文章编号：2096-1472(2021)-08-31-03
DOI:10.19644/j.cnki.issn2096-1472.2021.08.007

# 基于多个特征的多层次微博检索方法研究

范怡敏

(南昌理工学院计算机信息工程学院，江西 南昌 330044)
✉rowan521@163.com

**摘　要**：为了从大量微博信息中提取重要事件并预测发展趋势，基于微博的地理特征和时间特征，提出了一种对微博进行聚类和索引的多层次方法。该方法使用X均值聚类，根据用户输入的关键词建立索引，并根据索引自动评估聚类的数量。同时，基于情感特征对微博进行聚类，创建包含负面情感微博和正面情感微博的两个聚类。实验结果表明，所提索引机制不仅便于搜索，而且有利于检索任务。与其他微博聚类方法相比，所提方法在DBI指标和S系数两个指标上均有更好的表现，且时间复杂度较传统方法更低，与输入数据量的对数成正比。

**关键词**：微博检索；时间特征；地理特征；情感特征
**中图分类号**：TP391　**文献标识码**：A

# Research on Multi-level Microblog Retrieval Method based on Multiple Features

FAN Yimin

( *College of Computer Information and Engineering*, *Nanchang Institute of Technology*, *Nanchang* 330044, *China*)
✉rowan521@163.com

**Abstract:** In order to extract important events from a large amount of microblog information and predict the development trend, this paper proposes a multi-level method for clustering and indexing microblogs based on geographic and temporal characteristics of microblogs. X-mean clustering is used in this method, an index is built based on the keywords entered by the user, and the number of clusters is automatically evaluated based on the index. At the same time, the microblogs are clustered based on emotional characteristics, and two clusters containing negative emotional microblogs and positive emotional microblogs are created. Experimental results show that the proposed indexing mechanism is not only convenient for searching, but also conducive to retrieval tasks. Compared with other microblog clustering methods, the proposed method has better performance on both the DBI (Discriminated Bond Index) indicator and the S coefficient. The time complexity is lower than that of the traditional method, which is proportional to the logarithm of the input data volume.

**Keywords:** microblog retrieval; temporal characteristics; geographic characteristics; emotional characteristics



## 1 引言(Introduction)

过去几年中，网络媒体得到了飞速发展，越来越多的出版公司将重心从纸媒体转移到网络媒体。在线媒体通过社交网络平台完成点对点分享和广播。在博客和微博中，用户可以与特定人群共享信息，或向大量用户传播信息。由于微博的主体或元数据中包含了大量信息，因此，以微博时间、地理位置或空间特征为基础，可以提取重要事件及其发展趋势[1]。

微博的聚类检索是一个热门研究课题，已经有很多研究者对其进行了研究。王李冬等[2−3]提出了基于HowNet知识库系统的微博语义检索方法。杨震等[4]提出了一种微博检索结果的二次重排算法，基于微博内容相似关系构建关系图模型，利用PageRank算法对微博检索结果进行二次排序。SAMUEL等[5]提出了一个Lex−Rank算法的变体，以提取微博中存在的不同类型的时间信息，并将之用于摘要创建。韩中元等[6]提出

基金项目：江西省教育厅科学技术研究项目(GJJ191016).

了一种面向微博检索的基于词汇时间分布的查询扩展方法。DEMIRIZ等[7]提出了基于数据的空间和时间特征进行数据分析的方法，并使用模糊规则将该方法应用到欺诈检测任务中，表现出较好的性能。

本文的目标是开发一个含有微博时间、地理坐标和情感特征的框架，并使用这些特征进行聚类，建立起时间摘要处理的索引。本文提出了一个框架，以克服传统聚类(如K均值算法[8])算法的缺陷，并提出了一个多层级聚类方法，其中，空间特征进行1级聚类，时间特征完成2级聚类。同时，还可以基于情感对微博进行聚类。

## 2 提出的方法(Proposed method)

本文提出的方法主要以微博的时间、地理位置和情感特征为基础，对微博进行索引并创建聚类。以往的方法依靠用户指定的聚类数量，而本文的方法则基于建立的索引，自动评估聚类的数量。所提方法对K均值聚类做出了改进，有助于以微博的时间、地理位置和情感特征为基础，从微博中确定聚类的数量[9]。

首先，定义一个数据集 $D=\{d_1,d_2,d_3,...,d_n\}$，包含总计 $n$ 个文档，该数据集共 $m$ 维，有不同的模型 $M_j=\{C_1,C_2,...,C_k\}$，利用 $P(M_j\mid D)$ 完成对模型的评分。使用柯西−施瓦兹准则对后验进行逼近，如下所示：

$$BIC(M_j)=\hat{l}_j(D)-\frac{p_j}{2}\log n \tag{1}$$

式中，$\hat{l}_j(D)$ 为第 $j$ 个模型 $D$ 的似然对数，取最大似然点；$p_j$ 为 $M_j$ 中的参数数量，选择得分最高的模型。点概率的计算公式如下：

$$\hat{P}(x_i)=\frac{|C_{(i)}|}{n}\cdot\frac{1}{\sqrt{2\pi\hat{\sigma}^d}}\exp\left(-\frac{1}{2\pi\hat{\sigma}^2}\|x_i-\mu_{(i)}\|^2\right) \tag{2}$$

自由参数的数量为 $k-1+dk+1=(d+1)k$，X均值在全局用柯西−施瓦兹准则选择最佳模型，并在局部引导形心的分割。$k$ 的范围表示为 $[k_{\min},k_{\max}]$。开始时，X均值从 $k=k_{\min}$ 开始，并在需要时持续添加形心，直到达到上限为止。在该过程中，将得分最高的形心集合记录为最佳路线，并将之作为输出结果。对微博的定义如下：

$$T_{info}=(T_{ID},U,T,T_P,G,L,U_{ID},H,R_U,R_T,N_{RT}) \tag{3}$$

式中，$T_{ID}$ 为微博ID，$U$ 为用户名，$T$ 为微博正文文本，$T_P$ 为微博发表时间，$G$ 为发布微博的地理位置，$L$ 为微博语言，$U_{ID}$ 为用户ID，$H$ 为微博中包含的主题标签，$R_U$ 为回复微博，$R_T$ 为转发微博，$N_{RT}$ 为微博的转发数量。

每条微博中包含的特征数量不同，最高可能超过30 个特征。本文仅利用了少数几个特征，利用基于查询的方法完成对微博的索引，其中用户向系统提供搜索话题，利用该关键词建立一个索引。在建立索引的过程中，本文将首先对带噪数据的微博进行预处理，移除不包含原始内容的微博。

本文提出的基于时间和空间特征对微博进行聚类和索引的框架如图1所示。首先，移除时间和空间之外的其他特征，用包含微博用户所用的普通文本的最新词语和缩写形式的微博字典，对微博进行标准化，并从微博中移除停用词；然后，对微博进行词语切分，在微博上执行“词干”搜寻，将“词干”切分存储在数据库中，建立两个数据框架；最后，将查询与微博库进行匹配,如果数据框架中存在该词语，则该微博将被放入一个新的数据集中。利用X均值聚类算法[1−9]得出位置的数量和与该数量相对应形成的聚类数量，找出聚类的最优数量。完成初始聚类的形成后，在每个以地理位置特征形成的聚类上，完成基于微博时间特征的聚类，得到在地理位置特征中与微博的时间相关的2级聚类。

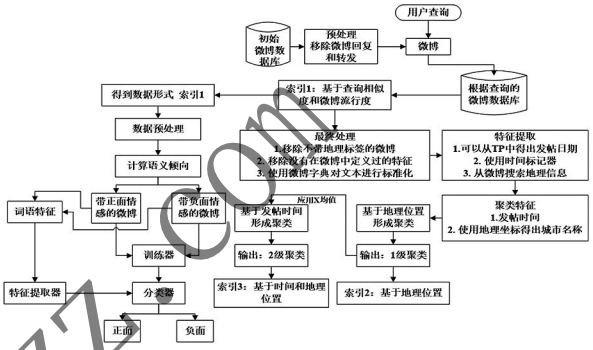


图1 本文提出的方法

Fig.1 The method proposed in this paper

## 3 实验与分析(Experiments and Analysis)

本文实验使用Intel Core i7处理器、RAM为16 GB的个人电脑作为实验平台，利用Fire−hose API得到所有的微博数据，包括地理位置信息的微博数量为134,540 条。数据收集于2019 年2 月至2019 年5 月。

基于Vincenty公式[9]，使用大圆距离计算出两个地理坐标之间的距离，以保证微博位置在用户设定的距离阈值内。如果该微博在阈值之外，则该微博形成一个单独的聚类。距离定义如下：

$$\Delta\sigma=\arctan\frac{\sqrt{(\cos\phi_2\cdot\sin(\Delta\lambda))^2+(\cos\phi_1\cdot\sin\lambda_1-\sin\phi_1\cdot\cos\lambda_2\cdot\cos(\Delta\lambda))^2}}{\sin\phi_1\cdot\sin\phi_2+\cos\phi_1\cdot\cos\phi_2\cdot\cos(\Delta\lambda)} \tag{4}$$

式中，$\phi_1$、$\lambda_1$为点1的纬度和经度；$\phi_2$、$\lambda_2$为点2的纬度和经度；$\Delta\sigma$为点之间的圆心角。

利用两个位置坐标，通过上述公式得出两个位置之间的距离。接着，进行如下实验：首先，计算两微博之间的距离，利用给定的阈值形成聚类；然后，利用微博的发帖时间对聚类内的微博再次进行聚类，即通过X均值完成该聚类；最后，利用微博的创建时间得出聚类。

### 3.1 评价分析

为了进行聚类评价，本文实验首先得出基于地理位置的第一个聚类，然后使用微博的创建时间对这些聚类再次进行聚类。基于地理坐标的聚类形成如图2所示，其中，“×”表

示聚类的中心。图3给出了聚类1中的聚类，基于微博事件再次形成聚类的结果。可以看出，相比于1级聚类，2级聚类具有更好的类间和类内的特征，特征样本更加清晰明了。

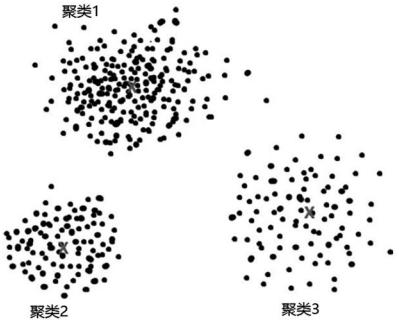


图2 1级聚类

Fig.2 Level 1 clustering

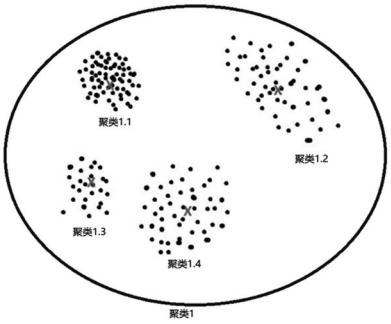


图3 2级聚类

Fig.3 Level 2 clustering

本文使用DBI指标和S系数两种方法进行评价，这两种指标数值越高，表示结果越好。不同方法的聚类评价结果如表1所示。实验中，每种方法在不同数量的微博上运行3次。由表1可知，在所有场景中，本文提出的系统均表现出超过其他聚类系统的性能。文献[5]提取微博中存在的不同类型的时间信息，并将之用于摘要创建，所用的元素比较少，获得的聚类结果较差。文献[8]使用较为传统的K均值聚类，在总体微博聚类过程中，使用的特征元素和层次较少。文献[7]将数据的空间和时间特征进行数据分析，取得了聚类结果最为接近本文的方法，优于文献[5]和文献[8]。总体来说，本文方法两种评价结果最优，其使用的特征元素和层次较为充分，因此，获得的聚类效果更好。

表1 聚类评价

Tab.1 Cluster evaluation

| 微博数量 | 56,000 | | 85,000 | | 100,000 | |
|---|---|---|---|---|---|---|
| 方法 | DBI | S系数 | DBI | S系数 | DBI | S系数 |
| 文献[5] | 0.503 | 0.511 | 0.531 | 0.533 | 0.547 | 0.579 |
| 文献[8] | 0.471 | 0.577 | 0.490 | 0.600 | 0.514 | 0.598 |
| 文献[7] | 0.545 | 0.547 | 0.565 | 0.561 | 0.575 | 0.591 |
| 本文方法 | 0.470 | 0.597 | 0.475 | 0.616 | 0.512 | 0.640 |

### 3.2 复杂度分析

本文提出框架的复杂度为 $O(N\log k)$，其中，$N$ 表示微博数量，$k$ 表示要形成的数据量。这表明所提方法的执行时间与输入数据的对数成正比，本文方法并不需要使用所有数据。传统微博K均值方法的复杂度为 $O(n^{dk+1}\log n)$，其中，$n$ 表示待聚类的项数，$k$ 表示要形成的聚类数，$d$ 表示维度。这表明其运行时间取决于因子数量，例如，待聚类的项数、要形成的聚类数和维度等。这证明与传统的微博聚类算法相比，所提方法的复杂度更低。

## 4 结论(Conclusion)

本文提出了一种基于微博的时间特征、地理位置和情感对微博进行聚类的方法，该方法能够对属于某个特定位置、某个特定的时间段或包含某种特定情感的微博进行聚类。在聚类之前，本文首先建立两个索引，分别用于非词干关键词和词干关键词，以达到有利于搜索过程和汇总过程的目的，使得微博的搜索工作量降低，搜索时间加快。

## 参考文献(References)

[1] 曹雾,张景鹏,胡含凯,等.基于文森特公式计算遥测天线理论跟踪弹道[J].探测与控制学报,2015,37(6):103-106.

[2] 王李冬,张慧熙.基于HowNet的微博文本语义检索研究[J].情报科学,2016,34(9):134-137.

[3] 王李冬,吕明琪.融合语义和时间因子的微博检索[J].情报杂志,2016,35(4):190-194.

[4] 杨震,张广源,范科峰.基于图模型决策的微博检索二次排序算法[J].北京工业大学学报,2017,43(1):94-99.

[5] SAMUEL A, SHARMA D K. Modified lexrank for tweet summarization[J]. International Journal of Rough Sets and Data Analysis (IJRSDA), 2016, 3(4):79-90.

[6] 韩中元,杨沐昀,孔蕾蕾,等.基于词汇时间分布的微博查询扩展[J].计算机学报,2016,39(10):2031-2044.

[7] DEMIRIZ A, LU B E. Fuzzy rule-based analysis of spatio-temporal ATM usage data for fraud detection and prevention1[J]. Journal of Intelligent & Fuzzy Systems, 2016, 31(02):805-813.

[8] 张云伟,宋安军.基于K-Means改进算法在微博话题发现中的应用研究[J].计算机系统应用,2016,25(10):308-311.

[9] 曹鹏,李博,栗伟,等.结合X-means聚类的自适应随机子空间组合分类算法[J].计算机应用,2013,33(2):550-553.

## 作者简介：

范怡敏(1981-)，女，硕士，讲师.研究领域：软件工程，大数据.