(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110297988 A
(43)申请公布日 2019.10.01

(21)申请号 201910606225.4

(22)申请日 2019.07.06

(71)申请人 四川大学
地址 610065 四川省成都市武侯区一环路南一段24号

(72)发明人 陈兴蜀 蒋术语 王海舟 王文贤 殷明勇 唐瑞 蒋梦婷 李敏毓

(74)专利代理机构 成都禾创知家知识产权代理有限公司 51284
代理人 裴娟

(51)Int.Cl.
G06F 16/9536(2019.01)
G06F 16/35(2019.01)

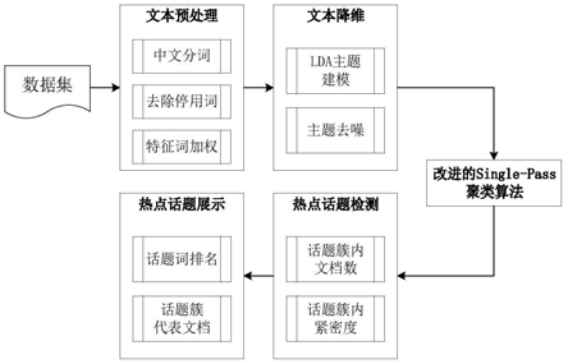
权利要求书2页 说明书9页 附图5页

(54)发明名称

基于加权LDA和改进Single-Pass聚类算法的热点话题检测方法

(57)摘要

本发明公开了一种基于加权LDA和改进Single-Pass聚类算法的热点话题检测方法，包括以下步骤：对文本数据进行预处理，包括中文分词、去除停用词和特征词加权；利用加权LDA主题模型对文本数据进行建模，通过挖掘其中的隐主题信息实现特征降维，并对向量化的结果进行过滤去噪；将经特征词加权的LDA主题模型处理后的文本向量化结果使用改进Single-Pass聚类算法进行聚类；利用话题簇规模和话题簇紧密度计算话题簇的热度值，识别热点话题。本发明检测方法具有算法复杂度低、对文本输入时间顺序依赖性较低等优点。



CN 110297988 A

1.一种基于加权LDA和改进Single-Pass聚类算法的热点话题检测方法,其特征在于,包括以下步骤:

步骤1:对文本数据进行预处理,包括中文分词、去除停用词和特征词加权;

步骤2:利用特征词加权的LDA主题模型对文本数据进行建模,通过挖掘其中的隐主题信息实现特征降维,并对向量化的结果进行过滤去噪;

步骤3:将步骤2中的经特征词加权的LDA主题模型处理后的文本向量化结果使用改进Single-Pass聚类算法进行聚类,即:

1)传入一个向量化后的文本数据d,如果d是数据集合中的第一篇文本,则新建一个话题簇,如果不是,则等待一个时间段T_n,对该时间段内的文本向量进行首先进行传统Single-Pass聚类;

2)将传统Single-Pass聚类后的结果与前一个时间段的聚类结果进行相似度对比:计算该批文本数据聚类得到的各个话题簇质心向量与已有的各个话题簇中的质心向量之间的相似度;

3)保留该批次文本向量各个话题簇的最大相似度并与阈值比较,如果大于阈值则归入与之相似度最大的原话题,否则新建一个话题;

4)更新话题簇,等待下一批向量化文本数据的传入;

步骤4:利用话题簇规模和话题簇紧密度计算话题簇的热度值,识别热点话题,即:

统计步骤3中每个话题簇中的文档数目,并对其进行归一化处理,再按以下方式获取话题簇k的规模c_k:

$$c_k = \frac{|D_k|}{|D_{\max}|}$$

其中,$|D_k|$是指话题簇k中包含的文档数目,$|D_{max}|$指最大话题簇中的文档总数;按以下方式获取话题簇k紧密度u_k:

$$u_k = \frac{\sum_{\vec{d_m},\vec{d_n}\in D_k} \cos(\vec{d_m},\vec{d_n})}{|D_k|*(|D_k|-1)/2}$$

$$\cos(\vec{d_m},\vec{d_n}) = \frac{\vec{d_m}*\vec{d_n}}{\|\vec{d_m}\|*\|\vec{d_n}\|}$$

其中,$\vec{d_m}$是话题簇k中第m篇文档利用“词频-逆话题频率”方法加权处理后的向量化表示;从话题簇规模和紧密度两个方面综合考虑,得到话题簇的热度,如下式:

$hot(k) = \eta * c_k + \lambda * u_k$

其中η是话题簇规模的权重,λ是话题簇紧密度的权重,η+λ=1。

2.如权利要求1所述的基于加权LDA和改进Single-Pass聚类算法的热点话题检测方法,其特征在于,在步骤1中,中文分词具体为:采用中科院汉语分词系统实现文本的分词、词性标注及命名实体识别工作。

3.如权利要求1所述的基于加权LDA和改进Single-Pass聚类算法的热点话题检测方法,其特征在于,第i个特征词t_i加权的具体方式为:

$$pos(t_i)=\begin{cases}3, & \text{如果}t_i\text{属于命名实体或者标签}\\ 1, & \text{其他}\end{cases}$$

其中pos(t$_i$)代表特征词t$_i$的词性权重。

4.如权利要求1所述的基于加权LDA和改进Single-Pass聚类算法的热点话题检测方法,其特征在于,还包括步骤5:基于话题词排序算法和文档距离计算对识别出的热点话题进行展示。

5.如权利要求4所述的基于加权LDA和改进Single-Pass聚类算法的热点话题检测方法,其特征在于,所述步骤5中的话题词排序算法具体为:

根据步骤4得到的不同热度话题簇,采用“词频-逆话题频率”的方法对每个话题簇内的话题词计算权重,再按权重排序;话题词权重得获取方式为:

$$w_{i,k}=\sqrt{n_k^{(w_i)}}\cdot\log\frac{|D_k|}{kf_{w_i}+1}$$

其中,w$_{i,k}$是文本中第i个单词w$_i$在话题簇k中的权重,$n_k^{(w_i)}$指的是单词w$_i$分配给话题簇k的次数,kf_{w_i}表示包含至少一次单词w$_i$的话题个数。

6.如权利要求4所述的基于加权LDA和改进Single-Pass聚类算法的热点话题检测方法,其特征在于,所述步骤5中的文档距离计算具体为:

采用Jensen-Shannon距离D$_{JS}$来度量d$_m$和d$_n$两个文档之间的相似度,其计算公式为:

$$D_{JS}(d_m\|d_n)=\frac{1}{2}D_{KL}(d_m\|Q)+\frac{1}{2}D_{KL}(d_n\|Q)$$

其中,Q=(d$_m$+d$_n$)/2,D$_{KL}$为文档向量之间的相对熵;由此得到话题簇中第m篇文档到簇内其它文档的总距离D(d$_m$)获取方式如下:

$$D(d_m)=\sum_{d_n\in D_k,d_n\neq d_m}D_{JS}(\theta_m\|\theta_n)$$

其中,θ$_m$是文档m的文档-主题分布,D$_k$为话题簇k的文档集合,d$_m$,d$_n$为D$_k$中的第m篇,第n篇文档。

基于加权LDA和改进Single-Pass聚类算法的热点话题检测方法

技术领域

[0001] 本发明涉及热点话题检测技术领域，具体为一种基于特征词加权的隐含狄利克雷分布(Latent Dirichlet Allocation,LDA)主题模型和改进Single-Pass聚类算法的热点话题检测方法。

背景技术

[0002] 热点话题是一段时间内，围绕某一事件的相关新闻报道、微博信息被大量用户讨论和分享，造成该事件被广泛关注，最终形成全网范围内的话题焦点。热点话题检测是舆情监控及引导工作中的重要任务之一，它通过对海量的实时数据进行及时有效的处理，挖掘文本数据中的话题结构，展示当前互联网中用户关注的话题焦点及其相关内容，为舆情监控者及普通用户掌握当前的热点话题发展趋势提供便捷准确的参考。

[0003] 近年来，互联网保持着高速发展的趋势，网络信息容量、网民数量都呈现出爆炸式的增长趋势，网络已经成为人们获取信息的主要渠道。根据中国互联网络信息中心(CNNIC)2019年2月发布的《第43次中国互联网络发展状况统计报告》显示，截至2018年12月，我国网民规模已经达到8.29亿，与2017年相比增长了5653万人，年增长率为3.8％，互联网普及率达到59.6％。随着网络成为人们日常生活中不可或缺的信息传播新媒体，互联网这一“虚拟社会”与真实社会之间的互动越来越频繁，互联网正逐渐呈现出社会化特征。通过互联网传播的信息包含了民众对当前社会各种热点现象及问题的观点和想法，主要涉及政治、军事、科技、经济、体育、娱乐等各个领域。

[0004] 但由于网络中的消息冗余繁杂，仅仅依靠人工查找新闻话题难以应对网络中海量信息的处理并对其中的敏感主题及时做出反应。尤其对于决策者，要监控网络中所有相关的信息是不现实的，如果没有自动化的工具支持，很难及时的做出正确的决断，所以人们希望可以通过计算机来自动获取热门新闻话题，从而提高网络监管能力及处置网络舆情突发事件的能力。更为重要的是，在一些安全机构针对网络犯罪的检测和预防过程中，能快速准确地检测出相关话题并及时应对就显得尤为重要。

发明内容

[0005] 本发明所要解决的技术问题是提供一种基于加权LDA和改进Single-Pass聚类算法的热点话题检测方法，其具有算法复杂度低、对文本输入时间顺序依赖性较低等优点。

[0006] 为解决上述技术问题，本发明采用的技术方案是：

[0007] 一种基于加权LDA和改进Single-Pass聚类算法的热点话题检测方法，包括以下步骤：

[0008] 步骤1:对文本数据进行预处理，包括中文分词、去除停用词和特征词加权；

[0009] 步骤2:利用特征词加权的LDA主题模型对文本数据进行建模，通过挖掘其中的隐主题信息实现特征降维，并对向量化的结果进行过滤去噪；

[0010] 步骤3:将步骤2中的经特征词加权的LDA主题模型处理后的文本向量化结果使用改进Single-Pass聚类算法进行聚类,即:

[0011] 1)传入一个向量化后的文本数据d,如果d是数据集合中的第一篇文本,则新建一个话题簇,如果不是,则等待一个时间段T_n,对该时间段内的文本向量进行首先进行传统Single-Pass聚类;

[0012] 2)将传统Single-Pass聚类后的结果与前一个时间段的聚类结果进行相似度对比:计算该批文本数据聚类得到的各个话题簇质心向量与已有的各个话题簇中的质心向量之间的相似度;

[0013] 3)保留该批次文本向量各个话题簇的最大相似度并与阈值比较,如果大于阈值则归入与之相似度最大的原话题,否则新建一个话题;

[0014] 4)更新话题簇,等待下一批向量化文本数据的传入;

[0015] 步骤4:利用话题簇规模和话题簇紧密度计算话题簇的热度值,识别热点话题,即:

[0016] 统计步骤3中每个话题簇中的文档数目,并对其进行归一化处理,再按以下方式获取话题簇k的规模c_k:

[0017] $$c_k = \frac{|D_k|}{|D_{\max}|}$$

[0018] 其中,$|D_k|$是指话题簇k中包含的文档数目,$|D_{max}|$指最大话题簇中的文档总数;按以下方式获取话题簇k紧密度u_k:

[0019] $$u_k = \frac{\sum_{\vec{d_m},\vec{d_n}\in D_k} \cos(\vec{d_m},\vec{d_n})}{|D_k| * (|D_k|-1)/2}$$

[0020] $$\cos(\vec{d_m},\vec{d_n}) = \frac{\vec{d_m} * \vec{d_n}}{\|\vec{d_m}\| * \|\vec{d_n}\|}$$

[0021] 其中,$\vec{d_m}$是话题簇k中第m篇文档利用“词频-逆话题频率”方法加权处理后的向量化表示;从话题簇规模和紧密度两个方面综合考虑,得到话题簇的热度,如下式:

[0022] $hot(k) = \eta * c_k + \lambda * u_k$

[0023] 其中η是话题簇规模的权重,λ是话题簇紧密度的权重,η+λ=1。

[0024] 进一步的,在步骤1中,中文分词具体为:采用中科院汉语分词系统实现文本的分词、词性标注及命名实体识别工作。

[0025] 进一步的,在步骤1中,第i个特征词t_i加权的具体方式为:

[0026] $$pos(t_i) = \begin{cases} 3, & \text{如果} t_i \text{属于命名实体或者标签} \\ 1, & \text{其他} \end{cases}$$

[0027] 其中$pos(t_i)$代表特征词t_i的词性权重。

[0028] 进一步的,还包括步骤5:基于话题词排序算法和文档距离计算对识别出的热点话题进行展示。

[0029] 进一步的,所述步骤5中的话题词排序算法具体为:

[0030] 根据步骤4得到的不同热度话题簇,采用“词频-逆话题频率”的方法对每个话题簇

内的话题词计算权重，再按权重排序；话题词权重得获取方式为：

[0031] $$w_{i,k}=\sqrt{n_k^{(w_i)}}\cdot\log\frac{|D_k|}{kf_{w_i}+1}$$

[0032] 其中，$w_{i,k}$是文本中第i个单词w_i在话题簇k中的权重，$n_k^{(w_i)}$指的是单词w_i分配给话题簇k的次数，kf_{w_i}表示包含至少一次单词w_i的话题个数。

[0033] 进一步的，所述步骤5中的文档距离计算具体为：

[0034] 采用Jensen-Shannon距离D_{JS}来度量d_m和d_n两个文档之间的相似度，其计算公式为：

[0035] $$D_{JS}\left(d_m\|d_n\right)=\frac{1}{2}D_{KL}\left(d_m\|Q\right)+\frac{1}{2}D_{KL}\left(d_n\|Q\right)$$

[0036] 其中，$Q=(d_m+d_n)/2$，D_{KL}为文档向量之间的相对熵；由此得到话题簇中第m篇文档到簇内其它文档的总距离$D(d_m)$获取方式如下：

[0037] $$D(d_m)=\sum_{d_n\in D_k,d_n\neq d_m}D_{JS}(\theta_m\|\theta_n)$$

[0038] 其中，θ_m是文档m的文档-主题分布，D_k为话题簇k的文档集合，d_m，d_n为D_k中的第m篇，第n篇文档。

[0039] 与现有技术相比，本发明的有益效果是：

[0040] 1)本发明对话题中的特征词(命名实体)赋予了相比于动词、名词更大的权重，增强了不同主题之间的可区分性和LDA模型的建模能力；

[0041] 2)本发明引入“话题中心”的概念来表示一个话题簇，将文本向量相似度的计算次数降低到话题簇个数的规模大小，算法复杂度与传统Single-Pass聚类算法相比普遍降低了至少十倍以上；

[0042] 3)本发明中改进Single-Pass聚类算法中的文件批处理的方法降低了Single-Pass聚类算法中文本输入顺序对聚类效果的影响，提高了聚类算法的稳定性；

[0043] 4)本发明从话题簇内的文档数目和文档紧密度两个方面考虑，计算话题的热度值，改进了话题的聚类效果。

附图说明

[0044] 图1为本发明的热点话题检测框架图；

[0045] 图2本发明的改进后的Single-Pass算法流程图；

[0046] 图3为本发明的新闻特征词加权与否的困惑度对比；

[0047] 图4为本发明的微博特征词加权与否的困惑度对比；

[0048] 图5为K-means算法、K-means++算法、传统Single-Pass算法和改进的Single-Pass聚类算法运行时间对比(日、周)；

[0049] 图6为使用本发明改进的方法与使用传统的Single-Pass方法的新闻数据困惑度对比；

[0050] 图7为使用本发明改进的方法与使用传统的Single-Pass方法的微博数据困惑度对比。

具体实施方式

[0051] 下面结合附图和具体实施方式对本发明做进一步详细说明。

[0052] 如图1所示，本发明方法输入为中文文本，输出为热点话题(包括排名后的话题词和话题簇代表文档)。首先对文本数据进行预处理，包括分词、停用词过滤、特征词加权等，然后利用LDA主题模型对其建模并对向量化的文本进行过滤去噪；接着基于改进的Single-Pass算法对降维后的文本进行聚类；最后通过热点话题检测方法识别话题簇中的热点话题，并采用话题词排名算法和文档距离计算公式对热点话题进行展示。详述如下：

[0053] 步骤1：文本预处理；本发明的文本预处理包括中文分词、去除停用词和特征词加权几个子步骤。

[0054] 1)中文分词

[0055] 中文句子与英文不同，句子中的词语往往是连接在一起的，为了便于利用LDA主题模型对其进行处理，分词成为文本处理的前提。本发明采用中科院汉语分词系统实现文本的分词、词性标注及命名实体识别工作。

[0056] 2)去除停用词

[0057] 停用词即是无区别能力也无描述能力的词，如“我”、“你”和虚词、介词等。本发明仅保留文档集合中的名词、动词和实体标注词汇，去掉常见的停用词和单个字的词语，利用“词频-逆文本频率”方法计算单词权重，每篇文本仅保留权重占比前75％的单词用于实现文本特征的降维。

[0058] 3)特征词加权

[0059] 利用LDA主题模型实现话题建模的过程实际上就是将文本集合从词空间降维到语义空间。在最初的LDA主题模型中，文本集合中的所有单词都被同等对待，这显然是不合理的，因此本发明在特征提取过程中对命名实体进行了加权处理，第i个特征词t_i加权的具体方式为：

[0060] $$pos(t_i)=\begin{cases}3, & \text{如果}t_i\text{属于命名实体}\\ 1, & \text{其他}\end{cases}$$

[0061] 其中$pos(t_i)$代表特征词t_i的词性权重。

[0062] 4)微博数据的预处理

[0063] 新闻文本采用以上方式预处理即可，针对微博数据由于更具特征性，可按如下方式更好的预处理：

[0064] a)使用中科院汉语分词系统提供的新词发现功能，利用采集到的微博历史数据，将其每3000条数据分为一组作为新词发现的一组文本输入，找到新词并存入词典文件中。

[0065] b)在调用分词功能之前，首先导入新词词典文件到系统的用户词典中，判断一条微博文本中是否包含标签符号(##)，如果存在，则提取出其中的主题信息，并对该主题信息和标签以外的其它文本信息分别进行分词，得到的结果利用停用词表进行过滤。

[0066] c)在计算特征词权重时，除了保留微博文本中的动词、名词及实体标注词汇以外，还考虑到文本内容中包含的标签信息。通常一条微博中的标签包含有该微博的主题信息，所以在利用“词频-逆文本频率”方法计算特征词权重时，赋予标签文本更高的权重。根据如下方式进行加权处理：

[0067] $weight(t_i)=\omega_1*pos(t_i)+\omega_2*tag(t_i)$

[0068] 其中，pos(ti)和tag(ti)分表代表第i个特征词ti的词性权重和标签权重，ω1和ω2代表权重因子，本发明取ω1＝ω2＝0.5。改进特征加权的处理方式如下：

[0069] $$pos(t_i)=\begin{cases}3, & 如果t_i\ 属于命名实体\\ 1, & 其他\end{cases}$$

[0070] $$tag(t_i)=\begin{cases}3, & 如果\ t_i\ 属于标签\\ 1, & 其他\end{cases}$$

[0071] d)去除文本长度小于5的微博，这种微博内容包含信息量往往很少且很难准确理解其语义信息。

[0072] e)去除内容只包含表情、链接、图片的微博。

[0073] f)对于转发的微博，它通常会在“//”符号后附带转发的原文信息，为了防止文本的重复出现，本发明过滤掉了转发的原文信息，只保留转发的文本内容。

[0074] 普通LDA模型和特征词加权处理后的LDA模型的建模效果对比：为了检测LDA模型通过特征词加权处理后建模的效果，使用困惑度(Perplexity)作为评价指标。困惑度越小表示模型的预测能力越强，模型的推广性能就越高。困惑度计算公式如下：

[0075] $$Perplexity(D_{test})=\exp\left\{\frac{\sum_{d=1}^{|D_{test}|}\log p(w_d)}{\sum_{d=1}^{|D_{test}|}N_d}\right.$$

[0076] 其中D_{test}表示测试集，$|D_{test}|$表示测试集中的文档数，N_d指文档d的单词数目，$p(w_d)$表示在测试集文档d中每个单词生成的概率。以天为时间片，从每个时间片的数据集中随机选择10％的文档作为测试集，随机选取实2017年12月23日至2017年12月29日的新闻报道和微博文本作为实验数据，分别使用特征词加权处理后的LDA模型和未对特征词加权的LDA模型对训练集建模分析，计算得到新闻困惑度如图3所示，微博困惑度如图4所示。从中可以看出利用特征词加权处理的LDA模型的困惑度均小于未对特征词加权的LDA模型困惑度。这表明对特征词进行加权处理可以提高LDA主题模型的建模能力。由于在特征词加权处理的过程中考虑到命名实体对文本语义的影响，所以利用LDA模型建模的过程中相应特征词的权重会增加，意味着主题-单词分布中对应特征词的分布值也会增大。表1列举了对特征词加权处理前后部分主题的特征词对比情况，从中可以看初对特征词进行加权处理可以有效增加不同主题之间的可区分性。

[0077] 表1特征词加权前后新闻话题对比

[0078]

事件	特征词加权前（前 10 个词）	特征词加权后（前 10 个词）
平昌冬奥会	科技 闭幕式 表演 技术 科技 观众 导演 天气 人工智能 演出	**北京** 闭幕式 表演 **张艺谋** 科技 **平昌** 导演 天气 人工智能 演出
第十三届全国人民代表大会	会议 全国人大 主席 代表团 秘书长 主持 推选 习近平 解放军 出席	全国人大 主席团 **习近平** 代表团 **张德江** 秘书长 **中国** 主持 推选 **王晨**

[0079] 步骤2：利用特征词加权处理的LDA主题模型对文本数据进行建模，通过挖掘其中

的隐主题信息实现特征降维，并对向量化的结果进行过滤去噪；

[0080] 使用步骤1中用特征词加权处理后的LDA主题模型对文本进行建模和采样，得到文档-主题分布参数θ。其中LDA主题在文档上的先验参数α、词语在主题上的先验参数β取经验值α＝50/r，β＝0.01；最优主题数r经贝叶斯方法确定为45。然后文档在各个主题上都会存在一个分布值，值越大表示文档对该话题的贡献越大。然后过滤掉文档-主题分布值小于该阈值的话题，本发明定义文档-主题分布值中最大分布值的一半作为阈值。过滤算法流程描述如下：

Algorithm 4.1: Document Distribution Filtering Algorithm

INPUT: document distribution $\boldsymbol{\theta}_m$

let $p_{upper} = \max\{\theta_{m,1}, \theta_{m,2}, \ldots, \theta_{m,K}\}$ and $p_{lower} = p_{upper}/2$

for $k = 1$ to K **do**

  if $\theta_{m,k} < p_{lower}$ **then**

    $\theta_{m,k} = 0$

[0081]

  end if

end for

calculate $pSum = \sum_k \theta_{m,k}$

for $k = 1$ to K **do**

  $\theta_{m,k} = \theta_{m,k} / pSum$

end for

OUTPUT: $\boldsymbol{\theta}_m$

[0082] 最后将文档-主题分布重新进行归一化处理。

[0083] 步骤3：将步骤2中的经特征词加权的LDA主题模型处理后的文本向量化结果使用本发明提出的改进的Single-Pass聚类算法进行聚类，实现基于文档的主题维度实现话题聚类。

[0084] 本发明中的改进的Single-Pass聚类算法实现的流程如图2所示，改进处在于：用“话题中心”来表示一个话题簇，降低算法计算代价和复杂度；用批量文本处理代替单文本处理，降低文本输入顺序对聚类效果的影响，提高算法稳定性。具体实施方法如下：

[0085] 为了更方便清楚的实施该聚类方法，此处先明确几个概念表示：d_i为第i篇文档；$D = \{d_1, d_2, \ldots, d_M\}$为M个文档的集合；$T_c$为相似度阈值，本发明中微博数据的阈值为0.45，新闻数据的阈值为0.32；两个文本向量d_1、d_2之间的相似度$sim(d_1, d_2)$获取方式如下：

[0086] $$sim(d_1, d_2) = COS(\vec{d_1}, \vec{d_2}) = \frac{\vec{d_1} \cdot \vec{d_2}}{\left|\vec{d_1}\right|\left|\vec{d_2}\right|}$$

[0087] 话题中心用质心向量表示，获取方式如下：

[0088] $$C_k = \frac{1}{N}\sum_{i=1}^{N} d_i$$

[0089] 其中，N表示该话题簇的文本总数。话题中心为C_k(k＝1,2,...,s)，它表示每个话题簇。

[0090] 首先，传入一个向量化后的文本数据d，如果d是数据集合中的第一篇文本，则新建一个话题簇。如果不是，则等待一个时间段T_n，对该时间段内的文本向量进行首先进行传统的Single-Pass聚类。再与前一个时间段的聚类结果进行相似度对比：计算该批文本聚类得到的各个话题簇质心向量与已有的各个话题簇中的质心向量之间的相似度，保留该批次文本向量各个话题簇的最大相似度并与阈值比较，如果大于阈值则归入与之相似度最大的原话题，否则新建一个话题。改进的Single-Pass聚类过程结束，更新话题簇，等待后续文档的传入。

[0091] 以特征词加权处理的LDA模型建模后得到的文本向量化结果作为输入，以漏检率、错检率及检测代价作为评价指标，本发明提出的改进算法与K-means、K-means++、传统Single-Pass算法在话题检测中的效果对比如表2。

[0092] 表2不同算法的话题检测效果对比

[0093]

	K-means 算法	K-means ++算法	传统的 Single-Pass 算法	改进的 Single-Pass 算法
话题数目	8（指定为 8）	8（指定为 8）	17	10

[0094]

漏检率	0.20	0.18	0.24	0.15
错检率	0.003	0.002	0.005	0.0008
检测代价	0.0043	0.0038	0.0053	0.0030

[0095] 从表2中可以得出，本发明提出的改进Single-Pass聚类算法比传统Single-Pass算法得到的话题数更接近真实情况，且漏检率和错检率均低于传统算法。

[0096] 再选3月15日这一日和3月12日至3月18日一周的新闻数据，对于一天的数据，改进算法以两小时为时间片进行一次话题聚类检测，如果两小时内新增数据量达到200条则立即进行一次话题聚类检测；对于一周的数据，则以天为时间片进行话题聚类检测。分别计算利用K-means算法、K-means++算法、传统Single-Pass算法和改进的Single-Pass聚类算法的运行时间，如图5所示。从图中可以看出，与K-means算法相比，利用改进的Single-Pass聚类算法进行热点话题检测的时间复杂度大大降低，主要是因为Single-Pass算法基于增量聚类的思想，不需要在输入新数据后对整个数据集重新聚类，因而提高了话题检测的效率，实验数据显示利用改进的聚类算法节省了约40％的时间。同时从图中也可以观察到，改进的Single-Pass算法运行时间比传统Single-Pass算法稍长一点，这主要是因为改进算法利用批处理的思想，文本数据按时间片分批输入，需要多次聚类，因而运行时间会稍长一点，但改进算法减少了传统算法对于文本输入顺序的依赖性，提高了算法稳定性，所以改进的Single-Pass聚类算法对于热点话题检测依然是有意义的。

[0097] 步骤4：利用话题簇规模和话题簇紧密度计算话题簇的热度值，识别热点话题。

[0098] 首先统计步骤3中每个话题簇中的文档数目，并对其进行归一化处理；然后按如下方式获取话题簇k的规模c_k：

[0099] $$c_k = \frac{|D_k|}{|D_{\max}|}$$

[0100] 其中，其中，|D_k|是指话题簇k中包含的文档数目，|D_{max}|指最大话题簇中的文档总数；按以下方式获取话题簇k紧密度u_k：

[0101] $$u_k = \frac{\sum_{\vec{d_m},\vec{d_n} \in D_k} \cos(\vec{d_m},\vec{d_n})}{|D_k| * (|D_k| - 1)/2}$$

[0102] $$\cos(\vec{d_m},\vec{d_n}) = \frac{\vec{d_m} * \vec{d_n}}{\|\vec{d_m}\| * \|\vec{d_n}\|}$$

[0103] 其中，$\vec{d_m}$是指话题簇k中第m篇文档利用“词频-逆话题频率”方法加权处理后的向量化表示；最后，从话题簇规模和紧密度两个方面综合考虑，得到话题簇的热度，如下式：

[0104] $hot(k) = \eta * c_k + \lambda * u_k$

[0105] 其中η是话题簇规模的权重，λ是话题簇紧密度的权重，η+λ＝1。

[0106] 步骤5：基于话题词排名算法和文档距离计算公式对识别出的热点话题进行展示。

[0107] 1)对每个话题簇内的话题词进行排序

[0108] 步骤4中的得到了不同热度的话题簇，然后再采用“词频-逆话题频率”的方法对每个话题簇内的话题词计算权重，再按权重排序。话题词权重得获取方式如下：

[0109] $$w_{i,k} = \sqrt{n_k^{(w_i)}} \cdot \log \frac{|D_k|}{kf_{w_i} + 1}$$

[0110] 其中，$w_{i,k}$是文本中第i个单词w_i在话题簇k中的权重，$n_k^{(w_i)}$指的是单词w_i分配给话题簇k的次数，kf_{w_i}表示包含至少一次单词w_i的话题个数。

[0111] 2)确定话题的代表性文档

[0112] 选择话题簇中最有代表性的文档来表示一个话题簇，即找到每个话题簇中与其它文档最为相似的文档，并用该文档的标题作为热点话题的展示。此处采用Jensen-Shannon距离(用$D_{JS}()$表示)来度量两个文档之间的相似度。Jensen-Shannon距离是基于KL(Kullback-Leibler)距离(即相对熵，用$D_{KL}()$表示)定义的计算公式，主要用于测量两个文档之间概率分布的相似性。KL距离也是用于测量概率分布之间相似性的方法，对于两个文档d_m和d_n，用KL距离计算其相似性是不对称的，即$D_{KL}(d_m||d_n) \neq D_{KL}(d_n||d_m)$。而Jensen-Shannon距离改进了KL距离不对称的缺点，其计算公式如下：

[0113] $$D_{JS}(d_m \| d_n) = \frac{1}{2} D_{KL}(d_m \| Q) + \frac{1}{2} D_{KL}(d_n \| Q)$$

[0114] 其中，$Q = (d_m + d_n)/2$，由此得到话题簇中第m篇文档到簇内其它文档的总距离$D(d_m)$获取方式如下：

[0115] $$D(d_m) = \sum_{d_n \in D_k, d_n \neq d_m} D_{JS}(\theta_m \| \theta_n)$$

[0116] 其中θ_m是文档m的文档-主题分布，θ_n是文档n的文档-主题分布，D_k为话题k的文档集合，d_m，d_n为D_k中的第m篇，第n篇文档。该公式的计算结果越小，表明该文档在话题簇中与其它文档的相似度越高。

[0117] 对步骤4和步骤5得到的3月15日的新闻和微博文本的代表性文档、话题热度、话题词进行展示，选取话题热度值排名前5的话题结果表3、表4所示。

[0118] 表3 3月15日新闻热点话题展示

[0119]

话题编号	话题（代表性文档）	话题词排序	话题热度
40	14%APP 可监听电话，手机厂商致力隐私把关	识别 手机 信息 隐私 厂商 支付 用户 技术 权限 侵犯	0.8656
36	《赛迪机器人 3·15 报告》揭示机器人产品质量 6 大痛点	机器人 认证 检测 产品 中心 漏洞 软件 质量 服务 功能	0.8229
16	美英轮番对俄采取措施	俄罗斯 英国 美国 制裁 总统 中毒 措施 关系 声明 特工	0.7894
37	百度 APP 整改！全国首例个人信息安全公益诉讼撤诉	信息 消费者 诉讼 百度 公益 江苏省 保护 权限 升级 整改	0.7346
29	Win7/8.1/10 集体推送补丁，修复 Intel 两大漏洞	漏洞 处理器 Windows 补丁 更新 Intel 公司 修复 发布 芯片	0.6986

[0120] 表4 3月15日微博热点话题展示

[0121]

话题编号	话题（代表性文档）	话题词排序	话题热度
12	美国宣布美国制裁俄互联网研究机构	俄罗斯 美国 制裁 实施 英国 干预 大选 逻辑 外交官 犯罪	0.7866
6	信息安全意识推广计划	网络 传播 宣言 志愿者 保障 软件 维护 舆论 糟粕 上网	0.7149
4	中国国防建设的整体布局战略思考	网络 强军 以色列 中国 较量 战略 突破口 应对 措施 威胁	0.6434
18	没有芯片安全就没有信息安全	中国 芯片 新闻 邓中翰 核心 人工智能 侵害 权限 标准 程序	0.5523
29	AMD 芯片曝光 12 个高危安全漏洞	AMD 处理器 Ryzen 分析 涉及 消费者 联合 发起 国际 曝光	0.4997

[0122] 图6和图7分别是以随机一周时间的新闻和微博数据为数据输入，基于结合特征词加权和Single-Pass算法改进两个方面对比困惑度的变化情况。通过这两个图可以看出，针对改进的Single-Pass聚类算法的输入文档集合，在其预处理过程中结合特征词加权后，话题检测模型的困惑度更小，也就意味着热点话题检测的效果会更好，从而证明了本发明提出的热点话题检测方法的有效性。

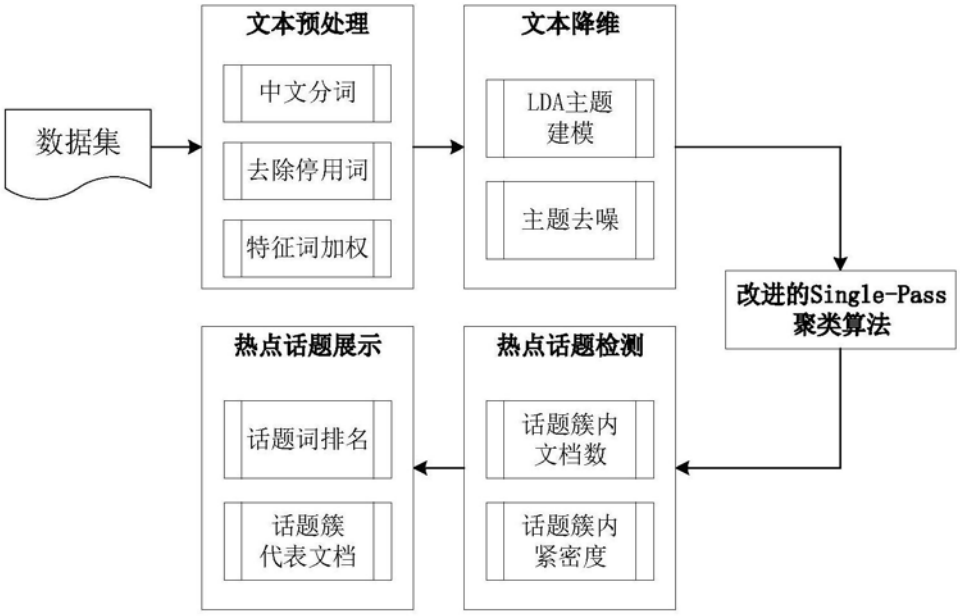

数据集
文本预处理
中文分词
去除停用词
特征词加权
文本降维
LDA主题建模
主题去噪
改进的Single-Pass聚类算法
热点话题检测
话题簇内文档数
话题簇内紧密度
热点话题展示
话题词排名
话题簇代表文档

图1

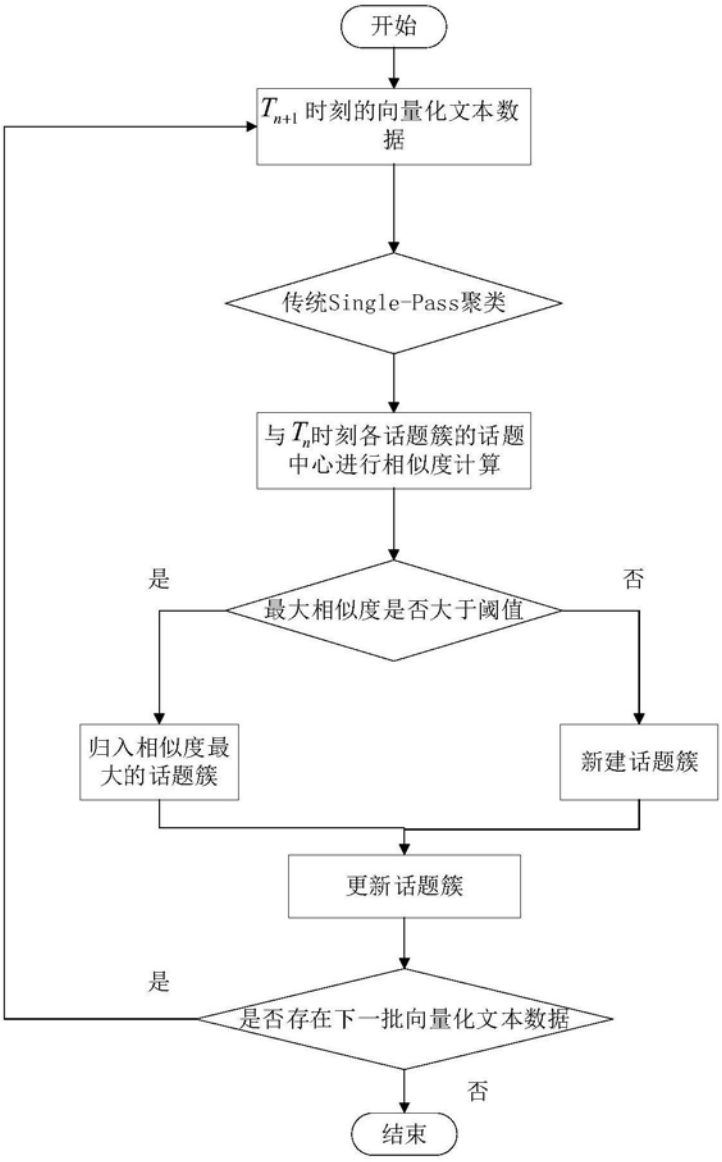

开始
T_{n+1} 时刻的向量化文本数据
传统Single-Pass聚类
与 T_n 时刻各话题簇的话题中心进行相似度计算
是
最大相似度是否大于阈值
否
归入相似度最大的话题簇
新建话题簇
更新话题簇
是
是否存在下一批向量化文本数据
否
结束

图2

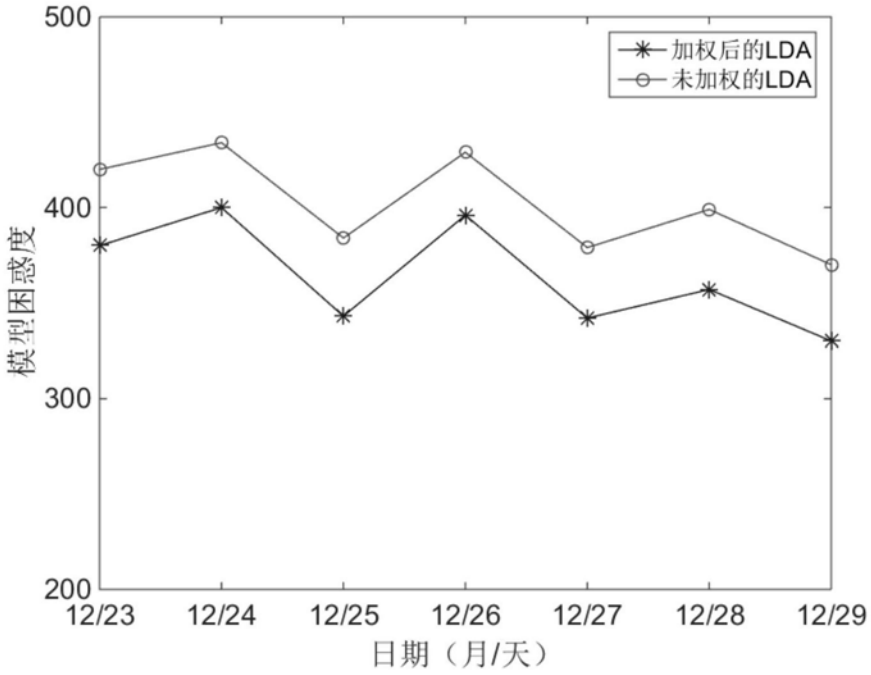

500
400
300
200
模型困惑度
加权后的LDA
未加权的LDA
12/23
12/24
12/25
12/26
12/27
12/28
12/29
日期（月/天）

图3

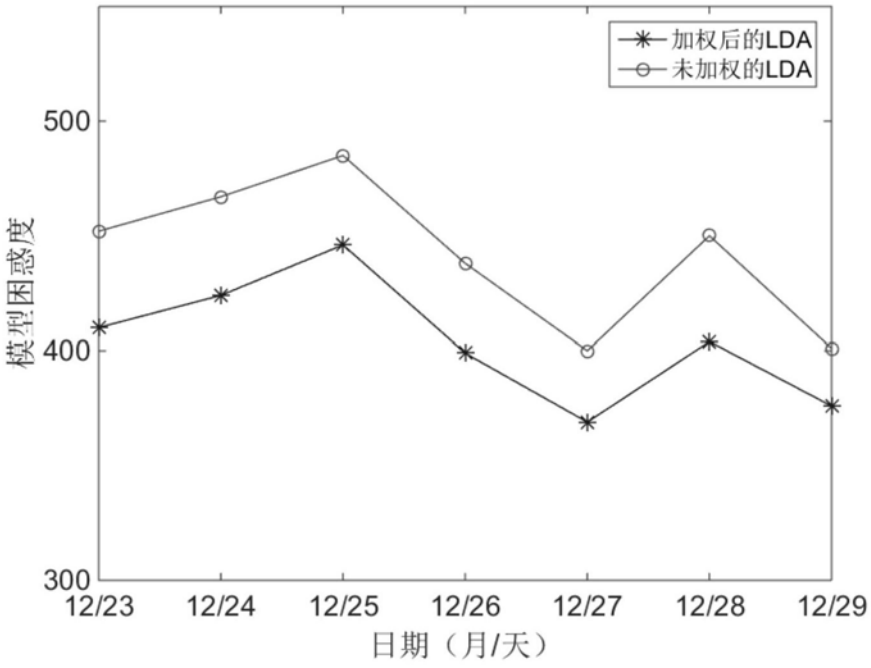

500
400
300
模型困惑度
加权后的LDA
未加权的LDA
12/23
12/24
12/25
12/26
12/27
12/28
12/29
日期（月/天）

图4

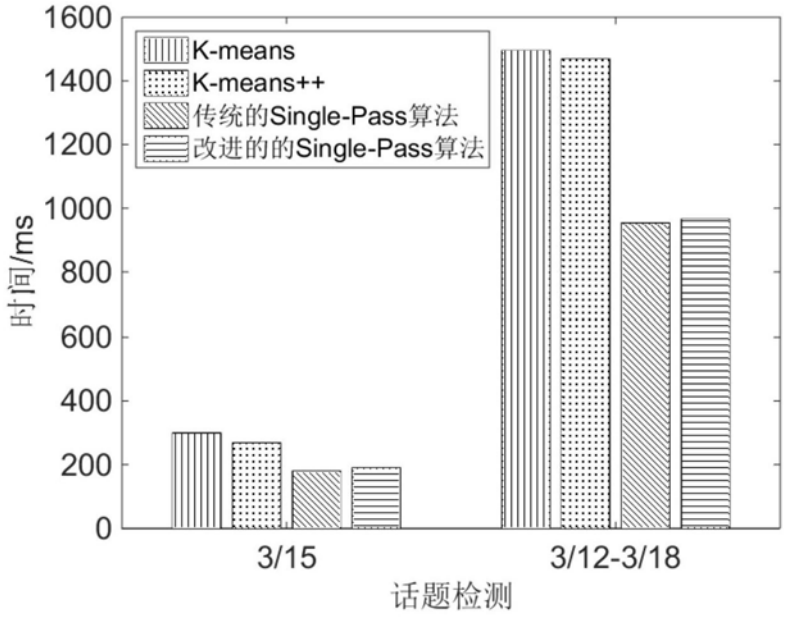

K-means
K-means++
传统的Single-Pass算法
改进的的Single-Pass算法
时间/ms
1600
1400
1200
1000
800
600
400
200
0
3/15
3/12-3/18
话题检测

图5

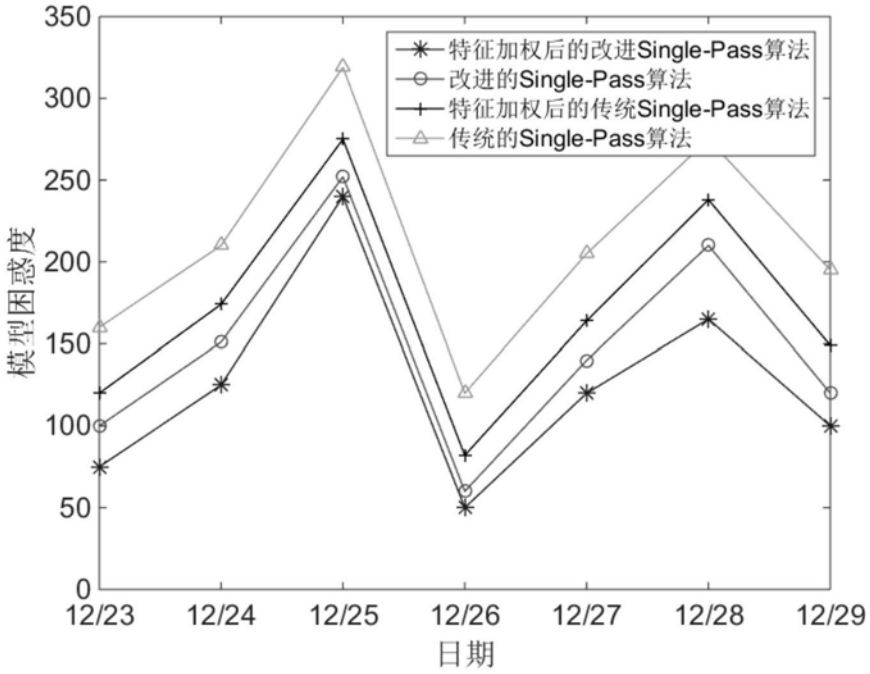

特征加权后的改进Single-Pass算法
改进的Single-Pass算法
特征加权后的传统Single-Pass算法
传统的Single-Pass算法
模型困惑度
350
300
250
200
150
100
50
0
12/23
12/24
12/25
12/26
12/27
12/28
12/29
日期

图6

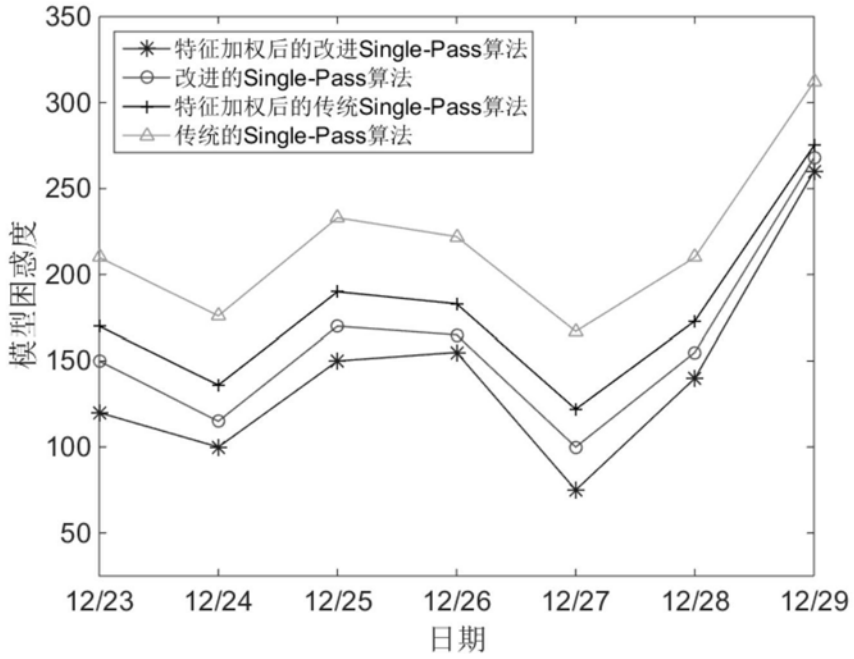

特征加权后的改进Single-Pass算法
改进的Single-Pass算法
特征加权后的传统Single-Pass算法
传统的Single-Pass算法
模型困惑度
日期
350
300
250
200
150
100
50
12/23
12/24
12/25
12/26
12/27
12/28
12/29

图7