E-Commerce Letters 电子商务评论, 2019, 8(1), 40-51
Published Online February 2019 in Hans. http://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2019.81006

Hans汉斯

Modeling and Optimization of Hot Topic Discovery in Social Media Based on Clustering of Word Kinetic Energy

Yingliang Wu[1,2], Kaimei Huang[1]

[1]Department of Electronic Business, School of Economics and Commerce, South China University of Technology, Guangzhou Guangdong
[2]Business Intelligence Research Center, Institute of Modern Services, South China University of Technology, Guangzhou Guangdong
Email: adylwu@scut.edu.cn, 849602425@qq.com

Received: Jan. 7th, 2018; accepted: Jan. 21st, 2019; published: Jan. 28th, 2019

Abstract

In the context of social media, social media public opinion has become a new perspective of social public opinion. Social big data analytics is showing an increasingly important social and business value. In the emerging field of social big data management, the discovery of hot topics in microblog is the basic and important issues for government public opinion analysis and data management; people have been studying and exploring advanced and applicable theories and methods of hot topic mining. However, when the traditional clustering algorithm is used in the microblog topic detection, the eigenvectors are too sparse and over-dimensioned, resulting in inaccurate clustering results. Therefore, this article explores the sudden features of words in the cycle of topic communication, and proposes the Word Kinetic Energy Clustering (WKEC) model and algorithm. The text clustering model, based on the topic life cycle feature, introduces the concept of kinetic energy theorem in physics, and calculates the kinetic energy of words with the maximum growth rate in the explosion period, which will be added to the weight of the words, modifying the classic TF-IDF model. Based on the algorithm design of Single-Pass and the real dataset from Sina Microblogs, the experimental results show that WKEC model can enhance the text features and improve the accuracy of topic discovery. In addition, due to the strong real-time of microblogging topic, in order to get closer to the real microblogging hot topic list, this article introduces the attenuation coefficient into the calculation of topic heat, and takes the tail time point of the explosion period as the decay moment of topic heat, proposing a more realistic method for the calculation of topic heat.

Keywords

Big Data Analysis, Topic Discovery, TF-IDF, Single-Pass, Word Suddenness, Word Kinetic Energy (WKE), Sina Microblogs

文章引用: 吴应良, 黄开梅. 基于词语动能聚类的社会化媒体热点话题发现建模与优化方法[J]. 电子商务评论, 2019, 8(1): 40-51. DOI: 10.12677/ecl.2019.81006

基于词语动能聚类的社会化媒体热点话题发现建模与优化方法

吴应良[1,2]，黄开梅[1]

[1]华南理工大学经济与贸易学院电子商务系，广东 广州
[2]华南理工大学现代服务业研究院商务智能研究中心，广东 广州
Email: adylwu@scut.edu.cn, 849602425@qq.com

收稿日期：2019年1月7日；录用日期：2019年1月21日；发布日期：2019年1月28日

摘　要

在社会化媒体情境下，社会化媒体舆情已成为社会舆情的新视域，社会化大数据分析正显现出日益重要的社会价值和商业价值。在新兴的社会化大数据管理领域，热点话题发现是网络舆情分析和数据治理基础而重要的课题，人们一直在研究和探索先进和适用的热点主题挖掘的理论和方法。针对传统的聚类算法用于微博话题检测时，存在特征向量过于稀疏和维度过高等问题，导致聚类结果不准确。本文通过对在话题传播周期中词语的突发性特征的研究，提出了一种基于传播周期的词语动能聚类(Word Kinetic Energy Clustering, WKEC)模型和算法。该文本聚类模型基于话题生命周期特性，引入物理学中的动能概念，用词语在话题爆发期的最大增长速度来表征词语的动能，并加入到词语权重的计算中，对经典的TF-IDF模型进行了改造。基于Single-Pass的算法设计和新浪微博真实数据集的实验结果表明，WKEC模型可以增强文本特征，提高话题发现的准确率。另外，由于微博话题实时性强，为了得到更接近真实的微博热点话题列表，本文在话题热度计算中引入衰减系数，并以爆发期尾部时间点作为话题热度开始衰减的时刻，给出了一种更加符合实际的话题热度计算方法。

关键词

大数据分析，话题发现，TF-IDF，Single-Pass，词语突发性，词语动能，新浪微博

1. 引言

近年来，以移动互联网与 Web2.0 为基础的社会化媒体工具和平台快速发展与广泛渗透，不仅成为最重要的大数据来源之一，更是形成了一种新的、重要的社会化媒体信息情境，深刻地改变了传统电子商务、电子政务的发展面貌与业务模式[1]。大数据和社会网络应用的发展使得当今的网络环境成为了一个巨大的、精准映射并持续记录人类行为特征的数字世界[2]。人们广泛关注社会化媒体的热门话题，并通过用户生成内容(UGC)、电子口碑(Electronic Word-of-Mouth, eWOM) [3]等来表达和传递自己的观点和思想，作为最重要的社会化媒体(社交媒体)之一，微博已融入经济活动与社会生活的方方面面，其内容涉及经济、政治、军事、文化、生活、体育与娱乐等各个领域，正显现出日益重要的商业价值、科学价值与

社会价值。在微博等社会化媒体工具或平台已经发展成为人们发表言论、传递公众情绪的重要渠道的时代背景下，对大数据时代政府的舆情管控与数据治理、企业的电子商务管理能力带来了新的挑战[4]。

回顾以往所发生的热点事件，几乎都能和微博相关联，微博等社会化媒体舆情已成为社会舆情的新视域，其社会影响力随着微博用户数量的增加而加强[5]。因此，微博热点话题发现是进行舆情分析与监管的基础和前提，对热点话题进行有效挖掘，可以对舆论尽早地进行疏导和管控，有助于维护社会稳定。另外，热点话题发现还具有重要的商业价值，企业可以通过微博等社会化媒体分析，随时和顾客保持有效沟通，深入了解顾客需求，以及发现用户兴趣和行为扩散的转变等，从而为企业在大数据环境下的营销决策提供有力支持，进而为企业带来商业价值。

因此，以微博为代表的社会化媒体热点话题的有效挖掘与发现，成为大数据分析和管理的一个热点课题，但一直面临特征向量过于稀疏、维度过高，导致聚类结果不准确等问题[6]。由于微博文本字数限制在 140 字以内，单条微博中每个单词出现的次数都很少，直接导致文本的特征不明显，特征向量稀疏，按照文本特征进行聚类困难。当前对于特征向量稀疏问题的研究主要集中在静态特征提取上，但是这些算法对于微博文本中的动态特征(词语突发性)无法进行有效测算，因而无法充分挖掘利用文本中所蕴含的特征信息。而在那些关注词语突发性的少数研究中，在词语突发性的衡量与评估上并没有结合传播生命周期特性，而传播生命周期正是微博动态特征和时间属性的最好表征。从当前来看，对于传播生命周期内词语突发性的衡量尚缺乏相应的研究理论和方法。并且，现有研究对于话题热度(人气度)计算多停留在微博用户特征和微博文本内容特征上，对于话题热度的时效性考量尚缺乏有效的评估技术，无法满足微博实时性强的特点。因此，在传播生命周期内针对微博短文本进行词语突发性衡量与话题人气度计算对提高聚类准确率以及提升政府和企业的决策能力具有重要意义。

2. 研究现状

随着网络信息爆炸式增长和满足网络舆情分析需求，基于微博平台的热点话题研究成为不同领域学者关注研究的热门领域。

唐晓波等[7]针对文本聚类和 LDA 主题模型的互补特征，综合考虑了微博特殊文体和短文本聚类效率问题，提出了基于频繁词集的文本聚类和基于类簇的 LDA 主题挖掘相融合的微博主题检索模型。徐雅斌等[8]提出并使用“频繁词集聚类”(Frequent Words Set Clustering, FWSC)算法来进行微博新话题的抽取，其有效解决了微博文本特征不明显问题。梁晓贺等[9]站在一个全新的视度，把超网络思想引进微博舆情主题发现中，构建了包含用户－观点－情感－时序阶段 4 层子网的舆情主题发现超网络模型。黄发良等[10]提出了一个基于多特征融合的微博主题情感挖掘模型 TSMMF (Topic Sentiment Model based on Multi-feature Fusion)，该模型将情感表情符号与微博用户性格情绪特征纳入到图模型 LDA 中，以实现微博主题与情感的同步推导。Yan 等[11]为了有效地检测微博文本中的热门话题，提出了一种基于潜在语义分析和结构特征相结合的微博话题检测方法，较大提高了微博话题检测模型和方法的性能。Liang 等[12]提出情感分布语言模型(Emotion Distribution Language Models, ELM)来模拟微博中的情感分布，并用它来检测热点发现的潜在时间间隔，根据检测到的时间间隔对话题进行分析，进而根据话题的内容和转发度估算每条微博的重要性，通过应用话题模型提取热点话题，将这两步法应用到新浪微博，实验结果表明该方法是有效的。Li 等[13]采用 BTM (Bi-term Topic Model)主题模型处理新浪微博短文本数据，以缓解稀疏问题，同时，将 K-means 聚类算法融入到 BTM 中，进一步进行主题发现，实验结果表明，该方法可以有效地发现话题。Zhao 等[14]基于新浪微博应用场景的研究，提出了一种“社交情感感知系统”(Social Emotional Perception System, SEPS)方法，进行每日热点话题检测，并分析这些话题的情感分布，可生成一个实时可视化的系统来监测社会情绪。袁华等[15]提出一种新的数据挖掘方法用于从海量 UGC 中分

析出其“热点话题词”和“局部特征词”之间的关联关系，首先从互联网上抓取大量文档内容，并进行分词抽取出包含语义的词作为最初数据集，然后利用域内热点名词词表，抓取出数据集中的热点话题词及其序列，接着利用一种新的词向量切分方法对数据集进行基于热点话题词的切分，并且通过最大置信度这一指标挖掘特定热点话题词下的局部特征词。牛奉高等[16]以优化 CLSVSM 为出发点，提出了一种新的文本表示语义核模型(Co-occurrence Latent Semantic Vector Space Model Kernel, CLSVSM_K)，具体针对 CLSVSM 维度较高、计算复杂度较大等诸多问题，构建了 CLSVSM 的语义核，其构建原理是基于潜在语义分析的思想，对特征词中的同义词进行了合并的同时又对共现矩阵进行了降维处理，大大降低了算法的复杂度，在数字文献资源上的实验结果表明该方法还具有良好的聚类效果，提高了文献资源主题聚合的精度。

在词语突发性概念与规律的研究中，薛薇[17]认为在一个随时间推进的信息流中，若某信息的出现密度突增则认为它具有突发特征。仲兆满等[18]根据微博的时空特点，在综合考虑微博博文及社交关系的基础上，利用词出现频率、词关联用户、词分布地域和词社交行为四类指标，提出了一种新颖的微博网络词突发值计算模型。郑斐然等[19]综合考虑词频以及词频在时间窗口的增长率来产生词语的权重，并通过增量聚类的方法从微博中挖掘新闻话题。林思娟等[20]提出一种基于词语能量值变化的微博热点话题检测方法，该方法引入物理学科中加速度的概念，用词语的加速度来刻画词语在相邻窗口之间速度的变化，并综合考虑词语的加速度和权重值来构造词语的综合能量值。金镇晟[21]结合物体的动能概念，用词语的突发值表示词语在某一时刻所具有的动能，将词语的动能加入到特征权重计算中，提高突发性词语的权重，提出了一种改进的特征提取算法，有利于更好地完成文本聚类。

关于话题的动态变化与热度计算，有学者开展了一些重要研究。Wu [22]对在线论坛话题人气度的建模与仿真研究表明，话题人气度(热度)的演化呈现出按一定的长尾幂律(Long-Tail Power Law)进行衰减的重要特性。何跃等[23]从微博用户特征、微博文本内容特征和微博信息传播特征三个维度出发，构建了一种评价微博热度的指标体系。为了得到更符合实际的话题热度计算方法，裴可锋等[24]引入 LDA 模型，结合话题内容和外在特征两个方面的热度因素，得到话题热度变化的时间序列。杨冠超[25]在微博影响力的计算中，对时间因素引起的影响力降低进行了考察和衡量，引入了影响力衰减系数。

西安交通大学的薛峰、周亚东等人[26]提出了综合的改进思想-在特征选择阶段动态地生成热点词特征库，在文本表示时给予突发性热点词更大的权重，然后进行聚类操作，经过实验证明该方案十分理想。本研究以薛峰、周亚东等人的综合改进思想为指导，从传播生命周期的角度，结合动能定理对词语的突发性进行衡量与评估，并将词语动能加入到 TF-IDF 权值计算中，提出一种改进的特征提取算法，提高突发性词语的权重，以便更好地完成文本聚类，并在话题热度(人气度)计算中引入衰减系数以对话题的时效性进行考量，研究结果对于辅助和指导企业的营销决策以及政府的舆情管控工作具有重要的参考意义。

3. 新的热点话题发现方法与建模的理论基础

3.1. 考虑时变因素的话题热度(人气度)的计算

考虑到微博热点话题的更新交替实时性很强，用户对于一个话题的关注会由于其他的因素如兴趣转移等发生自然的流逝，并且用户 u 的粉丝数和所发布微博的转发数和评论数是相对静态的，计算出来的热度值不能代表当前的热度值。所以在计算话题热度时要考虑时间因素。杨冠超[25]在微博话题影响力计算中将时间因素引起的影响力降低进行了考虑和衡量，引入了影响力衰减系数 d，则微博话题影响力随时间 t 的变化如式(1)所示。

$$\begin{aligned} Inf_{b,u}(t) &= -d(t-t_0)^2 + \left|u_{\text{followed}}\right| \\ H_T &= \sum \Delta Inf_T = \sum\left(\sum_{i=1}^{n} Inf_{b_i}\right) \end{aligned} \tag{1}$$

其中，$Inf_{b,u}$ 为用户 u 所发布的微博 b 的直接影响力；d 是衰减系数，根据经验取值；t_0 表示微博发布的时间；$\left|u_{\text{followed}}\right|$ 表示当前关注用户 u 的人数；ΔInf_T 表示给定一个时间段 Δt，话题 T 的影响力变化；H_T 表示话题 T 从开始发布的时间到当前时间的影响力变化。

3.2. 词语动能

“词语动能”(Word Kinetic Energy, WKE)的定义：词语因为运动而具有的能量，称为“词语动能”。在微博话题传播的情境中，“词语运动”是指词语随着文本数量增长而增长的这样一个动态过程，根据词语的增长速度计算而得到的一种值，即“词语动能”。词语的动能，是网络词语的突发性大小的一种度量，用词语的平均增长速度来进行计算[21]。词语动能的物理意义表示词语所具有的能量，是一种对词语突发性的衡量；而词语动能的社会意义，则表现为话题传播影响力的强度。

微博热点话题的传播有其特有的生命周期，其传播过程一般会经历潜伏期、爆发期、发展期、消亡期等阶段[27]。对于微博话题传播来说，随着时间的推进，相关文本数量会不断增长，即使文本形式和内容会出现一定的演变，但该话题的关键性词语仍然会出现在相关微博中。也就是说话题的关键词语和该话题文本数量具有相一致的增长特性。

词语的突发性大小与其在传播周期中的增长速度有关，并且其增长速度越大，其所呈现出的突发性也就越明显，其对于话题表述就越起着重要的特征描述作用，区分文本的能力也越强。而这和物理学中的动能定理较为相像，即一个物体因运动而具有能量，速度越大，它的动能也就越大。动能定理的基本公式如式(2)所示。

$$E=(1/2)mv^2 \tag{2}$$

其中，m 表示物体的质量，v 表示物体所具有的速度。

在这种理论的启发下，本文引入词语动能的概念，通过研究词语在不同时间窗口中的动能，来对该词语的突发性进行衡量：词语所具有的速度越大，其动能也就越大，词语的突发性也就越明显。本文用词语在不同时间窗口之间词频的变化情况来表示词语的速度 $v_{k,j}$，如式(3)、式(4)所示。

$$v_{k,j}=\begin{cases}0, k=0\\ \dfrac{m_{k,j}-m_{k-1,j}}{T_{k-1,k}}, k\neq 0\end{cases} \tag{3}$$

$$T_{k-1,k}=T_{k\_middle}-T_{k-1\_middle} \tag{4}$$

其中，$v_{k,j}$ 表示词语 j 在第 k 个时间窗口的速度；$m_{k,j}$ 表示词语 j 在第 k 个时间窗口的词频；$m_{k-1,j}$ 表示词语 j 在第 $k-1$ 个时间窗口的词频；T_{k_middle} 表示第 k 个时间窗口的中间时刻点；T_{k-1_middle} 表示第 $k-1$ 个时间窗口的中间时刻点；$T_{k-1,k}$ 表示第 k 个时间窗口与第 $k-1$ 个时间窗口的时间间隔。

物体因运动而具有能量，词语也因运动而具有能量。当词语具有了一定的速度时，也就具有了一定的动能。词语动能的计算由式(5)表示。

$$E_{k,j}=\alpha\cdot tf_{kj}\cdot v_{k,j}^2 \tag{5}$$

其中，$E_{k,j}$ 表示词语 j 在第 k 个窗口的动能；α 是动能权重系数，这是一个经验值，需要通过实验来进行确定；$tf_{k,j}$ 表示词语 j 在第 k 个窗口中出现的频率，它等于词语 j 在第 k 个窗口的词频除以第 k 个窗口的词频总数得到；$v_{k,j}$ 是词语 j 的速度。

4. 问题描述与模型构建

4.1. 问题定义

1) 研究问题一：如何基于传播周期对词语突发性进行衡量与评估？

在一个随时间推进的信息流中，若某词语的出现密度突增则认为它具有突发特征，那在话题传播所形成的话题相关文本信息流中，话题关键词与话题文本具有相一致的增长趋势，从词语突发性角度来看，不同词语有着不一样的突发性表征，话题关键词比普通词语突发性表征更明显，突发性表征越明显的词语，越有可能是话题关键词。而同一个词语在整个传播周期中，其突发性表征也会呈现出一定的周期性，在潜伏期和发展期相对比较平缓，在衰亡期呈现负的突发性，而相对于整个传播周期来说，词语在爆发期的突发性特征是比较明显的，并且词语突发性大小经历了先增大后减小的过程，所以，词语突发性最明显的地方出现在爆发期的偏中间部分，这个时候的词语突发性在整个传播周期中最具代表性。在词语突发性的衡量与评估方面，本文结合动能定理，采用词语在爆发期的最大增长速度来计算词语的动能，以表征词语的突发性，并将其加入到 TF-IDF 权重中。本文希望通过识别突发性词语来找到话题关键词，并加大其权重，以达到增强文本特征的目的，从而提高话题聚类准确率。

2) 研究问题二：如何基于传播周期对话题热度(人气度)进行时效性考量？

微博信息大多为实时性信息，热点话题的参与人数会随着讨论的深入逐渐增加，而到达讨论的高峰之后由于用户注意力的转移等因素，会逐渐进入一个平台期，其影响力保持不变或处于波动状态。之后，与新闻信息相同，随着时间的流逝，前一段时间内的热点话题的讨论热度将不断降低，而用户在平台上的活动又使得一些新的话题成为热点话题。把某一话题下的相关微博条目视为一个时间轴上的文本流，以上变化过程就表现为话题的传播周期特性。而热点话题发现建模要求我们判断话题是否为当前热点，而不是已经过时的热点，而话题本身所特有的传播周期特性使得判断话题是否为当前热点这一命题可以通过数学计算计算出来。

4.2. 基于词语动能的文本聚类模型

1) 对经典特征词权重计算方法 TF-IDF 的改进

由于词语在潜伏期和消亡期的增长速度很小甚至为 0 为负，其会拉低词语在整个传播周期内的平均增长速度，所以，根据平均增长速度计算得到的动能并不能很好地对词语突发性进行衡量。根据词语的传播周期曲线可知，词语在爆发期的增长速度，相对于整个传播周期来说，数值是比较大的，并且速度是先增大后减小，其最大增长速度出现在爆发期的偏中间部分，这也是词语突发性最明显的时候。因此，采用词语在爆发期的最大增长速度来计算其动能，并加入到其权重中，可以更有效的增强文本特征，其公式如式(6)所示。

$$\omega_{ij}=\frac{tf_{ij}\cdot\log\left(N/n_j+0.01\right)}{\sqrt{\sum_{p=1}^{k}\left[tf_{ij}\cdot\log\left(N/n_j+0.01\right)\right]^2}}+E_{k,j} \tag{6}$$

其中，公式(6)的前半部分是规范化的 TF-IDF，ω_{ij} 表示在文本 i 中特征词 j 的权重，tf_{ij} 是特征词 j 在文本 i 中出现的频率，N 是数据集中的文本总数，n_j 是数据集中包含特征词 j 的文本数，$E_{k,j}$ 表示特征词 j 的动能，采用特征词 j 在爆发期的最大增长速度计算得到。

2) 引入词语动能的文本相似度的计算

文本聚类的基础是文本间相似度的计算。本文选取应用最为广泛的余弦相似度公式来计算文本间的相似度值，其值一般在 0 到 1 之间，余弦法则公式如式(7)所示。

$$\cos\left(x,y\right)=\frac{\sum_{k=1}^{n}\omega_{xk}\times\omega_{yk}}{\sqrt{\sum_{k=1}^{n}\omega_{xk}^2\times\sum_{k=1}^{n}\omega_{yk}^2}} \tag{7}$$

其中，x 和 y 是进行相似度值计算的两个文本向量，ω_{xk} 表示向量 x 的第 k 个特征词的权重，ω_{yk} 表示向量 y 的第 k 个特征词的权重，n 表示 x 和 y 的特征词数量。

4.3. 优化算法设计

本文话题发现的核心算法是 Single-Pass 聚类算法[6]。Single-Pass 算法是单遍聚类算法，其对输入语料的顺序很敏感，具体表现为：该算法选择聚类中心时一般直接选取输入文档集的第一个文档成为第一个类别的质心，余下文本与该文档运算相似度，进而进行聚类。如果这篇文档不太具有代表意义，就不能全面的对一个类中的话题含义进行很好的阐述，那么最后的聚类效果不会让人满意。

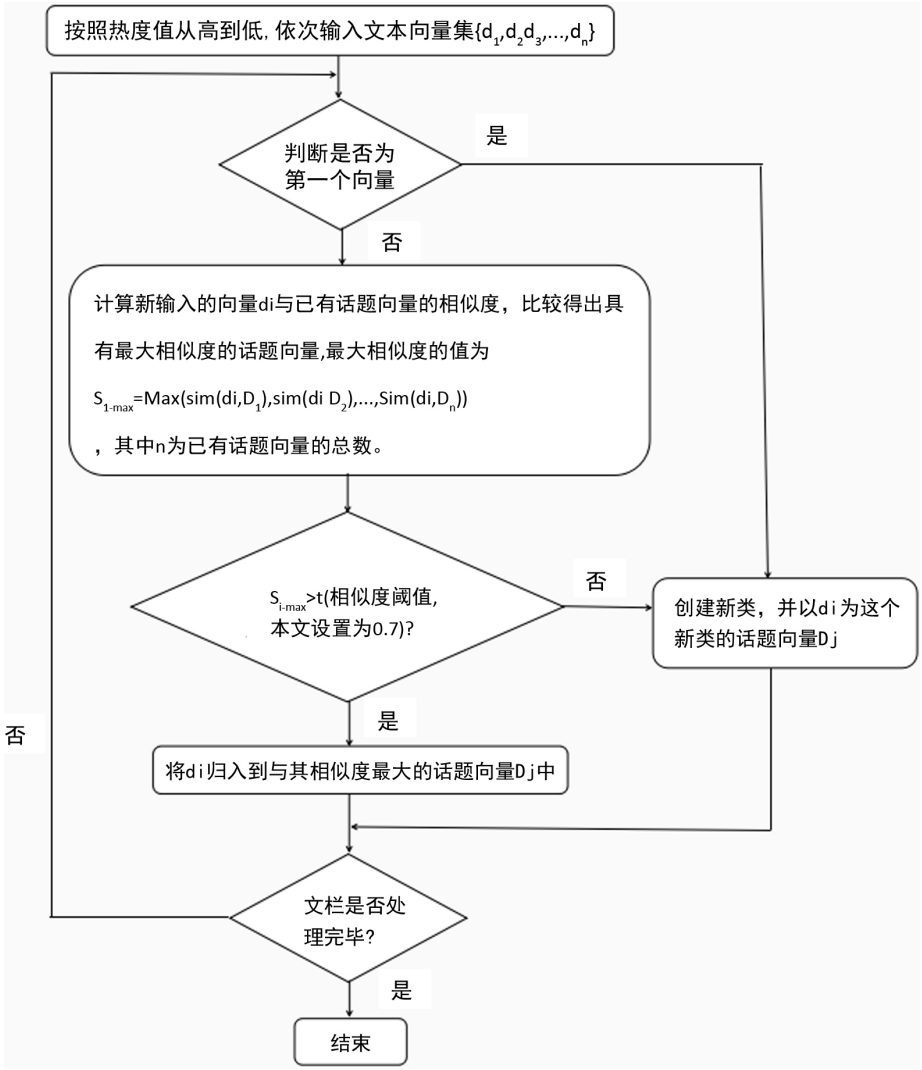

Figure 1. Flow chart of Single-Pass clustering algorithm
图 1. Single-Pass 聚类算法流程图

为解决这个问题，本文采用文献[28] (孙胜平，2011)提出的基于微博的转发数、评论数和用户粉丝数的微博热度计算公式来计算微博语料的热度。并把微博语料按照热度从高到低排序，依次输入到 Single-Pass 聚类算法中，以便更好地适应该算法对语料输入顺序需求[27]。Single-Pass 算法会顺序地处理输入的文本，每次处理一篇，增量地更新聚类结果。算法具体步骤如图 1 所示。

聚类算法完成后，会形成若干独立的话题向量，每个话题向量由一组特征词及其权重值构成。把每个话题向量中的特征词按照权值从大到小进行排序，即可辨识出话题的主要内容。

4.4. 基于传播周期的话题热度(人气度)计算

根据话题传播周期可知，话题在爆发期和发展期是处于热度持续上升的阶段，但为了计算方便，本文假定发展期话题热度是衰减的，所以以话题爆发期尾部时间点为话题热度开始衰减的时刻，引入热度衰减系数 β，提出更符合实际的话题热度计算方法，如式(8)所示。

$$TH_i = \sum_{k=0}^{n} bh_{ik} - \beta\left(t - t_0\right)^2 = \sum_{k=0}^{n}\left(2re_k + cm_k + \sqrt{fl_k}\right) - \beta\left(t - t_o\right)^2 \tag{8}$$

其中，TH_i 表示第 i 个话题的热度；bh_{ik} 表示第 i 个话题内的第 k 条微博的热度值，其计算基于孙胜平提出的微博博文热度计算方法[28]，re_k 代表第 k 条微博的转发数，其权重系数为 2，cm_k 代表第 k 条微博的评论数，其权重系数为 1，fl_k 代表第 k 条微博的博主粉丝数，其权重系数为 1/2；β 是热度衰减系数，根据经验取值；t 表示计算话题热度时的时间；t_0 表示话题 i 的爆发期尾部时间点。

公式(8)结合了话题在传播周期中的发展趋势，把时间因素所引起的影响力降低进行了更好的考虑和衡量，所计算出来的话题热度值更加接近现实，代表话题当前的热度，符合微博话题实时性强的特点。

5. 实验及结果分析

为了检验模型改进的效果，实验总体设计分为以下 4 个部分：1) 实验语料准备及模型性能评测标准；2) 动能权重系数 α 最佳取值的探讨；3) 为了验证本文提出的话题热度计算方法的有效性，将其与文献[25]的算法同同一时刻新浪推荐热门话题列表进行对比；4) 为了验证词语动能聚类算法在微博话题检测中的改进效果，将其与文献[21]提出的采用平均增长速度来计算词语的动能并加入到特征词权重中的改进算法进行对比实验。

5.1. 实验语料准备

图 2 是获取本文实验语料的流程图，其中，去除停用词所采用的停用词表是根据《百度停用词列表》、《哈工大停用词表》和《四川大学机器智能实验室停用词库》这三个比较权威的停用词库整理得到的。

Figure 2. Flow chart for acquiring experimental corpus
图 2. 获取实验语料的流程图

5.2. 模型和算法性能评测

TDT 会议对话题检测制定了一种规范的评价标准，评测指标主要有：召回率、准确率、漏检率、错检率、F 值以及误测开销[29]。该评价标准同样适用于微博平台的话题检测。本文选用话题检测常用的评价指标即召回率、错检率来对模型与算法的性能进行测试。

召回率(R)是系统正确检测出属于某一个话题的文本数(D)和所有应该被检测出的属于某一个话题的文本数(T)的比值。

$$R = D/T * 100\% \tag{9}$$

错检率(P_{FA})是系统判断错误的也就是错误检测出的属于某一个话题的文本数(FA)和所有不属于该话题的文本数(NT)的比值。

$$P_{FA} = FA/NT * 100\% \tag{10}$$

本文的召回率 R 值和错检率 P_{FA} 值都是根据计算出来的每一个话题的 R 值与 P_{FA} 值，然后求平均得到。

5.3. 实验结果

5.3.1. 动能权重系数 α 的取值

在进行真正的实验之前，先对动能权重系数 α 的取值进行探讨，α 是一个经验值，需要通过实验来确定它的最佳取值。在其他实验条件相同的情况下，α 取值不同，实验效果也不同，其召回率和错检率随 α 变化的结果如图 3、图 4 所示。

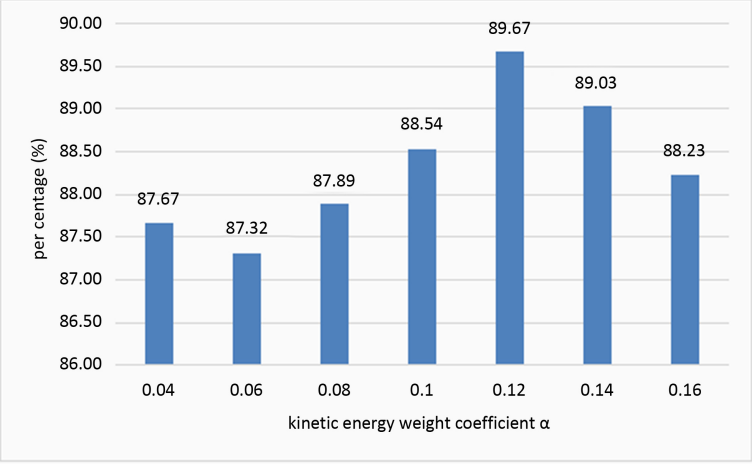

Figure 3. Recall rate changing with α
图 3. 随 α 变化的召回率

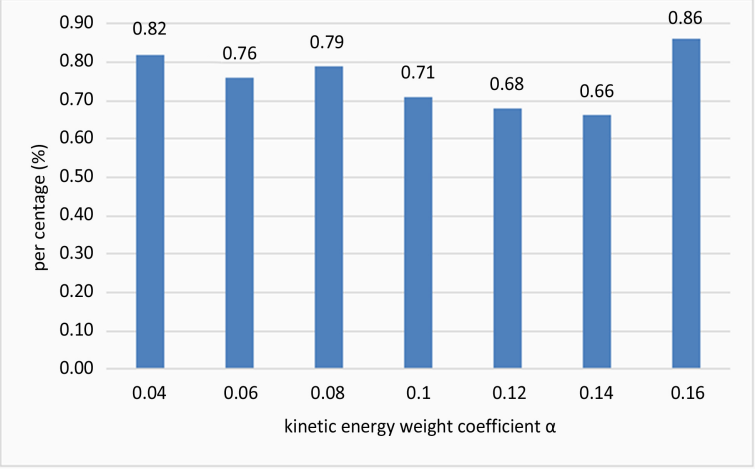

Figure 4. Error detection rate changing with α
图 4. 随 α 变化的错检率

从图 3 和图 4 中可以看出，当 α 取值为 0.12 时，实验效果较好，召回率较高，错检率也较低。所以本文的动能权重系数 α 设置为 0.12，进行话题发现实验。

5.3.2. 实验结果及算法评测

本文一共检测出 39 个话题，为了简单清晰地展现实验结果，只对热度排名前 4 的话题进行实验结果展示、对比和算法评测。表 1 是热度排名前 4 的话题向量：

Table 1. Experimental results of topic discovery algorithm
表 1. 话题发现算法的实验结果

话题 1		话题 2		话题 3		话题 4	
特征词	权重	特征词	权重	特征词	权重	特征词	权重
杨洁	0.096	自杀	0.086	幸存	0.097	苦撑	0.083
去世	0.087	袭击者	0.082	小女孩	0.089	决不放弃	0.075
导演	0.084	叙利亚	0.079	天真一笑	0.082	一步之遥	0.072
西游记	0.082	引诱	0.075	爆炸	0.075	拔河比赛	0.067
回忆	0.076	薯片	0.072	袭击	0.072	只剩	0.063
快乐	0.071	儿童	0.070	回眸	0.065	一人	0.061
一路走好	0.069	引爆	0.064	挥舞	0.050	出乎意料	0.058

为了检验本文提出的话题热度计算方法的有效性，把文献[25]提出的方法记为算法 1，然后把本文算法和算法 1 分别与同一时刻新浪推荐热门话题列表进行对比。对比结果如表 2 所示：

Table 2. Comparison of experimental results
表 2. 实验结果对比

话题排名	本文算法实验结果	算法 1	新浪推荐热门话题
1	《西游记》导演杨洁去世	自杀袭击者引诱儿童	《西游记》导演杨洁去世
2	叙利亚自杀袭击者用薯片引诱儿童	独生女留学嫁老外不归	天舟一号在海南文昌发射
3	一个在爆炸袭击中幸存的叙利亚小女孩的微笑	白百何出轨事件	拔河比赛只剩一人苦撑
4	小学拔河比赛只剩他一人苦撑，决不放弃	天舟一号发射	自杀袭击者用薯片引诱儿童

由表 2 可以看出，本文算法检测出来的前 4 个热点话题在新浪推荐的热门话题列表中有 3 个，并且检测出来的第一个热点话题就是新浪推荐中的第一个，检测结果的准确性较高。而算法 1 所得到的热点话题列表，只有 2 个话题在新浪推荐热门话题列表中，并且“白百何出轨事件”发生在 2017 年 4 月 12 日，时至 2017 年 4 月 16 日，该事件已经基本平息，而本文做实验的时间点是 2017 年 4 月 19 日，说明该话题虽曾经很热门，但现已经不再是人们关注的焦点。所以本文提出的话题热度计算方法更有效，所计算出来的话题热度值也更接近真实情况。

文献[21]用突发值这一概念来衡量和表示微博的时间属性，结合动能定理，采用平均增长速度，将词语的突发值表示成动能并加入到特征词权重计算中。本文基于微博意见领袖进行话题发现，并对词语动能的计算进行了改进，在词语动能计算中考虑了话题的传播周期特性，速度采用词语增长速度的最大值而不是平均增长速度。本文在基本同一实验环境下对以上两种算法同时进行实验，目的是验证本文算法是否有显著的改进效果。把文献[21]提出的算法记为算法 2，比较两个算法的召回率见表 3、错检率见表 4。

Table 3. Topic recall rate obtained by two algorithms
表 3. 两种算法得到的话题召回率

序号	话题	算法 2(%)	本文算法(%)
1	《西游记》导演杨洁去世	80.82	86.35
2	叙利亚自杀袭击者用薯片引诱儿童	81.98	85.21
3	一个在爆炸袭击中幸存的叙利亚小女孩的微笑	80.21	87.31
4	小学拔河比赛只剩他一人苦撑，决不放弃	81.24	85.24

Table 4. Topic error detection rate obtained by two algorithms
表 4. 两种算法得到的话题错检率

序号	话题	算法 2(%)	本文算法(%)
1	《西游记》导演杨洁去世	0.81	0.67
2	叙利亚自杀袭击者用薯片引诱儿童	0.79	0.65
3	一个在爆炸袭击中幸存的叙利亚小女孩的微笑	0.82	0.72
4	小学拔河比赛只剩他一人苦撑，决不放弃	0.83	0.61

通过对比实验表明，本文算法表现出较高的召回率和较低的错检率，这说明本文提出的结合话题传播周期特性，用词语在爆发期的增长速度的最大值来计算词语的动能，并加入到词语的权重中，增强了微博文本特征，从而提高了话题检测的准确率。

6. 结论

社会计算与大数据时代的到来[30]，对社会化媒体分析和大数据治理提出了许多具有重要价值和时代挑战性的课题[31]。利用计算机技术以及社会化媒体平台上即时产生的海量数据来精准发现热点话题，从而制定有针对性的传播策略，对于任何社会组织而言都非常重要[32]。为了解决微博等社会化媒体文本短小所带来的文本特征不明显问题，提高话题聚类的准确率，进而为有效进行舆情分析与监管提供更好的决策支持以及为企业在大数据环境下的营销决策提供有力支持，本文针对现有典型话题检测模型与算法存在的不足，基于对话题传播周期特性的认知，对词语的突发性特征进行了深入的探讨与刻画，以此为基础，提出了一种新的用于微博热点话题发现的词语动能聚类(WKEC)模型和算法。为了证明 WKEC 模型与算法的有效性，将本文算法和现有典型算法进行了对比，微博真实数据集上的对比实验结果表明，本文算法在召回率和错检率指标上都表现的更好，即召回率更高、错检率更低，说明本文提出的模型与算法聚类准确率更高。另外，本文针对微博话题实时性强的特点，提出了一种新的更符合实际的微博话题热度计算方法，把其和相关的已有典型模型和算法热点话题检测结果与同一时刻新浪热门话题推荐结果进行对比，比较结果表明，本文提出的模型与方法所计算出来的话题热度值更接近真实情况。本文研究通过对词语权重计算的改造与优化，对经典的文本聚类模型 TF-IDF 做出了改进，由此提出的热点话题发现算法，为进一步提高微博等社会化媒体分析方法的质量与效果提供了一种新的途径，具有重要的理论意义与应用价值。

当然本文所提出的 WKEC 模型，只是简单地把词语动能加入到 TF-IDF 权重中，可能存在不合理的地方，在以后的研究中可讨论如何将词语动能以更合理的方式来改进 TF-IDF 权重，以更为准确地发现热点话题。

基金项目

国家自然科学基金项目“管理科学理论和方法的综合集成研究”(70440011)，国家社会科学基金项目“分享经济下基于 TRIZ 理论的网络约租车服务创新研究”(16BGL190)，国家社会科学基金项目“基于关联数据的政府数据开放研究”(14BTQ009)。

参考文献

[1] Lee, I. (2017) Big Data: Dimensions, Evolution, Impacts, and Challenges. *Business Horizons*, **60**, 293-303. https://doi.org/10.1016/j.bushor.2017.01.004

[2] 冯芷艳, 郭迅华, 曾大军, 等. 大数据背景下商务管理研究若干前沿课题[J]. 管理科学学报, 2013, 16(1): 1-9.

[3] Ahmad, S.N. and Laroche, M. (2017) Analyzing Electronic Word-of-Mouth: A Social Commerce Constructs. *International Journal of Information Management*, **37**, 202-213. https://doi.org/10.1016/j.ijinfomgt.2016.08.004

[4] 郑大庆, 黄丽华, 张成洪, 等. 大数据治理的概念及其参考架构[J]. 研究与发展管理, 2017, 29(4): 65-72.

[5] 刘社瑞, 唐双. 自媒体时代微博舆情演化与应对策略[J]. 求索, 2011(10): 86-87, 171.

[6] 杜治娟. 社会媒体大数据分析研究综述[J]. 计算机科学与探索, 2017, 11(1): 1-23.

[7] 唐晓波, 房小可. 基于文本聚类与 LDA 相融合的微博主题检索模型研究[J]. 情报理论与实践, 2013, 36(8): 85-90.

[8] 徐雅斌, 李卓, 吕非非, 等. 基于频繁词集聚类的微博新话题快速发现[J]. 系统工程理论与实践, 2014, 34(S1): 276-282.

[9] 梁晓贺, 田儒雅, 吴蕾, 等. 基于超网络的微博舆情主题挖掘方法[J]. 情报理论与实践, 2017, 40(10): 100-105.

[10] 黄发良, 冯时, 王大玲, 等. 基于多特征融合的微博主题情感挖掘[J]. 计算机学报, 2017, 40(4): 872-888.

[11] Yan, X. and Zhao, H. (2013) Chinese Microblog Topic Detection Based on the Latent Semantic Analysis and Structural Property. *Journal of Networks*, **8**, 917-923. https://doi.org/10.4304/jnw.8.4.917-923

[12] Yang, L., Lin, H.F., Lin, Y., *et al.* (2016) Detection and Extraction of Hot Topics on Chinese Microblogs. *Cognitive Computation*, **8**, 577-586. https://doi.org/10.1007/s12559-015-9380-6

[13] Li, W.J., Feng, Y.M., Li, D.J., *et al.* (2016) Micro-Blog Topic Detection Method Based on BTM Topic Model and K-Means Clustering Algorithm. *Automatic Control and Computer Sciences*, **50**, 271-277. https://doi.org/10.3103/S0146411616040040

[14] Zhao, Y.Y., Qin, B., Liu, T., *et al.* (2016) Social sentiment Sensor: A Visualization System for Topic Detection and Topic Sentiment Analysis on Microblog. *Multimedia Tools and Applications*, **75**, 8843-8860. https://doi.org/10.1007/s11042-014-2184-y

[15] 袁华, 徐华林, 钱宇, 等. 域内海量数据中热点话题及其特征词抽取方法[J]. 管理工程学报, 2018, 32(4): 133-140.

[16] 牛奉高, 张亚宇. 基于共现潜在语义向量空间模型的语义核构建[J]. 情报学报, 2017, 36(8): 834-842.

[17] 薛薇. 基于突发性诊断的网络热点事件识别方法[J]. 统计与决策, 2015(15): 8-12.

[18] 仲兆满, 管燕, 李存华, 等. 微博网络地域 Top-K 突发事件检测[J]. 计算机学报, 2018, 41(7): 1504-1516.

[19] 郑斐然, 苗夺谦, 张志飞, 等. 一种中文微博新闻话题检测的方法[J]. 计算机科学, 2012, 39(1): 138-141.

[20] 林思娟, 林柏钢, 许为, 等. 一种基于词语能量值变化的微博热点话题发现方法研究[J]. 信息网络安全, 2015(10): 46-52.

[21] 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]: [硕士学位论文]. 北京: 北京理工大学, 2015: 26-29.

[22] Wu, Y. and Wu, W. (2015) Modeling Topic Popularity Distribution and Evolution in an Online Discussion Forum. *Journal of Computational Information Systems*, **18**, 6797-6810.

[23] 何跃, 蔡博驰. 基于因子分析法的微博热度评价模型[J]. 统计与决策, 2016(18): 52-54.

[24] 裴可锋, 陈永洲, 马静. 基于 DTPM 模型的话题热度预测方法[J]. 情报杂志, 2016, 35(12): 52-57.

[25] 杨冠超. 微博客热点话题发现策略研究[D]: [硕士学位论文]. 杭州: 浙江大学, 2011: 26-27.

[26] 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): 64-69 + 116.

[27] 姚海波. 微博热点话题检测与趋势预测研究[D]: [硕士学位论文]. 广州: 华南理工大学, 2013: 23-38.

[28] 孙胜平. 中文微博客热点话题检测与跟踪技术研究[D]: [硕士学位论文]. 北京: 北京交通大学, 2011: 38-40.

[29] 洪宇, 张宇, 刘挺, 等. 话题检测与跟踪的评测及研究综述[J]. 中文信息学报, 2007(6): 71-87.

[30] 孟小峰, 李勇, 祝建华. 社会计算: 大数据时代的机遇与挑战[J]. 计算机研究与发展, 2013(12): 2483-2491.

[31] 吴应良, 黄媛, 王选飞. 在线中文用户评论研究综述: 基于情感计算的视角[J]. 情报科学, 2017(6): 159-163, 170.

[32] 王玮, 温世阳. 情感分析在社会化媒体效果研究中的应用——基于分类序列规则的微博文本情绪分析[J]. 国际新闻界, 2017, 39(4): 63-75.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2168-5843，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：ecl@hanspub.org