况下敏感话题的出现都给网民带来不良影响,给社会舆论的健康发展带来压力。一些不法分子利用网络的开放自由性恶意传播不法信息,严重影响了社会的稳定[1]。因此,对于网络敏感话题的识别是当今学者面临的一项紧迫而又重要的课题。

Zhang 等人[2]提出基于索引树和命名实体的新事件检测来提高话题检测的性能;薛晓飞等人[3]提出基于新闻要素的新事件检测方法能够将文本按时间、地点、人物等要点分别进行相似度计算以检测新事件;王颖颖等人[4]提出的检测系统中的性能提升策略能够通过预处理文档来提升检测性能。Alsumit[5]提出基于生成主题模式的在线话题检测可以有效检测到新话题。闵可锐等人[6]设计了一个互联网话题识别与跟踪系统,可以将网络上发布的海量文本信息聚合成各个话题,但其对获取敏感信息的能力相对较弱;鲁明羽等人[7]提出的一种基于模糊聚类的网络论坛(BBS)热点话题挖掘算法可以发现聚类中的孤立点以解决网络中单个帖子的话题多样性问题,但对于篇幅较大的文本性能较弱,并且对敏感性话题的发现效果也并不明显。本文在上述研究成果的基础上将 CRFs 模型引入敏感话题识别中,提出了基于 CRFs 模型的敏感话题识别方法。a)将随机挑选出的一篇待检测文本 s 和剩余的待检测文本分别作为 CRFs 模型的观察序列和状态序列来计算文本 s 和其余待检测文本间的相关性概率值;b)将相关性最高的那篇文本和文本 s 合并表征一个类别,同时,将相关性最低的那篇文本作为另一个类别;c)将这两个类别作为 CRFs 模型新的状态序列,剩余的待检测文本作为新的观察序列进行迭代,据此实现敏感话题的识别。在数据集上进行的实验中,该方法的代价函数以及宏平均 F 度量都取得了很好的效果。

收稿日期: 2013-07-18; **修回日期**: 2013-08-27 **基金项目**: 国家语委"十二五"科研规划资助项目(YB125-49);国家教育部科学技术研究重点资助项目(212167);中央高校基本科研业务费专项资金科技创新资助项目(SWJTU12CX096);国家级大学生创新创业训练计划资助项目(201210694017)

作者简介:翟东海(1974-),男,山西芮成人,副教授,博士,主要研究方向为海量数据挖掘、数字图像处理(dhzhai@swjtu.edu.cn);聂洪玉(1989-),女,硕士研究生,主要研究方向为海量数据挖掘;崔静静(1988-),女,硕士研究生, 主要研究方向为海量数据挖掘;杜佳(1991-),男,本科生,主要研究方向为海量数据挖掘.

1 敏感话题识别

条件随机场模型是一种判别式概率无向图学习模型,它是 Lafferty 于 2001 年在最大熵模型(maximum entropy model, MEM)和隐马尔科夫模型(hidden Markov model, HMM)的基础上提出来的,该模型克服了其他算法的标注偏置的问题,更好地拟合了真实世界,是近几年自然语言处理领域常用的算法之一[8]。CRFs 模型可以用于计算给定输入序列时指定输出的概率值,本文基于这种思想构造了一种敏感话题识别模型。该模型参照敏感词库将每一篇待检测文本用向量空间模型表示,并采用 CRFs 模型进行计算得到一系列的概率值,利用这些概率值进行敏感话题识别,其流程框图如图 1 所示。

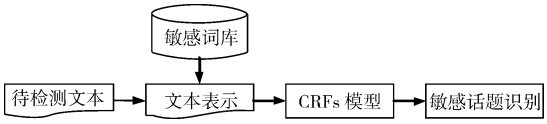

图 1 敏感话题识别流程

1.1 文本表示

网络中的所有待检测文本可以采用向量空间模型并参考敏感词库表示为 $T_n=\{t_1, w_1; t_2, w_2; \cdots; t_i, w_i\}$。其中,特征词 t_i 必须同时出现在待检测文本和敏感词库中,其权重 w_i 采用传统的 TF-IDF[9] 来计算;同时,参照文献[10]中引入了突发性因子的思想为该权重引入敏感因子 β_i:

$$w_{ni}=\frac{tf_{ni}\log\left(\frac{K}{k_i}+0.01\right)}{\sqrt{\sum_{i=1}^{t}(tf_{ni})^2\left[\log\left(\frac{K}{k_i}+0.01\right)\right]^2}}\beta_i \tag{1}$$

其中:tf_{ni} 表示第 n 篇文档中敏感词 t_i 出现的频率;K 表示总文档数;k_i 表示含有敏感词 t_i 的文档数。敏感因子 β_i 用信息增益来表示,如式(2)所示。

$$\beta_i=-P(C_m)\log P(C_m)+P(t_i|C_m)\log P(t_i|C_m)+P(\overline{t_i}|C_m)\log P(\overline{t_i}|C_m) \tag{2}$$

其中:$P(C_m)$ 表示属于第 m 类敏感话题的文本数;$P(t_i|C_m)$ 表示属于第 m 类敏感话题并包含敏感词 t_i 的文本数;$P(\overline{t_i}|C_m)$ 表示属于第 m 类敏感话题但不包含敏感词 t_i 的文本数。

1.2 CRFs 模型

在 CRFs 模型中最简单最常用的是线性链式结构(linear-chain CRFs),如图 2 所示。

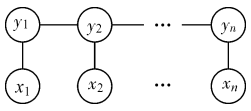

图 2 线性链式 CRFs 结构

其中,$y_1, y_2, \cdots, y_n$ 是敏感话题类别特征,$\boldsymbol{y}=\{y_1, y_2, \cdots, y_i\}$ 是用向量空间模型表示的一个敏感话题的类别;$x_1, x_2, \cdots, x_n$ 是待检测文本的特征,$\boldsymbol{x}=\{x_1, x_2, \cdots, x_i\}$ 是用向量空间模型表示的一个待检测文本。根据条件随机场的基本理论,观察序列 $\boldsymbol{x}$ 对应参数集合 $\Lambda=\{\lambda_1, \cdots, \lambda_j\}$ 的指定状态 $\boldsymbol{y}$ 的条件概率如式(3)所示[9]。

$$P(\boldsymbol{y}|\boldsymbol{x},\lambda)=\frac{1}{Z(\boldsymbol{x})}\exp\left(\sum_{i=1}^{n}\sum_{j}\lambda_j f_j(y_{i-1}, y_i, \boldsymbol{x}, i)\right) \tag{3}$$

其中:f_j 为特征函数,是转移特征函数和状态特征函数的统一表示;λ_j 为特征函数的权值,通过训练得到;$Z(\boldsymbol{x})$ 为归一化因子,如式(4)所示。

$$Z(\boldsymbol{x})=\sum_{y_i\in y}\exp\left(\sum_{i=1}^{n}\sum_{j}\lambda_j f_j(y_{i-1}, y_i, \boldsymbol{x}, i)\right) \tag{4}$$

1.3 敏感话题识别

基于 CRFs 的敏感话题识别过程首先要将待检测文本表示为模型中的观察输入序列和输出类别序列。从 K 篇待检测文本中随机挑出 1 篇作为 CRFs 模型中的观察输入序列 s,剩余的 $K-1$ 篇待检测文本作为 CRFs 模型中的 $K-1$ 个输出类别序列。这样就可以通过 CRFs 模型来计算输入序列中的文档和输出序列中文档之间的概率值,以后的步骤用类似方法进行迭代,直到识别出所有敏感话题的类别。其识别过程如图 3 所示。

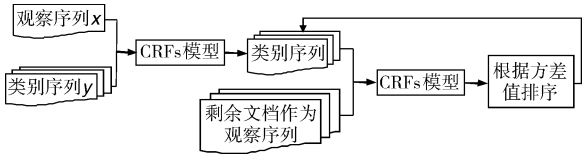

图 3 CRFs 模型敏感话题识别流程

表 1 概率方差表

	文本	类别		
		C_1	C_2	
↓	d_1	$P_{1,1}$	$P_{1,2}$	δ_1
	d_2	$P_{2,1}$	$P_{2,2}$	δ_2
↑	⋮	⋮	⋮	⋮
	d_{n-3}	$P_{n-3,1}$	$P_{n-3,2}$	δ_{n-3}

a)将得到的 $K-1$ 个概率值排序,最大概率值所对应的文本与输入观察序列中那篇文本归并为一类并记做类 C_1,同时将最小概率值对应的文本记做类 C_2。

b)将剩余的 $K-3$ 个待检测文本作为输入观察序列,C_1 和 C_2 作为输出类别序列,这样通过 CRFs 模型可以得到待检测文本隶属于 C_1 和 C_2 类的两个概率值。

c)对每篇待检测文本与输出类别序列的各个概率值求方差并排序(如表 1 所示),方差值越大说明该文本和类别有很大的区分度;反之,方差值越小就说明该文本和类别没有区分度或相关性不大。

d)查看最小方差值所对应的那篇文本的所有概率值,若其中最小的概率值小于某一阈值 θ 就将其作为一个新的类 C_3;否则,查看方差值第二小的那篇文本。依此类推,直到找到概率值小于阈值 θ 的文本。同时将最大方差值所对应的文本归并到最大概率所对应的类别。

e)重复步骤 b)~d),直到所有的文本都被归类。

阈值 θ 用于控制是否需要增加新的类别,若 θ 值越大,类别间的区别越不明显,从而使得到的类别数越多,会将属于一个类别的文本错分出来;若 θ 值越小,得到的类别数将会越少,从而会将文本错分为一个类别。根据表 1 描述的报道之间的概率值以及方差值,在此基础上通过观察类别间的距离随 θ 的变化趋势对 θ 进行估计。

1.4 CRFs 模型训练[11]

训练 CRFs 模型需要调整其参数 $\Lambda=\{\lambda_1, \cdots, \lambda_j\}$,使得它与训练集 $T=(y_k, x_k)$ 相吻合。参数 λ 的选取采用最大似然估计法,即假设 $P(\boldsymbol{y}|\boldsymbol{x},\lambda)$ 为 λ 的函数并使得 $P(\boldsymbol{y}|\boldsymbol{x},\lambda)$ 的对数值最大的 λ 值为估计值,并且为了缓解模型的过度拟合引入了高斯先验分布作为平滑因子,如式(5)所示。

$$L_\Lambda=\sum_T\sum_{i=1}^{n}\sum_j\lambda_j f_j(y_{i-1}, y_i, \boldsymbol{x}, i)-\sum_T\log Z(\boldsymbol{x})-\sum_j\frac{\lambda_j^2}{2\sigma^2} \tag{5}$$

其中:σ 为 λ 的标准差。由于 $L_\wedge$ 为凸函数,导数为零的点为最大值点,对 $L_\wedge$ 求偏导,如式(6)所示。

$$\frac{\partial L_\wedge}{\partial \lambda_j} = \sum_T \sum_{i=1}^{n} f_j(y_{i-1}, y_i, \boldsymbol{x}, i) - \sum_T \sum_{y_i \in y} P(y|\boldsymbol{x}) \sum_{i=1}^{n} f_j(y_{i-1}, y_i, \boldsymbol{x}, i) - \frac{\lambda_j}{\sigma^2} \tag{6}$$

在 CRFs 模型中 f_k 是一个任意的特征函数,对于待检测文本的每一个特征,属于敏感词库并且权值大于某一阈值时为 1,其余情况为 0。式(6)中,第一项为特征 f_k 在经验分布下的期望值,只需计算特征 f_k 在输入观察序列和输出状态序列中出现的次数;第二项是 f_k 在模型分布的期望值,需要使用 Viterbi 算法计算[12]。由于引入了高斯平滑因子,参数 λ 的估计问题可以使用 L-BFGS (limited-memory Broyden-Fletcher-Goldfarb-Shanno)算法解决[13]。可以将 L-BFGS 算法看做是一个黑盒优化过程,只需要为其提供目标函数的一阶偏导函数即可。

2　实例验证

在实验中,本文使用从国内各大网站包括搜狐、新浪、雅虎、腾讯、网易和天涯论坛上发表的 2012 年 1 月 1 日到 2012 年 12 月 1 日期间的 20 000 个新闻文本作为本文实验的数据语料库。实验分为训练阶段和测试阶段,在训练阶段中对 CRFs 模型进行训练,数据采用上述网站中已经定义好的 16 个大的话题类别共 10 000 个文本,包含色情、暴力等敏感信息。测试阶段中对算法进行比较实验,数据集采用包括 105 个话题(由 10 000 篇报道组成),使用天涯网站上的敏感词汇库作为本文的敏感词库。将每篇报道经过内容提取、分词和向量表示后用于实例验证。

本文采用文本聚类中广泛使用的漏检率和错检率以及耗费函数[14]来评价 CRFs 模型的敏感话题识别的性能。假设数据集中与某类话题相关的报道文本数是 A,不相关的文本数是 B,其中,A 篇报道中有 a 篇报道被聚到该类话题中,B 篇报道中有 b 篇被聚到该类话题中,那么漏检率如式(7)所示,错检率如式(8)所示。

$$P_{\text{miss}} = (A - a)/A \tag{7}$$

$$P_{\text{fault}} = b/B \tag{8}$$

代价函数建立在系统漏检率和错检率的基础上,综合考虑了漏检和错检得到的性能损耗代价,其计算如式(9)所示。

$$C = C_{\text{miss}} P_{\text{miss}} P_{\text{target}} + C_{\text{fault}} P_{\text{fault}} P_{\text{non-target}} \tag{9}$$

其中:C_{miss} 和 C_{fault} 分别代表漏检一个报道和错检一个报道的代价系数,不同的评测中取值不一样,根据 TDT 的评测标准本文取 $C_{\text{miss}} = 1$,$C_{\text{fault}} = 0.1$;P_{target} 和 $P_{\text{non-target}}$ 是文本属于某类话题的先验概率($P_{\text{non-target}} = 1 - P_{\text{target}}$)。

为了更加客观地评价本文提出的基于 CRFs 模型的敏感话题识别模型的性能,本文将经典 single-pass 算法[15,16]和 DBSCAN 算法[17]以及 GSP(similarity pagerank)算法[18]作为比较对象进行实验。经过训练阶段,已经得到了 CRFs 模型需要的特征模板以及参数,经过测试阶段,检测出了包括色情、暴力、政治等 12 个敏感话题,在测试阶段得到的实验结果如表 2 和图 4 所示。观察图表可知,因为本文提出的 CRFs 模型考虑了每篇文档特征之间的关系,更加准确地拟合了真实世界,聚类后话题类别数更接近实际。而经典 single-pass 算法对输入待检测文本的顺序比较敏感,当输入文本的顺序发生变化时聚类结果也会发生很大的变化;由于 DBSCAN 算法没有考虑类别距离和数据密度大小的不均匀性,不容易得到高质量聚类结果;GSP 算法对敏感性文档的识别度不高,因此,对敏感话题的发现效果不如 CRFs 模型。

表 2　两种算法性能指标比较表

算法	评价指标			
	话题数	漏检率 P_{miss}	错检率 P_{fault}	耗费函数
CRFs 模型	95	0.001 09	0.098	0.019 43
经典 single-pass 算法	85	0.005 12	0.189	0.029 734
DBSCAN 算法	89	0.004 65	0.137	0.025 734
GSP 算法	90	0.003 78	0.129	0.020 815

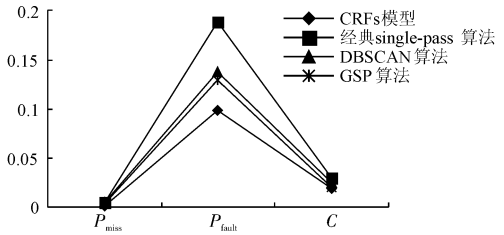

图 4　三种性能指标比较

在本文实验中,还使用文献检索中广泛使用的正确率、召回率及 F 度量[19]对 CRFs 模型和经典 single-pass 算法以及 GSP 算法进行了比较。假设一个文本聚类器的各文本聚类结果表示如表 3 所示,其中,A 表示被正确聚到类 C_m 的文本数,B 表示属于类 C_m 但未聚到该类的文本数,D 表示不属于类 C_m 但被错误聚到该类的文本数,E 表示不属于类 C_m 且未被聚到该类的文本数。那么,类 C_m 准确率 P_{Cm} 如式(10)所示。

$$P_{Cm} = \frac{A}{A + D} \tag{10}$$

类召回率 R_{Cm} 如式(11)所示。

$$R_{Cm} = \frac{A}{A + B} \tag{11}$$

F_{Cm} 度量如式(12)所示。

$$F_{Cm} = \frac{2R \times P}{R + P} \tag{12}$$

表 3　文本聚类输出结果

分类	标记为类 C_m	未标记为 C_m
属于类 C_m	A	B
不属于类 C_m	D	E

准确率反映了聚类后结果中含有实际属于该类文本的比例,召回率反映了聚类后结果中错误聚合到该类文本的比例。在聚类结果中,每个类别都有对应的正确率、召回率和 F 度量,因此需要每个类别的聚类结果来综合评价聚类器的性能。本文采用宏平均(macro average)的方式,即计算类别的正确率、召回率以及 F 度量的平均值[20],计算公式如下:

$$\text{macro average } P = \frac{\sum_{Cm \in C} P_{Cm}}{|C|} \tag{13}$$

$$\text{macro average } R = \frac{\sum_{Cm \in C} R_{Cm}}{|C|} \tag{14}$$

$$\text{macro average } F = \frac{\sum_{Cm \in C} F_{Cm}}{|C|} \tag{15}$$

依然采用上述数据集作为实验的数据基础,105 个话题 2 657篇报道,结果如表 4 和图 5 所示。

表 4　CRFs 模型和经典 single-pass 算法实验结果对比表

算法	评价指标		
	macro average R	macro average P	macro average F
CRFs 模型	0.852 6	0.802 59	0.823 5
经典 single-pass 算法	0.750 1	0.702 43	0.701 98
DBSCAN 算法	0.790 6	0.753 8	0.724 32
GSP 算法	0.801 3	0.754 7	0.783 6

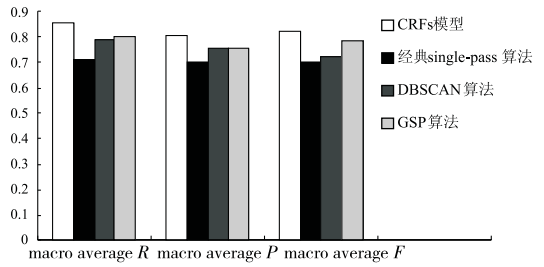

图 5 CRFs 模型和经典 single-pass 算法实验结果对比表

实例验证结果表明,本文提出的基于 CRFs 模型的敏感话题识别研究算法能够更加准确地对网络文本进行敏感话题识别,其漏检率 P_{miss}、错检率 P_{fault} 以及耗费函数 C 都低于经典 single-pass 算法和 DBSCAN 算法以及 GSP 算法,且准确率 P、召回率 R 以及 F 度量的宏平均值也比经典 single-pass 算法和 DBSCAN 算法以及 GSP 算法高。

3 结束语

本文将敏感话题识别问题看做是 CRFs 的计算问题,提出了基于 CRFs 模型的敏感话题识别方法。首先,将随机挑选出的一篇待检测文本 s 和剩余的待检测文本分别作为 CRFs 模型的观察序列和状态序列来计算文本 s 和其余待检测文本间的相关性概率值;然后,将相关性最高的那篇文本和文本 s 合并表征一个类别;同时,将相关性最低的那篇文本作为另一个类别,这样将这两个类别作为 CRFs 模型新的状态序列,剩余的待检测文本作为新的观察序列进行迭代,据此实现敏感话题的识别。在数据集上进行的实验中,该方法的代价函数以及宏平均 F 度量都取得了很好的效果。

参考文献:

[1] 冯颖. 网络舆情敏感话题发现平台的研究[D]. 北京:北京交通大学电子信息工程学院,2009.

[2] ZHANG Kuo, ZI Juan, WU Li-lang. New event detection based on indexing-tree and named entity[C]//Proc of the 30th Annual International ACM SIGIR Conference on Research and Development in Information on Retrieval. New York: ACM Press, 2007: 215-222.

[3] 薛晓飞,张永奎,任晓东. 基于新闻要素的新事件监测方法研究[J]. 计算机应用, 2008, 28(11): 2975-2977.

[4] 王颖颖,张骞,胡乃静. 在线新事件检测系统中国的性能提升策略[J]. 计算机工程, 2008, 34(15): 72-74.

[5] ALSUMAIT L S. Online topic detection, tracking and significance ranking using generative topic models[M]. Fairfax: Georage Mason University, 2009.

[6] 闵可锐,赵迎宾,刘昕,等. 互联网话题识别与跟踪系统设计与实现[J]. 计算机工程,2008,34(19):344-352.

[7] 鲁明羽,姚晓娜,魏善岭. 基于模糊聚类的网络论坛热点话题挖掘[J]. 大连海事大学学报,2008,34(4):146-153.

[8] 宁伟,蔡东风,张桂平,等. 基于条件随机场的冠词选择研究[J]. 中文信息学报, 2008, 22(6): 116-122.

[9] SALTON G, BUCKLEY B. Term-weighting approach in automatic text retrieval [J]. Information Processing and Management, 1988,24(5): 513-523.

[10] 薛峰,周亚东,高峰,等. 一种突发性热点话题在线发现与跟踪方法[J]. 西安交通大学学报,2011,45(12):64-70.

[11] LAFFERTY J, McCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]//Proc of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2001: 282-289.

[12] 周俊生,戴新宇,尹存燕,等. 基于层叠条件随机场模型的中文机构名自动识别[J]. 电子学报, 2006,34(5):804-809.

[13] NOCEDAL J. Updating quasi Newton matrices with limited storage [J]. Mathematics of Computation, 1980,35(151):773-782.

[14] ALLAN J, FENG Ao, BOLIVAR A. Flexible intrintic of evaluation hierarchical clustering for TDT[C]//Proc of the 12th International Conference on Information and Knowledge Management. 2003: 263-270.

[15] ALLAN J, CARBONELL J, DODDINGTON J, *et al.* Topic detection and tracking pilot study final report[C]//Proc of DARPA Broadcast News Transcription and Understanding Workshop. San Francisco: Morgan Kaufmann Publishers, 1998: 194-218.

[16] 殷风景,肖卫东,葛斌,等. 一种面向网络话题发现的增量文本聚类算法[J]. 计算机应用与研究, 2011, 28(1):54-57.

[17] ESTER M, KRIEGEL H P, SANDER J, *et al.* A density-based algorithm for discovering clusters in large spatial databases[C]//Proc of International Conference on Knowledge Discovery and Data Mining. 1996: 226-231.

[18] 薛玮. 网络舆情信息挖掘系统的研究[D]. 北京:北京交通大学, 2008.

[19] VITERBI A J. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm[J]. IEEE Trans on Information Theory, 1967, 13(2): 260-269.

[20] 罗铁坚,王文杰,刘务华. 文本聚类算法的质量评价[J]. 中国科学院研究生院学报,2006, 23(5): 640-646.

(上接第 980 页)

[10] 朱勤, 彭希哲, 吴开亚. 基于投入产出模型的居民消费品载能碳排放测算与分析[J]. 自然资源学报,2012,27(12):2018-2029.

[11] JOHANSSON J. Risk and vulnerability analysis of interdependent technical infrastructures[D]. Lund: Lund University, 2010.

[12] 顾洁, 秦玥, 包海龙,等. 基于熵权与系统动力学的配电网规划动态综合评价[J]. 电力系统保护与控制, 2013(1):1-7.

[13] LEONTIEF W. Input-output economics[M]. Oxford: Oxford University Press, 1986.

[14] EUSGELD I, KROGER W, SANSAVINI G, *et al.* The role of network theory and object-oriented modeling within a framework for the vulnerability analysis of critical infrastructures[J]. Reliability Engineering & System Safety, 2009, 94(5): 954-963

[15] JOHANSSON J, JÖNSSON H, JOHANSSON H. Analysing the vulnerability of electric distribution systems: a step towards incorporating the societal consequences of disruptions[J]. International Journal of Emergency Management, 2007, 4(1): 4-17.

[16] LITTLE R G. A socio-technical systems approach to understanding and enhancing the reliability of interdependent infrastructure systems [J]. International Journal of Emergency Management, 2004, 2(1): 98-110.

[17] EARL E L, JOHN E M, WILLIAM A W. Restoration of services in interdependent infrastructure systems: a network flows approach[J]. IEEE Trans on Systems, Man, and Cybernetics, Part C: Application and Reviews, 2007, 37(6): 1303-1317.

[18] WINKLER J, DUEÑAS-OSORIO L, STEIN R, *et al.* Interface network models for complex urban infrastructure systems[J]. Journal of Infrastructure Systems, 2011, 17(4): 138-150.

[19] OUYANG Min, HONG Liu, MAO Zi-jun, *et al.* A methodological approach to analyze vulnerability of interdependent infrastructures[J]. Simulation Modelling Practice and Theory, 2009, 17(5): 817-828.

[20] MOTTER A, NISHIKAWA T, LAI Ying-cheng. Cascade-based attacks on complex networks[J]. Physical Review E, 2002, 66(6): 065102.