



基于峰值密度聚类的电信业投诉热点话题检测方法

江俊^{1,2,3}, 黄骅⁴, 任条娟¹, 张登辉¹

(1. 浙江树人大学信息科技学院, 浙江 杭州 310015; 2. 浙江大学信息与电子工程学院, 浙江 杭州 310058;
3. 东方通信股份有限公司, 浙江 杭州 310053; 4. 万向集团公司万向研究院, 浙江 杭州 311215)

摘 要: 针对电信业对投诉热点话题缺乏有效的检测方法问题, 提出一种基于峰值密度聚类算法的投诉热点话题检测方法。首先建立电信业专用词库用于投诉样本的文本分词, 采用向量空间模型表示文本分词, 然后通过计算文本分词相似度和密度, 并运用密度峰值聚类算法对分词进行聚类分析。最终通过类簇关键词选取并排序, 从而得到热点话题描述。将本方法应用到某电信企业投诉热点话题检测中, 结果表明本方法有效并具有实际应用价值。

关键词: 热点话题检测; 文本分词; 聚类分析

中图分类号: TP391

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2019076

Telecom complaint hot topic detection method based on density peaks clustering

JIANG Jun^{1,2,3}, HUANG Hua⁴, REN Tiaojuan¹, ZHANG Denghui¹

1. College of Information Science and Technology, Zhejiang Shuren University, Hangzhou 310015, China

2. College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310058, China

3. Eastern Communications Co., Ltd., Hangzhou 310053, China

4. Wan Xiang Research Institute of Wan Xiang Group Corporation, Hangzhou 311215, China

Abstract: In view of the lack of effective detection methods for hot topics in telecom industry, a method of complaint hotspots detection based on density peaks clustering algorithm was proposed. Firstly, a special vocabulary for telecommunication industry was established to segment the complaint samples. The vector space model was used to represent the text segmentation. Then, the similarity and density of the text segmentation were calculated, and the clustering analysis of the words was carried out by using the density peaks clustering algorithm. Finally, keywords were selected and sorted by clustering. This method was applied to the complaint hotspots detection tasks within a telecom company. The results show that this method is effective and has practical application value.

Key words: hot topic detection, text segmentation, cluster analysis

1 引言

近年来, 电信行业业务信息量已呈爆炸式增

长趋势。对于热点话题检测的需求也与日俱增, 以业务投诉为例, 如何及时从海量的投诉中发现话题、跟踪话题, 并根据话题的变化趋势了解相

收稿日期: 2018-12-15; 修回日期: 2019-04-12

基金项目: 浙江省自然科学基金 (No. LGF18F030004, No. LGF19F010005)

Foundation Items: The Natural Science Foundation of Zhejiang Province of China (No. LGF18F030004, No. LGF19F010005)



关业务情况以及用户关注点,已成为各大电信运营商重点关注的问题之一。因此,针对电信业投诉业务信息进行分析、甄别,进而从中检测出有意义的热点话题,已成为当前亟待解决的难点问题。

话题检测与跟踪(topic detection and tracking, TDT)技术^[1]正是为解决通过文本聚类以及关键词提取,并从大规模的文档中提炼出有意义的话题为个人或企业提供决策依据等一系列问题而产生的^[2-3]。近年来,国内外学者针对文本话题检测进行了深入研究,取得了一系列成果,如 Mohd 等^[4]针对新闻话题检测中的话题结构进行了讨论,并采用 Single-Pass 聚类的方式对话题进行聚类。Zhang 等^[5]针对新闻文档提出一种基于增量型聚类的自动话题检测算法,采用改进的权重算法计算特征权值,通过自适应地提炼具有较强辨别能力的文本特征来提高聚类准确性。Wang 等^[6]针对网络话题提出了基于多中心模型的网络热点话题算法以处理大规模动态数据。Cheng 等^[7]针对新闻报道具有较多冗余且话题易漂离的特点,提出一种在线检测算法。Li 等^[8]针对 Tweet 数据流,筛选其中的热词,继而通过词聚类的方式得到热点话题。但与新闻、网络话题相比,电信业投诉文本结构更加复杂短小,且包含大量的专业词汇,因此加大了话题检测的难度。

本文从电信行业对热点话题检测的实际需求出发,提出一种基于峰值密度聚类的电信业投诉热点话题检测方法。首先,建立电信专业词库对投诉样本进行文本分词预处理,并剔除无关词汇。其次,采用向量空间模型表示文本分词,计算每个文本分词的高斯截断密度并使用峰值密度聚类算法对文本集合进行聚类。最后,提取每个类簇关键词,并结合 N-gram 模型对关键词进行排序以更准确的语义描述热点话题,同时对热度进行评估。实验将本方法应用于某电信企业投诉热点话题检测中,结果证明本方法有效且具有实用价值。

2 方法与模型分析

针对电信行业热点话题检测业务的难点问题,本文分别选择了峰值密度聚类方法和 N-gram 语言模型,解决具有行业特色的文本聚类以及语义化关键词排序问题。

2.1 峰值密度聚类方法

文本聚类是热点话题检测中的重要环节,目前话题检测技术中广泛运用的多为经典文本聚类算法,如 DBSCAN^[9-12]、Single-Pass 等。这类算法普遍存在参数敏感、计算量大、难以处理不均匀分布数据集等问题。峰值密度聚类(clustering by fast search and find of density peaks, CFSFDP)方法^[13]的核心思想是聚类中心的选取,该方法生成的聚类中心是由一些局部密度比周围点更大,且距其他高密度点较远的点组成的。考虑待聚类的数据集 $S = \{x_i\}_{i=1}^N$, 对于任意 $x_i \in S$, 分别定义其局部密度 ρ_i 和距离参数 σ_i , 得到每个点的 ρ_i 和 σ_i 之后,通常选择 ρ_i 和 σ_i 较大的点作为聚类中心,该类中心点具有较高的局部密度,且和其他高密度点距离较大。将这些点作为聚类中心,并将剩下的点划分至对应的簇中。因此,该方法能解决电信业投诉内容在无先验知识的情况下,根据投诉文本数据对象之间的相似性来完成文本分类,无需预先确定聚类数目。同时聚类处理过程中对文本中可能包含的多种语言的文本数据或形状结构复杂的类都有不错的效果,并能较好地防止噪声点干扰。

2.2 N-gram 语言模型

传统语言模型仅使用无序关键词集合来描述热点话题,即关键词之间的顺序和原本语料中的次序无关,这在新闻或网络热点话题检测中是可行的。如“一带一路 峰会 中国”,对关键词次序进行调整不会影响对于热点话题的语义理解。然而在电信行业业务相关热点话题中,关键词多为业务相关词汇,且词序不同有可能导致语义明显差别,如“用户 滞纳金 减免”的语义表达相对

“滞纳金 减免 用户”就更为明确,因此需要根据原本语料中的词序对关键词进行排序,从而有助于投诉热点话题相关的处理业务人员对于话题描述关键词语义的准确理解。

因此,本文引入 N -gram 模型^[14]对关键词进行排序,从而获得热点话题更合理的语义描述方式。

N -gram 模型也称为 $N-1$ 阶 Markov 模型,它基于如下假设:当前词出现的概率仅与前面的词相关。假设句子 S 由单词序列 $w_1, w_2, \dots, w_n$ 组成:

$$S = w_1, w_2, w_3, \dots, w_n \quad (1)$$

则根据条件概率公式, S 的出现概率可以表示为:

$$\begin{aligned} p(S) &= p(w_1) p(w_2 | w_1) \dots p(w_n | w_1 w_2 \dots w_{n-1}) \\ &= \prod_{k=1}^n p(w_k | w_1 w_2 \dots w_{k-1}) \end{aligned} \quad (2)$$

根据假设,式(2)等价于:

$$p(S) = \prod_{k=1}^n p(w_k | w_{k-n+1} w_{k-n+2} \dots w_{k-1}) \quad (3)$$

特别地,当 N 取 2 时,有:

$$p(S) = p(w_1) p(w_2 | w_1) p(w_3 | w_2) \dots p(w_n | w_{n-1}) \quad (4)$$

式(4)中的条件概率即 N -gram 模型的参数,一般通过语料库进行训练得到,即:

$$p(w_k | w_{k-n+1} w_{k-n+2} \dots w_{k-1}) = \frac{C(w_{k-n+1} w_{k-n+2} \dots w_{k-1} w_k)}{C(w_{k-n+1} w_{k-n+2} \dots w_{k-1})} \quad (5)$$

其中, $C(\cdot)$ 表示序列在语料库中的出现频率。训练语料规模越大,参数估计越准确。

N -gram 模型中的 N 值可取 1、2、3,不同 N 值对应模型名称分别为 Unigram 模型、Bigram 模型和 Trigram 模型。虽然 N 值增大可以更好地刻画语言结构,但同时会造成计算量大、数据稀疏等问题^[15]。针对本文的排序问题,在计算中取 $N=2$ 或 $N=3$ 都可满足实际应用,综合考虑计算效率和排序语义描述准确性,本文建议取 $N=2$ 。

3 算法描述

3.1 算法构成与流程

本文提出的基于峰值密度聚类的电信业投诉

热点话题检测方法通过构造相应的算法来实现最终的热点话题检测功能。该算法主要由预处理、文本聚类以及热点话题描述 3 部分组成,预处理主要实现对电信行业投诉样本文档进行分词及特征向量表示;文本聚类主要通过文本相似度计算以及峰值密度聚类算法对经预处理的投诉文本集合进行聚类;热点话题描述则实现关键词筛选并结合 N -gram 模型进行热点话题的关键词语义化描述。基于峰值密度聚类的电信业投诉热点话题检测算法流程如图 1 所示。

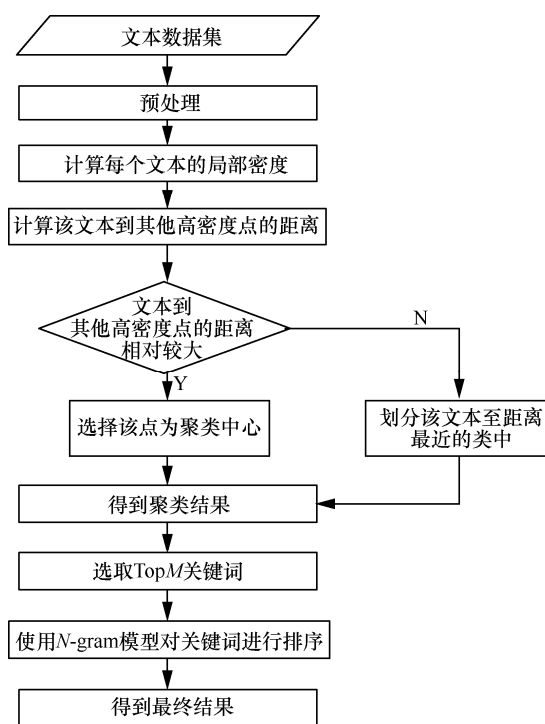


图 1 算法流程

算法首先对文本集进行预处理,采用向量空间模型描述文本,在相似度矩阵的基础上计算每个文本的局部密度 ρ_i 和距离参数 σ_i ,选取聚类中心并进一步得到聚类结果。在聚类结果基础上,对每个类簇提取关键词,引入 N -gram 模型对关键词进行排序,得到最终结果。

3.2 预处理

本文算法主要应用于电信投诉热点话题检测,与传统的网络热点话题检测不同,投诉话题



包含大量的业务用语及专业术语, 因此在进行热点话题检测之前, 需要对文本进行预处理, 主要包含以下工作。

(1) 为了提升分词的精确性, 建立了与电信相关的专用词典以及停用词词典, 共包含数千个重点业务相关词汇, 并标注词性。

(2) 去除投诉文本中包含的工单号、手机号码等数字。

(3) 采用 HanLp 分词工具对投诉话题样本进行分词, 并去除停用词以及无用高频词 (如“谢谢”“请处理”等)。

采用向量空间模型^[16]表示分词后的文本, 在该模型下, 每个文本可用规范化特征向量表示如下:

$$d = \{(t_1, w_1), (t_2, w_2), \dots, (t_n, w_n)\} \quad (6)$$

其中, t_i 表示第 i 个特征项, w_i 表示第 i 个特征项的权重。本文采用 tf-idf 计算特征项的权重, 表达式如下:

$$w_i = \text{tf}_i \times \text{idf}_i \quad (7)$$

其中, tf_i 为第 i 个特征项的词频, idf_i 为第 i 个特征项的逆文档频率, 表达式为:

$$\text{idf}_i = \log \frac{N+1}{N_i+1} \quad (8)$$

其中, N 为总文档数, N_i 为包含词语 i 的文档数。

3.3 相似度计算及文本聚类

为了对文本进行聚类, 首选需要解决相似度计算的问题, 常用的文本相似度计算方式包括欧氏距离、余弦相似度、Jaccard 相似度等, 本文采用余弦相似度进行计算, 如下所示:

$$\text{sim}(d_i, d_j) = \frac{\sum_{k=1}^n w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^n w_{ik}^2} \cdot \sqrt{\sum_{k=1}^n w_{jk}^2}} \quad (9)$$

其中, d_i 代表第 i 个文本, w_{ik} 代表第 i 个文本第 k 个分量的权值, 由式 (9) 可知, 文本相似度等于其特征向量之间的夹角, 夹角值越大, 文档相似度

越高, 反之则越低。

文本之间的距离可以通过相似度得到, 本文采用式 (10) 计算距离:

$$d_{ij} = \begin{cases} \frac{1}{\text{sim}(d_i, d_j)}, & \text{sim}(d_i, d_j) \neq 0 \\ \gamma, & \text{sim}(d_i, d_j) = 0 \end{cases} \quad (10)$$

其中, d_{ij} 代表文本 d_i 和 d_j 之间的距离, 当 $\text{sim}(d_i, d_j) \neq 0$ 时, d_{ij} 为 $\text{sim}(d_i, d_j)$ 的倒数, 当 $\text{sim}(d_i, d_j) = 0$ 时, $d_{ij} = \gamma$, γ 为足够大的正数, 本文取 $\gamma = 10 \max \{d_{ij}, 0 \leq i \leq N, 0 \leq j \leq N\}$, 即 γ 等于 d_{ij} 最大值的 10 倍。

使用聚类算法之前, 首先要计算局部密度 ρ_i , 包括截断密度和高斯密度两种计算方式, 本文选取高斯密度, 表达式如下:

$$\rho_i = \sum_{j \in S \setminus \{i\}} e^{-\left(\frac{d_{ij}}{d_c}\right)^2} \quad (11)$$

其中, d_{ij} 为 d_i 和 d_j 之间的距离, d_c 为距离阈值。由定义可知, 相对离散的截断密度来说, 连续的高斯密度不易产生冲突, 且同样满足与 d_i 的距离小于 d_c 的点越多, ρ_i 的值越大这一条件。采用如下方法确定 d_c , 将数组 $d = \{d_{ij}, 0 < i < j < N\}$ 按降序排列, 取 $d_c = d[\lambda h]$, 其中 h 为数组 d 的总长度, λ 为小于 1 的系数, 一般为 0.01~0.05。

σ_i 的计算表达式如下:

$$\sigma_i = \begin{cases} \min_{j \in 1, \dots, n \setminus \{i\} \wedge \rho_j > \rho_i} d_{ij}, & x_i \text{ 为局部密度最大的点} \\ \max_{j \in 1, \dots, n \setminus \{i\}} d_{ij}, & x_i \text{ 不为局部密度最大的点} \end{cases} \quad (12)$$

由式 (12) 可知, 当文本 d_i 具有最大的局部密度时, σ_i 表示集合 S 中点与 d_i 的最大距离; 否则, σ_i 表示所有局部密度大于 d_i 的点与 d_i 距离的最小值。

对于每个文本 d_i , 生成对应的 (ρ_i, σ_i) 组合, 根据决策图选择 ρ_i 和 σ_i 都相对较大的点作为聚类中心, 将剩余点划分至距离最近的聚类中心所在的类簇中, 完成文本聚类过程。

3.4 热点话题描述

在聚类结果基础之上, 通过 tf-idf 值提取每个类

中的关键词,选取排名前 M 的关键词作为话题描述,并在 N -gram 模型的基础上对关键词进行排序。

通过给定的语料训练 N -gram 模型,确定模型参数,计算每个关键词集合概率最大的排列方式。例如对于由{用户, 滞纳金, 减免} 3 个关键词组成的集合,共存在 6 种排列方式。需要依次计算各种方式的概率,并取出现概率最大的排列方式。

在实际计算中,由于参数空间过大,会导致数据稀疏问题^[17],即许多参数在训练语料中未出现,从而导致概率为 0 的情况,本文引入平滑技术以解决这一问题,将式 (5) 转为如下形式:

$$p(w_k | w_{k-n+1} w_{k-n+2} \dots w_{k-1}) = \frac{C(w_{k-n+1} w_{k-n+2} \dots w_{k-1} w_k) + \pi}{C(w_{k-n+1} w_{k-n+2} \dots w_{k-1}) + \pi |V|} \quad (13)$$

其中, π 为小于 1 的正数,本文取 $\pi=0.1$, V 为所有可能不同的 N -gram 数量。采用式 (13) 代替式 (5) 解决了数据稀疏的问题。进一步,为了防止概率过小导致的数据下溢问题,采用取自然对数进行计算的方式,即令:

$$\begin{aligned} \ln p(S) &= \ln \prod_{k=1}^n p(w_k | w_{k-n+1} w_{k-n+2} \dots w_{k-1}) \\ &= \sum_{k=1}^n \ln p(w_k | w_{k-n+1} w_{k-n+2} \dots w_{k-1}) \end{aligned} \quad (14)$$

综上,热点话题描述主要在 N -gram 模型的基础上,首先得到关键词的全排列组合,然后计算每种组合的生成概率,取概率最大的组合方式作为热点话题描述,其完整过程如算法 1 所示。

算法 1 求热点话题描述

输入 N -gram 模型, 关键词集合 φ

输出 关键词排序 answer

对于给定关键词集合 φ , 获取其对应全排列集合 Φ , 定义空集合 ψ ;

for each $\Phi_i \in \Phi$

next = 1

count $\ln(\Phi_i)$ with (13) and (14)

temp $\leftarrow \ln(\Phi_i)$ and add temp to ψ

next++

end for

$\psi_{\max} \leftarrow \max(\psi_i)$, index $\leftarrow i$, $\max(\psi_i)$ donates the max value in ψ and i donates the index of max value in ψ

answer $\leftarrow \Phi_{\text{index}}$

使用第 2.2 节给出的关键词组合{用户、滞纳金、减免}对算法进行具体描述。首先获取该组合的 6 种排序方式: {用户, 滞纳金, 减免}、{用户, 减免, 滞纳金}、{减免, 用户, 滞纳金}、{减免, 滞纳金, 用户}、{滞纳金, 减免, 用户}、{滞纳金, 用户, 减免}。通过式 (13) 以及式 (14) 分别计算各组合的出现概率, 取概率最大的组合方式作为热点话题描述。

4 实验结果及分析

4.1 实验数据说明

本文所用数据来源于某电信公司业务部门提供的 2016 年 1—12 月的投诉工单, 共包含投诉单 13 218 余件, 从中选取工单得到 3 个数据集 data1、data2 以及 data3, 其中 data1 共包含 2 592 条工单 14 个事件话题, data2 共包含 5 196 条工单 24 个事件话题, data3 包含 7 408 条工单 32 个事件话题。采用剩余投诉单作为训练集对 N -gram 模型进行训练。

采用准确率、召回率和 F 值^[18]作为评价指标。准确率和召回率是信息检索领域中两个较为重要的指标, 准确率反映算法对不同主题的区分能力, 准确率越高, 每个类中的内容就越集中; 召回率主要用来衡量算法结果与人工评价的吻合程度, 召回率越高, 则算法结果越符合实际情况; F 值是综合准确率和召回率的评价指标, 其值介于 0 到 1 之间, 反映了算法的综合性能。本文采用的精确度与召回率的计算式如下:

$$\text{准确率} = \frac{\text{真实发生的事件数量}}{\text{检测出的总事件数量}} \quad (15)$$



$$\text{召回率} = \frac{\text{不重复出现的真实事件数量}}{\text{被检测出的总事件数量}} \quad (16)$$

$$F\text{值} = \frac{2 \times \text{准确率} \times \text{召回率}}{\text{准确率} + \text{召回率}} \quad (17)$$

4.2 实验数据及测评方法

针对以上测试数据集,采用本文算法、Jarvis-Patrick 算法、DBSCAN 算法以及 ADACC K -means 聚类算法进行相关热点话题检测,在文本聚类领域获得了广泛的应用。Jarvis-Patrick 算法^[19-20]是一类基于 SNN (spiking neuron network) 相似度的聚类方法,能够处理带有噪声的数据集合,对于高维数据的分析输出拥有良好的效果。 K -means 算法是一类应用广泛的聚类算法,ADACC K -means 算法^[21]是 K -means 算法的改进,该算法需预设以下参数:领域半径 r ,高密度点个数 M 。实验环境为:处理器为 Intel Core i5-6300HQ 2.3 GHz,8 GB 内存,Win10 64 位操作系统。

各算法均测试多次,取结果较好的参数组,结果除去噪声点(仅包含一个文本的类簇),算法参数设置以及测试结果分别见表 1 和表 2。

表 1 算法参数设置

	本文 算法	DBSCAN 算法		Jarvis-Patrick 算法	ADACC K -means 算法	
	λ	minPts	Eps	K	r	M
data1	0.05	5	4.0	5	0.4	10
data2	0.03	6	4.0	6	0.3	20
data3	0.01	6	4.0	7	0.1	30

表 2 测试结果

算法	数据集	准确率	召回率	F 值
本文算法	data1	0.667	0.714	0.690
	data2	0.652	0.625	0.639
	data3	0.677	0.719	0.697
DBSCAN 算法	data1	0.643	0.529	0.580
	data2	0.484	0.625	0.546
	data3	0.687	0.500	0.579
Jarvis-Patrick 算法	data1	0.526	0.714	0.606
	data2	0.609	0.583	0.596
	data3	0.739	0.531	0.618
ADACC K -means 算法	data1	0.563	0.643	0.600
	data2	0.636	0.583	0.608
	data3	0.630	0.567	0.597

由表 2 可知,本文算法整体效果优于其他 3 类比较算法。从各算法的整体表现来看,本文算法最优,Jarvis-Patrick 算法优于 DBSCAN, Jarvis-Patrick 算法存在将簇割裂的情况,这在 data1 和 data3 的测试中表现得较为明显,在 data1 中,算法将大量原本聚集在一起的簇割裂,造成准确率不高,在 data3 中,算法产生了大量噪声点,影响算法的召回率。DBSCAN 算法的参数需要多次调整,且结果对参数变化较为敏感,本文算法对参数选择较为顽健。而 ADACC K -means 算法同样存在结果与参数选择关联性较大的问题。综合来看,本文算法优于 DBSCAN 以及 Jarvis-Patrick,表 3 给出了 data2 中排名前 5 ($M=5$) 的热点话题及热度值,与设定值

表 3 热点话题及热度值

事件编号	特征词	关键词	热度	对应人工标注的话题名称	对应人工标注的热度
1	用户、要求减免、滞纳金、金额、欠费、违约、拆机、帮忙、处理...	用户、要求减免、滞纳金、金额、欠费	0.189	客户要求滞纳金减免	0.081
2	翼支付、账户、周末、活动、消费、未减免、享受、没有...	翼支付、账户、周末、活动、未减免	0.101	翼支付周末活动未减免	0.085
3	手机号码、实名登记、实名、违章停机、复机、核实、号码、小户名、实名制	手机号码、实名登记、实名、违章停机、复机	0.076	实名制违章停机复机	0.069
4	号码、交费、发票、无法打印、提示、前台、充值、销账、11888...	号码、交费、发票、无法打印、提示	0.050	发票无法打印	0.046
5	预订单、取消、提示、关联、完成、导购、已完成、提示...	预订单、取消、提示、关联、完成	0.050	预订单关联取消	0.030

接近,进一步说明了本文算法的实用价值。

5 结束语

本文提出了一种基于密度峰值聚类的热点话题检测方法,在文本预处理的基础上,应用密度峰值聚类算法对文本进行聚类,实验结果表明,与其他聚类算法比较,本文算法取得了较高的准确率、召回率和 F 值,热点话题检测结果更好,如何对算法效率进行进一步的优化,使之能够适用于更大规模的文本集以及在线检测,是下一步的研究目标。

参考文献:

- [1] LI B L, YU S W. Research on topic detection and tracking[J]. Computer Engineering and Applications, 2003, 39(17): 7-10.
- [2] 关晓惠, 钱亚冠, 孙欣欣. 一种改进的基于局部密度的聚类算法[J]. 电信科学, 2016, 32(1): 54-59.
- [3] GUAN X H, QIAN Y G, SUN X X. An improved clustering algorithm basing on local density[J]. Telecommunications Science, 2016, 32(1): 54-59.
- [3] 周坚, 石永革, 何美斌. 基于 A-D 模型的 K-means 算法在通话异常客户挖掘中的应用[J]. 电信科学, 2018, 34(4): 81-89.
- [4] ZHOU J, SHI Y G, HE M B, et al. Application of K-means algorithm based on A-D model in calling abnormal customer mining[J]. Telecommunications Science, 2018, 34(4): 81-89.
- [4] MOHD M, CRESTANI F, RUTHVEN I. Construction of topics and clusters in topic detection and tracking tasks[C]// International Conference on Semantic Technology and Information Retrieval, June 28-29, 2011, Putrajaya, Malaysia. Piscataway: IEEE Press, 2011: 171-174.
- [5] ZHANG X M, LI Z J, CHAO W H. Research of automatic topic detection based on incremental clustering[J]. Journal of Software, 2012, 23(6): 1578-1587.
- [6] WANG W, YANG W, QI H F. Network hotspot topic detection algorithm based on multi-center model[J]. Journal of Nanjing University of Science and Technology, 2009, 33(4): 422-421.
- [7] CHENG W, LONG Z Y. Online topic detection algorithm for internet news[J]. Computer Engineering, 2009, 35(18): 28-38.
- [8] LI C, SUN A, DATTA A. Tweepent: segment-based event detection from tweets[C]//International Conference on Information and Knowledge Management, Oct 29-Nov 2, 2012, San Francisco, USA. Piscataway: IEEE Press, 2012: 155-164.
- [9] VISWANATH P, SURESH B V. Rough-DBSCAN: a fast hybrid density based clustering method for large data sets[J]. Pattern Recognition Letters, 2009, 30(16): 1477-1488.
- [10] KUMAR K M, RAMA M R A. A fast DBSCAN clustering algorithm by accelerating neighbor searching using groups method[J]. Pattern Recognition, 2016(58): 39-48.
- [11] 谢静瑶, 解思江, 焦阳, 等. 一种改进的启发式自适应 DBSCAN 聚类算法的研究及其在电力系统信息安全预警分析中的应用[J]. 电信科学, 2017, 33(Z1): 117-122.
- [11] XIE J Y, XIE S J, JIAO Y, et al. Research on an improved heuristic auto-adapted DBSCAN clustering algorithm and the application in power system information security early-warning analysis[J]. Telecommunications Science, 2017, 33(Z1): 117-122.
- [12] SUN M X, LIU C Q. Research on hot topic detection based on DBSCAN algorithm and inter sentence relationship[J]. Library & Information Service, 2017(12): 113-121.
- [13] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.
- [14] WEI M, XU W R, JUN G. A chinese text classifier based on N-gram language model and chain augmented naïve bayesian classifier[J]. Journal of Chinese Information Processing, 2006, 20(3): 29-35.
- [15] WU Y, WEI G, LI H. Word segmentation algorithm for chinese language based on N-gram models and machine learning[J]. Journal of Electronics and Information Technology, 2001, 23(11): 1148-1153.
- [16] TANG G, XIA Y Q, ZHANG M, et al. Cross-lingual document clustering based on similarity space model[J]. Journal of Chinese Information Processing, 2016, 26(2): 116-120.
- [17] WU C Y. Model of Chinese words rough segmentation based on bi-gram and N-most-probability method[J]. Journal of Computer Applications, 2007, 27(2): 2902-2905.
- [18] MA C L, SHAN H, MA T. Improved density peaks based clustering algorithm with strategy choosing cluster center automatically[J]. Computer Science, 2016, 43(7): 255-280.
- [19] CAO Z W, ZHOU Y. Design and implementation of JP algorithm based on MapReduce[J]. Computer Engineering, 2012, 38(24): 14-16.
- [20] VATHY-FOGARASSY A, KISS A, ABONYI J. Improvement of Jarvis-Patrick clustering based on fuzzy similarity[C]//International Workshop on Fuzzy Logic and Applications: Applications of Fuzzy Sets Theory, July 7-10, 2007, Camogli, Italy. New York: ACM Press, 2007: 195-202.
- [21] CHENG Y Y, ZHOU P. Improved K-means clustering algorithm for dynamic allocation cluster center[J]. Computer Technology and Development, 2017, 27(2): 33-36.

[作者简介]



江俊(1983—),男,博士,东方通信股份有限公司工作站及浙江大学流动站博士后,浙江树人大学讲师,主要研究方向为人工智能、数据科学与大数据处理、信息检索等。

黄骅(1983—),男,博士,浙江大学流动站博士后,主要研究方向为文本挖掘、信息检索和自然语言处理等。

任条娟(1965—),女,浙江树人大学信息科技学院院长、教授,中国电子学会理事,主要研究方向为无线传感网和车联网、数据科学与大数据处理等。

张登辉(1970—),男,浙江树人大学信息科技学院副院长、教授,主要研究方向为机器学习与社交计算、数据科学与大数据处理等。