第 32 卷 第 6 期
2017 年 12 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 32 No. 6
Dec. 2017

DOI:10.13364/j.issn.1672-6510.20160432

基于 Single-Pass 的在线话题检测改进算法

马永军[1,2]，刘 洋[1]，李亚军[1]，汪 睿[1]
（1. 天津科技大学计算机科学与信息工程学院，天津 300457；
2. 天津科技大学食品安全管理与战略研究中心，天津 300222）

摘 要：现有话题检测的主要方法是利用 Single-Pass 及其改进算法进行聚类分析，没有考虑文本的结构特点，相似度计算方法单一，从而影响准确度. 针对此问题，改进了 Single-Pass 的相似度计算方法，综合考虑文本的标题、摘要、时间、地名以及来源等要素，采用层次分析法计算并赋以不同权重，提出一种多相似度计算组合策略. 考虑到食品安全是一个广受关注的话题，实验通过网络爬虫抓取并筛选了最近 3 年食品安全方面的媒体信息，以此作为数据进行分析，结果表明，采用本文提出的改进 Single-Pass 聚类算法，话题检测准确度更高.

关键词：网络舆情；Single-Pass；相似度计算；食品安全
中图分类号：TP391 **文献标志码**：A **文章编号**：1672-6510(2017)06-0073-06

An Improved Algorithm Based on Single-Pass for Online Topic Detection

MA Yongjun[1,2], LIU Yang[1], LI Yajun[1], WANG Rui[1]
(1. College of Computer Science and Information Engineering, Tianjin University of Science & Technology, Tianjin 300457, China; 2. Food Safety Management and Strategic Research Center, Tianjin University of Science & Technology, Tianjin 300222, China)

Abstract: At present, the main research method of existing topic detection is to use Single-Pass and its improved algorithm for clustering analysis. However, these algorithms use a single similarity calculation method without considering the structural characteristics of the text, which affects the clustering accuracy. This research has improved the similarity calculation method of Single-Pass and proposed a multi-similarity computation combination strategy which toke the title, abstract, time, place names and source into consideration, and used the analytic hierarchy process to calculate and assign them different weights. As food safety is a widely concerned topic, we analyzed the data about food safety in the last three years which we could get with the web crawler. The results show that the improved Single-Pass clustering algorithm proposed in this paper has a higher topic detection accuracy.

Key words: internet public opinion; Single-Pass; similarity calculation; food safety

与传统的信息传播渠道相比，网络具有更大的开放性和虚拟性，随着自媒体时代的到来，各种不同的观点、议论和情绪通过网络空间不断发酵放大，有的形成了网络舆情事件. 在当今我国建设网络强国的背景下，网络舆情分析得到高度关注[1]. 国外对网络舆情的研究始于 19 世纪中期，我国对网络舆情的研究要稍晚一些，相关研究学者提出网络舆情是指公众通过互联网表达和传播不同的情绪、态度和意见[2–3]. 食品安全作为一个人们普遍关心的民生话题被广为关注，对食品安全舆情分析显得尤为重要[4–5].

话题检测是网络舆情分析的关键一环，目前研究方法主要集中在对文本聚类算法的选取上，例如：采用 K-means 聚类算法计算待聚类的文本与已存在的类簇之间的距离，通过比较距离与阈值的大小完成话题检测[6]；采用凝聚层次聚类算法完成话题检测任务[7]；采用基于密度的聚类算法 OPTICS 结合 LDA 模型以达到话题检测的目的[8]；采用基于网格和矩阵的高维数据流聚类算法[9]等；其中常用的聚类算法是

收稿日期：2016–12–30；**修回日期**：2017–04–28
基金项目：天津市教委重大项目(2014ZD22)；天津市应用基础与前沿技术研究计划(14JCQNJC00300)
作者简介：马永军（1970—），男，山东日照人，教授，yjma@tust.edu.cn

Single-Pass 算法[10]. Yang 等[11]首次提出采用 Single-Pass 聚类找回时间间隔较大的话题相关文本. 针对 Single-Pass 算法的不足，陆续出现了一些改进算法，例如：以 Single-Pass 算法为核心，采用语义相似度度量文本相似性[12]或识别命名实体[13]，从而实现话题检测；有些算法在提高聚类准确性方面进行改进[14−15]；还有算法采用概率潜在语义方法计算文本相似度[16]，计算文本被正确划分到某一类别的概率[17]，都获得了更好的话题识别度.

上述改进 Single-Pass 聚类算法均采用单一的相似度计算方法，没有考虑文本的结构特点，因而影响了聚类的准确度. 针对此问题，本文在现有 Single-Pass 算法的基础上，通过分析文本结构特点，综合考虑文本标题、摘要、时间、地名以及来源等要素，提出了一种多相似度计算组合策略.

1　话题检测原理介绍

话题检测通常需要文本预处理和文本聚类两个步骤. 文本预处理将非结构化的文本信息转化为可以计算的向量，文本聚类对各个数据点进行聚类，输出话题集，完成话题检测任务.

1.1　文本预处理

文本预处理包括中文分词、特征选择、特征加权以及构建文本表示模型. 首先，借助分词系统将文本信息分成若干个词语，它是文本预处理中的重要环节. 其次，对分词后的词语进行特征选择，去除无用词和无效词. 常用的特征选择方法为文本频数(document frequency，DF)[18]，其从特征的候选集中选择一些对文本表征能力更强的词语作为特征项，当一个特征的文本频数低于某个阈值时，即它在很少的文本中出现过，那么就认为它含有较少的类别信息，应当被舍弃掉. 然后，对选取出来的特征项分配其权重，以突出对文本表现能力强的特征. TF-IDF 是目前使用最广泛的一种特征加权方法[19]，它的原理是：某个词语在一篇文本中出现的频率越高 TF 值就越大，而该词语在其他文本中出现的次数越少 IDF 值就越大，那么就认为该词语可用于区分类别，应当分配其较大的权重值. 最后，构建文本表示模型，目的是以计算机能够识别的方式来表示一个文本. 目前向量空间模型(vector space model，VSM)是被广泛应用的一种文本表示模型[20]. VSM 将文本映射到多维空间中的一个点上，每一个向量表示一个文本，向量的每一个维度都是一个特征项，特征项的频率为向量的值. 通过建立 VSM 将文本信息表示成 n 维空间中的一个向量.

1.2　Single-Pass 聚类算法

将文本向量化之后，文本转化为了可计算的向量空间模型，通过计算向量之间的相似度得出两个文本之间的相关程度. 基于 VSM 的文本相似度计算有多种方法，Single-Pass 聚类算法采用夹角余弦公式计算两个文本相似度

$$S\left(d_i,d_j\right)=\cos\theta=\sum_{k=1}^{D}\omega_{ik}\omega_{jk}\Big/\sqrt{\sum_{k=1}^{D}\omega_{ik}{}^2\sum_{k=1}^{D}\omega_{jk}{}^2} \tag{1}$$

式中：ω_{ik} 为文本 i 的特征向量的第 k 个权重；ω_{jk} 为文本 j 的特征向量的第 k 个权重；D 为文本特征项的维度数量. 两个文本向量夹角越小，余弦值就越大，表明两篇文本的内容越相似.

Single-Pass 算法每次对一个文本进行聚类分析，将待聚类的文本与已有话题的各个文本依次进行比较，分别计算文本之间的相似度，选取最大的相似度及文本所在的话题类，如果该相似度大于预先设定的阈值，则将其划分到该话题类中，否则创建一个新话题并将该文本加入到此话题中. 通过多次迭代计算，可以将文本集聚类为多个话题类，实现话题检测.

2　改进的 Single-Pass 聚类算法

2.1　相似度计算方法

经典 Single-Pass 聚类算法采用单一的相似度计算方法，没有考虑文本的结构特点，无法很好地判别两篇文本的相似程度. 具体分析文本的结构特点可知：标题是一篇文本的核心，通过标题可以了解文本的中心内容；摘要是一篇文本的主体，它可以辅助了解文本的相关信息，占有重要地位；相似的事件可能会在不同时间段发生，在时间距离相近的两篇文本中包含的特征项非常相似或相近，如果不考虑两篇文本的时间距离，那么 Single-Pass 聚类算法不会将其归为同一类别中，所以时间也是一个必不可少的要素；考虑到地名在文本中的重要性和特殊性，在相同事件中经常会出现不同的地名术语，利用传统的夹角余弦方法计算这些术语之间的相似度为 0，但是其在地理空间上可能是有关联的，因此需要将地名术语分离出来单独计算；此外，自媒体时代下的网络信息具有大量转载、重复性较高的特点，同样的一篇文本内容可

能会通过网络不断转载，造成文本信息重复，因此在计算相似度时非常有必要考虑文本的来源.

经过上述分析，本文改进相似度的计算方法，选取文本的标题、摘要、时间、地点以及来源作为相似度比较的重要依据.

(1) 标题、摘要相似度：采用夹角余弦公式计算标题、摘要相似度，计算方法见式(1).

(2) 时间距离：引入时间距离概念，定义文本 d_i 和 d_j 的时间距离为

$$S_{\text{time}}\left(d_i,d_j\right)=\begin{cases}1-t/m & t<m\\ 0 & t\geqslant m\end{cases} \tag{2}$$

式中：$t=\left|t_i-t_j\right|$，为两篇文本的时间之差；m 则为时间间隔，本文设定为 7 d.

(3) 地名相似度：构建一棵以中国为根节点的地理树，利用地名之间的下属关系将每个地名表示成树中的一个节点.

计算两个地名的相似度需要考虑：地理树中的每一个子节点是父节点的一个分支，两个子节点之间的距离，两个子节点的公共深度，每个节点距离根节点的深度. 综合考虑以上因素对计算两个地名相似度的影响，定义地名 p_i,p_j 的相似度为

$$S_{\text{place}}\left(p_i,p_j\right)=\frac{2deep\left(p_i\bigcap p_j\right)}{deep\left(p_i\right)+deep\left(p_j\right)} \tag{3}$$

式中：$deep\left(p_i\bigcap p_j\right)$ 为地名 p_i 与地名 p_j 距离根节点的公共深度；$deep\left(p_i\right)$ 为地名 p_i 距离根节点的深度；$deep\left(p_j\right)$ 为地名 p_j 距离根节点的深度.

(4) 来源相似度：PageRank 是一种基于链接分析用来计算页面权威性的算法，传统的 PageRank 算法在计算 *PR* 值时，将页面的 *PR* 值平均分配给该页面的所有链接，却忽略了判断这些页面权威性，本文采用改进的 PageRank 算法来计算页面的 *PR* 值，用于计算两个文本来源的相似度，改进的 PageRank 计算公式为

$$PR_{(\text{p})}=\left(1-d\right)+d\left[a\sum_{i\in v_1}\frac{PR_i}{C_i}+(1-a)\sum_{j\in v_2}\frac{PR_j}{C_j}\right] \tag{4}$$

式中：d 为阻尼系数，通常取值为 0.85；a 为判断链出站点是否为站外链接的比重系数，站外页面相比于同一个站内的页面更能体现页面所属站点的重要性，取值为 0.75. v_1 为链出页面与 p 页面不是同一个站点的集合；C_i 表示页面 i 全部链出页面的数量；v_2 为链出页面与页面 p 属于同一个站点的集合；C_j 表示页面 j 全部链出页面的数量.

(5) 总相似度：将以上各相似度加权求和，得到两个文本之间的总相似度为

$$S_{\text{total}}(d_i,d_j)=\alpha S_{\text{title}}+\beta S_{\text{body}}+\gamma S_{\text{time}}+\lambda S_{\text{place}}+\omega\frac{PR_{d_i}}{PR_{d_j}} \tag{5}$$

式中：S_{title}、S_{body}、S_{time}、S_{place} 分别为标题、摘要、时间以及地名的相似度；最后一项为文本 d_i 和 d_j 来源的 *PR* 值之比；α、β、γ、λ、ω 分别为标题、摘要、时间、地名和来源的权重因子，5 个权重因子取值均在 0 到 1 之间，且相加总和等于 1.

采用层次分析法确定 5 个权重因子的取值[21]，具体计算方法为：根据标度方法，确定 5 个影响因子(标题、摘要、时间、地名以及来源)两两因素相比的标度值，构造出成对比较矩阵

$$\boldsymbol{A}=\begin{bmatrix}1 & 2 & 3 & 2 & 3\\ 1/2 & 1 & 1/2 & 1/5 & 5\\ 1/3 & 2 & 1 & 1 & 5\\ 1/2 & 5 & 1 & 1 & 7\\ 1/3 & 1/5 & 1/5 & 1/7 & 1\end{bmatrix} \tag{6}$$

为了验证矩阵 $\boldsymbol{A}$ 归一化后的权向量是否为所求的特征向量，需要计算矩阵 $\boldsymbol{A}$ 是否可以通过一致性检验. 利用一致性指标和随机一致性指标的数值表，计算得出一致性比率＜0.1，通过一致性检验，从而得到特征向量 $\boldsymbol{\omega}=(0.35,0.3,0.15,0.1,0.1)^{\text{T}}$，即确定 5 个权重因子的值为 $\alpha=0.35$，$\beta=0.3$，$\gamma=0.15$，$\lambda=0.1$，$\omega=0.1$.

2.2　话题检测算法流程

首先输入一个文本及阈值，采用改进的相似度方法计算待聚类文本与已有话题各个文本的相似度并记录最大相似度以及文本所在的话题类，如果最大相似度大于阈值则将其加入到该话题中，否则新建一个话题并将该文本加入到新话题中，经过多次迭代计算，可以得到多个话题类，完成话题检测.

基于改进后的 Single-Pass 聚类算法完成话题检测的具体流程如下：

(1) 输入一个文本 d 及阈值 T_{c}；

(2) 判断 d 是否为第一个文本，如果是则转至(3)，否则转至(4)；

(3) 创建一个新话题并将文本 d 加入到新话题，转至(7)；

(4) 对文本 d 进行预处理、构建向量空间模型；

(5) 计算文本 d 与已有话题各个文本的相似度，包括标题、摘要、时间、地名以及来源，赋予每个权重因子不同的权值并加权求和得到两文本之间的总相似度，记录最大相似度 $S_{\max}$ 和文本所在的话题类 T；

(6) 如果最大相似度 $S_{\max}$ 大于预先设定的阈值 T_c，则将文本 d 聚类到话题类 T 中，否则转至(3)；

(7) 完成一次聚类，直到所有的数据点均已被划分到各个类别中.

3　实　验

3.1　实验平台搭建和数据获取

为了适应大数据的需求，便于后期的应用扩展，实验采用 1 台服务器和 3 台 PC 搭建 Hadoop 分布式集群，每台机器安装 64 位 CentOS 系统，硬盘为 500 GB，内存为 4 GB. Hadoop 分布式总体架构见图 1.

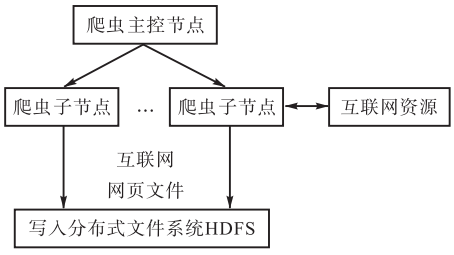

图 1　Hadoop 分布式总体架构

Fig. 1　Hadoop distributed overall architecture

实验采用基于 Hadoop 的分布式主题爬虫抓取并筛选了最近 3 年食品安全方面的媒体信息，共计 98 000 条，其数据集具有一定代表性. 通过手工整理标注出 1 080 条有效事件报道，共分为 5 个有关食品安全事件的话题，见表 1.

表 1　食品安全话题

Tab. 1　Topics of food safety

话题编号	食品安全话题	文本数量/个
T1	重金属小龙虾	192
T2	发霉大米	218
T3	工业明胶	209
T4	福喜食品	225
T5	假冒奶粉	236

下面是“重金属小龙虾”话题的实验样例：

标题为“小龙虾虾头的重金属含量比虾肉高得多”、“小龙虾头部重金属最多营养专家建议最好自己做”、“小龙虾用于污水处理、重金属超标”.

摘要为“很多人爱吃的小龙虾，虾头到底能不能吃?某高一同学现场做实验测出：虾头里有重金属铜离子残留，比虾肉中多很多……”，“小龙虾头部重金属最多，营养专家建议最好自己做……”，“近年来，有网文屡屡宣称，小龙虾喜欢脏污环境、被用来清污、体内重金属超标，还能不能食用. 因为虾是高蛋白食物，部分过敏体质者会对小龙虾产生过敏症状……”.

时间上选取近 3 年的数据，格式为 XXXX 年 XX 月 XX 日，例如，2016 年 8 月 15 日.

地名中包含 4 个直辖市以及各大省市，取到区级别，例如天津市、保定市、济南市、浦东新区、宣化区、大庆市.

数据来源自 www.cfsn.cn(中国食品安全网)、www.foodmate.net(食品伙伴网)、www.cfsiw.com(中国食品安全信息网)、www.cfqn.com.cn(中国食品网)和 www.cnfdn.com(中国食品监督网).

3.2　实验设计

实验首先采用由中科院研发的 ICTCLAS 分词系统对采集到的数据进行分词，然后进行文本预处理，最后采用本文改进后的 Single-Pass 聚类算法进行聚类分析，得到多个话题类.

在评价话题检测性能时，通常采用话题检测评测标准[22]，本文采用漏检率、误检率以及耗费函数值作为评价指标.

计算上述 3 个指标之前，先定义 4 个量值：A 为被系统检测到并聚合到话题中并且人为判定为与话题相关的文本数；B 为被系统检测到且被聚合到话题中但人为判定与话题不相关的文本数；C 为未被系统检测到却与话题相关的文本数；D 为未被系统检测到并且人为判定与话题不相关的文本数.

确定好 4 个量值之后，定义漏检率 $R_E = C/(A+C)$，误检率 $R_F = B/(B+D)$，耗费函数值是对漏检率和误检率的综合评价指标，定义为

$$H_{\text{cost}} = C_E \times R_E \times P + C_F \times R_F \times (1-P) \tag{7}$$

式中：C_E 和 C_F 分别为漏检率和误检率的系数；P 为文本归属某一类话题的先验概率. 根据评测标准，取 $C_E = 1$，$C_F = 0.2$，$P = 0.02$.

为了验证本文改进算法的有效性，实验选取经典 Single-Pass 算法以及文献[16]、文献[17]提出的基于 Single-Pass 的改进方法，对实验结果进行分析对比.

3.3　结果分析

图 2 为 4 种方法的漏检率–误检率变化曲线. 由

图 2 可知：漏检率和误检率会随着其中一个量的变化而变化，两者呈负相关关系，即随着误检率的升高漏检率就会降低；反之，当误检率降低时漏检率就会升高.

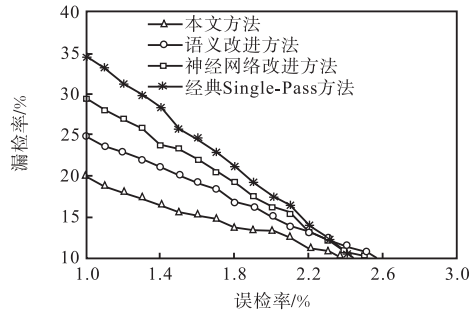

图 2　漏检率–误检率变化曲线

Fig. 2　Curves of miss rate-false rate changes

当误检率一定时，漏检率越小则表明聚类效果越优. 取相同的误检率为 1%，比较 4 种方法的漏检率大小. 经典 Single-Pass 聚类算法采用单一的相似度计算方法，没有考虑特征项之间的语义关系以及文本的结构特点，其漏检率接近 35%；采用语义度量相似度的 Single-Pass 方法在原有算法的基础上，采用概率潜在语义分析方法，综合考虑词汇与文本的共现形式，通过用训练集对模型进行训练，获得文本与词汇的概率分布，在计算文本之间相关度方面，提高了对话题相关报道的认知性，相比于经典 Single-Pass 聚类算法，其漏检率有所下降；借鉴神经网络思想改进后的 Single-Pass 聚类算法，采用每个单元存储权重值，一组权重集就可以表示自身对每个输出类别的重要性，在聚类分析时，减少了与话题相关但是未被系统检测到的文本数量，漏检率明显下降；本文方法在漏检率方面均低于上述两种改进算法.

当漏检率一定时，误检率越小则表明聚类效果越好. 取相同的漏检率为 15%，比较 4 种方法的误检率大小. 采用经典的 Single-Pass 聚类算法，选取某一话题的所有文本来表示话题类，随机选取聚类中心，通常选取第一个文本成为第一个类别的聚类质心，如果第一个文本不太具有代表意义，那么便不能全面地对一个类中的话题进行很好的阐述，随着文本聚类的进行，被错分的可能性逐渐增加，其误检率接近 2.2%；采用语义度量相似度的 Single-Pass 方法加强了语义相近的特征项的权重值，区别传统的文本相似度计算，以文本的语义相关度计算作为话题聚类标准，提高话题检测的准确性，误检率有所下降；借鉴神经网络思想改进后的 Single-Pass 聚类算法，采用网格结构存储权重值，通过检索求和每个相关单元的权重列表值计算下一个输出单元被正确聚类到某一类别的概率，被错分的可能性会随着聚类的进行而逐渐降低；本文改进算法误检率为 1.8% 左右，明显低于上述两种改进算法.

在实验数据一致的前提下，当 4 种方法获得最优聚类效果时，比较各项评测指标，见图 3. 由图 3 可知：经典 Single-Pass 聚类算法，采取单一的相似度计算方法，采用所有的话题文本表示话题类，随着文本聚类的进行，特征项分布不均匀，被错分的可能性较大，其漏检率和误检率较高；采用语义度量相似度的 Single-Pass 方法区别于传统的文本相似度计算，将文本的语义相关度计算作为话题聚类标准，虽然各项指标值有所下降，但是没有考虑半结构化网页中不同位置的特征项重要程度的不同，会影响相似度计算，从而导致误检率偏高；借鉴神经网络思想改进后的 Single-Pass 聚类算法在 Single-Pass 聚类算法的基础上，通过计算待聚类文本被正确分到某一类别的概率完成聚类分析，其漏检率要明显低于经典 Single-Pass 聚类算法；本文方法分析了文本结构的特点，提出多相似度计算的组合策略，其耗费函数值为 4 种方法中最小的，聚类效果最优.

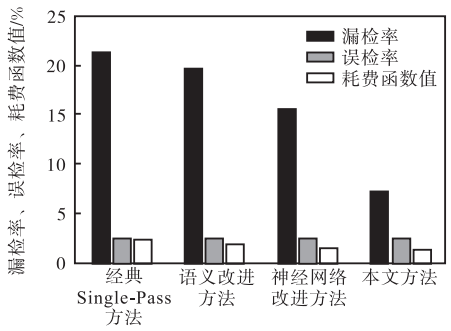

图 3　各项评测指标比较

Fig. 3　Comparison of the evaluation indicators

4　结　语

本文在经典 Single-Pass 聚类算法的基础上，详细分析文本结构的特点，改进相似度计算方法，提出了一种多相似度计算组合策略. 采用夹角余弦公式计算标题、摘要的相似度；引入时间距离概念；构建一棵以中国为根节点的地理树，利用地名之间的下属关系计算地名相似度；采用改进的 PageRank 算法估算文本来源的网站 *PR* 值，用于计算文本来源的相似度，通过赋予不同权重因子的权重值并加权求和以得

到两文本的总相似度. 实验表明,本文改进的 Single-Pass 聚类算法,在漏检率、误检率以及耗费函数值上均有所下降. 在特征加权时,特征项之间的语义关系可能有助于提高算法性能,值得进一步研究.

参考文献:

[1] 《总体国家安全观干部读本》编委会. 总体国家安全观干部读本[M]. 北京:人民出版社,2016:147–152.

[2] 刘毅. 网络舆情研究概论[M]. 天津:天津人民出版社,2007:51–54.

[3] 王来华. 论网络舆情与舆论的转换及其影响[J]. 天津社会科学,2008(4):66–69.

[4] 林萍,黄卫东,洪小娟. 全媒体时代我国食品安全网络舆情构成要素研究[J]. 现代情报,2013,33(11):12–16.

[5] 任立肖,张亮. 食品安全突发事件网络舆情的分析模型:基于利益相关者的视角[J]. 图书馆学研究,2014(1):65–70.

[6] Fang C,Jin W,Ma J. K-means algorithms for clustering analysis with frequency sensitive discrepancy metrics [J]. Pattern Recognition Letters,2013,34(5):580–586.

[7] 潘大庆. 基于层次聚类的微博敏感话题检测算法研究[J]. 广西民族大学学报:自然科学版,2012,18(4):56–59.

[8] 李琮,袁方,刘宇,等. 基于 LDA 模型和 T-OPTICS 算法的中文新闻话题检测[J]. 河北大学学报:自然科学版,2016,36(1):106–112.

[9] 胡国辉. 基于不规则网格的高维数据流聚类算法研究[D]. 秦皇岛:燕山大学,2014:34–42.

[10] Yi X L,Zhao X,Ke N,et al. An improved Single-Pass clustering algorithm Internet-oriented network topic detection[C]//Proceedings of the Fourth International Con ference on Intelligent Control and Information Processing. Piscataway:IEEE,2013:560–564.

[11] Yang Y,Pierce T,Carbonell J. A study on retrospective and online event detection[C]//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York:Association for Computing Machinery,1998:28–36.

[12] 陈红阳. 中文微博话题发现技术研究[D]. 重庆:重庆理工大学,2015:44–50.

[13] 唐琳. 面向食品安全的在线新闻话题监测技术的研究与应用[D]. 广州:中山大学,2015:17–21.

[14] Gong Z,Jia Z,luo S,et al. An adaptive topic tracking approach based on Single-Pass clustering with sliding time window[C]//Proceedings of International Conference on Computer Science and Network Technology. Piscataway:IEEE,2011:1311–1314.

[15] 叶施仁,杨英,杨长春,等. 孤立点预处理和 Single-Pass 聚类结合的微博话题检测算法[J]. 计算机应用研究,2016,33(8):2294–2297.

[16] 师伟. 基于语义相关的在线话题发现算法的研究与应用[D]. 西安:西安石油大学,2014:24–29.

[17] Greer K. A Single-Pass classifier for categorical data [J]. Computer Science,2016,3(1/2):1–15.

[18] Gunawan H,Zhang B,Wang Y,et al. Mutual information based method for selecting informative feature sets [J]. Pattern Recognition,2013,46(12):3315–3327.

[19] 金镇晟. 基于改进的 TF-IDF 算法的中文微博话题检测与研究[D]. 北京:北京理工大学,2015:20–21.

[20] Salton G,Wong A,Yang C S. A vector space model for automatic indexing[J]. Communication of the ACM,1975,18(11):613–620.

[21] 徐志明,李栋,刘挺. 微博用户的相似性度量及其应用[J]. 计算机学报,2014,37(1):207–218.

[22] Wang Z M,Zhou X S. A topic detection method based on bicharacteristic vectors[C]//Proceedings of the International Conference on Networks Security,Wireless Communications and Trusted Computing. Piscataway:IEEE Computer Society,2009:683–687.

责任编辑: 常涛