文章编号：2095-2163(2023)09-0164-04 中图分类号：TP274 文献标志码：A

基于流量时间序列的社交网络事件聚类分析

刘 静，张 鸽
（山西农业大学 软件学院，山西 太谷 030801）

摘 要：社交网络上用户发布的在线内容非常不稳定，用户对每个事件的关注度也随时变化。虽然每个事件的关注程度各不相同，但某些具有共同特征的事件会呈现出相似的流量模式，本文旨在根据社交网络事件的流量时间序列对事件进行聚类，找到事件的共性特征。首先，利用皮尔逊相关系数来确定各事件的主题标签；然后，利用各事件的主题标签获得每隔固定时间有关该事件的推文总量，即该事件的流量时间序列；最后，利用 K-SC（K-Spectral Centroid）聚类算法对事件的流量时间序列进行聚类，并分析聚类结果中每一类事件的共性特征。利用推特上 2020 东京奥运会期间场地自行车比赛事件的推文数据，验证了本文方法对基于流量时间序列的社交网络事件进行聚类分析的有效性。

关键词：社交网络；流量模式；流量时间序列；K-SC 聚类算法

Clustering analysis of social network events based on traffic time series

LIU Jing, ZHANG Ge
（School of Software, Shanxi Agricultural University, Taigu Shanxi 030801, China）

【Abstract】 The online content posted by users on social networks is highly unstable, and users´ interest in each event varies over time. Although the level of attention differs for each event, some events with common characteristics exhibit similar traffic patterns. This paper aims to cluster the events of social network based on their traffic time series to identify common features among them. Specifically, Pearson correlation coefficient is used to determine the thematic labels of each event. Then the total number of tweets related to each event at fixed intervals are calculated to represent the traffic time series of each event. Finally, the K-SC clustering algorithm is employed to cluster the traffic time series of the events, and the common features of events in the same cluster are analyzed. In the experiments, Twitter data of track cycling races during the 2020 Tokyo Olympics are used to validate the effectiveness of the proposed method in clustering social network events based on traffic time series.

【Key words】 social network; traffic pattern; traffic time series; ; K-SC clustering algorithm

0 引 言

社交网络是一个用户可以即时发布内容、分享信息并与他人进行交流互动的互联网平台，例如微博、推特等平台。社交网络上每时每刻都会产生用户发布的大量内容，表现出丰富的时间动态性。由于社交网络中个体用户行为难以预测，因此分析用户行为非常复杂。但一些研究表明，互联网在线内容的流量模式可以反映群体用户的行为[1-2]。事件随时间变化的流量时间序列可以显示出用户群体何时对事件产生关注度，以及关注度如何随时间增长消退的变化模式。对于某种类型的事件，群体用户的流量时间序列往往会形成一些特定的规律。

现有研究大多根据社交网络中的流量模式分析用户的行为，文献[3]根据用户在电商平台购买及评论的时间序列分析用户的消费行为习惯；文献[4]通过用户行为的时间序列分析来检测社交网络中的欺诈行为。而本文旨在利用事件的流量模式对事件进行聚类分析。具有共性特征的事件可能会呈现出相似的流量时间序列，可以根据事件的流量时间序列对事件进行聚类，找到事件的共性特征。K-means 算法是聚类领域最常使用的聚类算法，该算法通过在每次迭代中交替执行簇分配步骤和簇中心更新步骤，来最小化簇内样本到簇中心的距离平方，以找到最终的簇划分结果[5]。但该算法通常运用在数值型数据上，无法在时间序列数据上体现出数据的时序性；K-SC 算法是一种在 K-means 算法基础上改进的，针对时间序列数据进行聚类的算法，用来揭示在线内容时间动态性的规律[6]。该算法与 K-means算法流程相同，都是交替执行簇分配步骤

基金项目：山西省重点实验室开放基金（CICIP2021005）。
作者简介：刘 静（1990-），女，硕士，讲师，主要研究方向：聚类分析；张 鸽（1990-），女，硕士，讲师，主要研究方向：信息融合技术。
收稿日期：2023-07-03

和簇中心更新步骤，只不过根据时间序列数据的特点重新定义了距离和簇中心。本文利用 K-SC 算法对社交网络上热门事件的流量时间序列进行聚类分析，发现具有相似流量模式的事件的共同特征，并分析同类事件流量时间序列的共同特点。首先需要获得事件的流量时间序列，而确定事件的主题标签是获得事件流量时间序列的关键，本文利用皮尔逊相关系数来确定事件的主题标签，利用事件的主题标签每隔固定时间检索出包含该主题标签的推文，并计算推文数量，该数量形成的序列即为该事件的流量时间序列。然后就可以利用 K-SC 算法对多个事件的流量时间序列进行聚类，找到具有相似时间序列形状的事件，并分析其共性特征。

利用推特上 2020 东京奥运会期间场地自行车比赛事件的推文数据进行实验，获取事件的流量时间序列，并对其进行聚类分析，验证了本文方法可以对基于流量时间序列的社交网络事件进行有效聚类，从而发现同类事件的共性特征。

1　方法描述

1.1　获取事件的主题标签

社交网络平台上的事件是用主题标签来确定的，因此找到每个事件的主题标签是检测事件流量模式的关键。首先，根据事件最关键的主题标签获取包含该主题标签的推文；然后，查看这些推文中包含其他哪些主题标签，并将所有主题标签保存到一个初始标签池 H 中，记为 $H=\{H_1, H_2, H_3, \cdots, H_n\}$，其中 n 为当前主题标签的个数。然而，并非 H 中的所有主题标签都能直接与某个特定事件相关联，因此使用皮尔逊相关系数对 H 中的主题标签进行重新检查。首先，从 H 中选择那些必定与事件相关的主题标签，如事件的名称，并将其保存到标签池 R 中，且 $R \in H$；然后为了确定 H 中的各个主题标签是否与事件相关，需要将其逐个与 R 中的主题标签进行比较，计算 H 中每个主题标签与 R 之间的皮尔逊相关系数。如果相关性大于某个阈值，则来自 H 中的某个主题标签被认为是与事件相关的，并保留在 H 中；如果相关性小于阈值，则将其从 H 中删除。计算了相关性后，最终的 $H=\{H_1, H_2, H_3, \cdots, H_m\}$（其中 $m \leqslant n$）被视为事件的主题标签。

1.2　创建事件的流量时间序列

获取到每个事件的主题标签 H 后，就可以根据主题标签来计算事件流量随时间变化的流量时间序列。每隔一定时间 t，包含 H 的推文总量被认为是该事件的流量，但在 H 中可能会存在一些噪声主题标签与 H 中真正的主题标签非常相似但又不属于该主题。为了减少噪声，需要确保那些包含与 H 相似的噪声主题标签的推文不被收集。首先按照目标时间间隔查询数据库，收集与 H 相似但不包含在 H 中的主题标签，结果集记为 h；然后，在目标时间间隔 t 上选择包含 H 但不包含 h 的推文，记录每隔 t 分钟的推文数量，该事件即可获得一个离散的时间序列 $E_i(mt)$（其中 $m=1, 2, 3, \cdots, M$，M 为时间序列的长度），即事件 i 在以 t 为时间间隔的 mt 时刻上被提及的次数，这个离散的时间序列即为事件的流量时间序列。

1.3　利用流量时间序列对事件进行聚类

获取了社交网络中事件的流量时间序列后，根据流量时间序列对事件进行聚类。本文利用经典的 K-SC 算法对事件的时间序列进行聚类。K-SC 算法通过在每次迭代中交替执行两个步骤，即簇分配步骤和簇中心更新步骤，来最小化簇内样本到簇中心的距离平方和，即式(1)：

$$L=\sum_{k=1}^{K}\sum_{E_i\in C_k} d(E_i,\mu_k)^2 \tag{1}$$

其中，K 为簇的个数；C_k 为第 k 个簇的样本集合；$d(E_i,\mu_k)$ 为时间序列 E_i 到簇中心 μ_k 的距离。

K-SC 算法将时间序列样本到簇中心的距离定义为式(2)：

$$d(E_i,\mu_k)=\min_{\alpha_i,q_i}\frac{\|\alpha_i E_{i(q_i)}-\mu_k\|}{\|\mu_k\|} \tag{2}$$

其中，α_i 为用于匹配两个时间序列形状的缩放系数，$E_{i(q_i)}$ 为将时间序列 E_i 平移 q_i 个时间单位的结果，使得 E_i 和 μ_k 在相同的时间达到峰值。

在 K-SC 的算法流程中，首先随机选择 K 个初始样本作为簇中心；在簇分配步骤中，将每个样本分配到与簇中心距离最近的簇；在簇中心更新步骤中，新的簇中心 μ_k^* 应该使得对于所有样本的 $d(E_i,\mu_k)$ 的和最小，即式(3)：

$$\mu_k^*=\arg\min_{\mu_k}\sum_{E_i\in C_k} d(E_i-\mu_k)^2 \tag{3}$$

经过推导可得出公式(4)：

$$\mu_k^*=\arg\min_{\mu_k}\frac{\mu_k^{\mathrm{T}} M\mu_k}{\|\mu_k\|^2} \tag{4}$$

其中：

$$M=\sum_{E_i\in C_k}\left(I-\frac{E_i E_i^{\mathrm{T}}}{\|E_i\|^2}\right) \tag{5}$$

因此求解 μ_k^* 等价于求解矩阵 $\boldsymbol{M}$ 的最小特征向

量。根据上述迭代步骤,K-SC 的算法流程见算法 1。

算法 1 K-SC 聚类算法

输入 时间序列 $E_i, i = 1,2,\cdots,N$;簇个数 K;初始簇中心 $\mu_k, k = 1,2,\cdots,K$

重复

$C = \{C_1,\cdots,C_K\} \leftarrow \phi$

for $i = 1$ to N :(簇分配步骤)

$k^* \leftarrow \arg\min_k d(E_i,\mu_k)$

$C_{k^*} \leftarrow C_{k^*} \cup \{i\}$

end for

for $k = 1$ to K:(簇中心更新步骤)

$$M \leftarrow \sum_{E_i \in C_k} \left(I - \frac{E_i E_i^{\mathrm{T}}}{\|E_i\|^2}\right)$$

$\mu_k \leftarrow M$ 的最小特征向量

end for

直到 L 不变

输出 $C,\mu_1,\cdots,\mu_K$

2 实验分析

2.1 数据集

实验采用推特上 2020 东京奥运会的数据,创建比赛事件的流量时间序列,并对事件进行聚类分析,找到同类事件的一般特性。本文实验通过推特流媒体 API 获取本届奥运会期间与场地自行车决赛项目相关的公开推文,对表 1 中的 8 项场地自行车比赛事件的流量时间序列进行聚类。首先,获取各事件的初始主题标签,其中将“比赛名称”和“获胜者名字”视为最关键的主题标签;然后,获取各事件最终的主题标签,其中皮尔逊相关系数的阈值设定为 0.8;最后,创建各事件的流量时间序列,将时间间隔 t 设置为 5 min,并为各事件的时间序列选取合理且相同长度的时间区间。

表 1 本实验选取的场地自行车比赛名称

Tab. 1 Names of the selected track cycling races in this experiment

场地自行车决赛项目
Women Sprint Final(女子竞速赛决赛)
Women Keirin Final(女子凯琳赛决赛)
Women Team Sprint Final(女子团体竞速赛决赛)
Women Team Pursuit Final(女子团体追逐赛决赛)
Men Sprint Final(男子竞速赛决赛)
Men Keirin Final(男子凯琳赛决赛)
Men Team Sprint Final(男子团体竞速赛决赛)
Men Team Pursuit Final(男子团体追逐赛决赛)

2.2 确定簇的个数

对事件进行聚类前,首先需要确定簇的个数。本实验在不同的簇个数(即令 $K = 1, 2, 3, 4, 5, 6$)下执行 K-SC 算法,得到目标函数值 L(即距离平方和)收敛后的值,目标函数 L 的值随簇个数 K 的变化如图 1 所示。从图 1 中可以看出,在 $K = 3$ 后,目标函数值的下降率变得平缓,簇个数对聚类目标值影响不大,因此将簇的个数设置为 3。

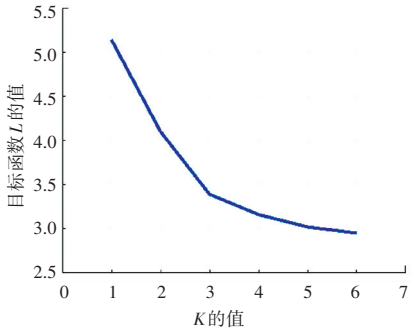


图 1 目标函数 L 的值随簇个数 K 的变化

Fig. 1 The change of the value of objective function L with respect to the number of clusters K

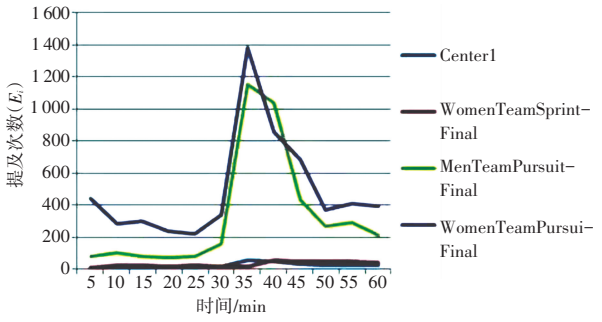
2.3 聚类结果分析

获取了场地自行车比赛事件的流量时间序列并确定了簇的个数后,用基于时间序列的 K-SC 算法对事件进行聚类。聚类结果见表 2,可以看出这些比赛事件基本按照是否为团体项目被划分开;每个簇中每个事件的流量时序折线图如图 2 所示,可以看出除了簇 2 中事件的时间序列形状较为特殊之外,团体项目和个人项目显示出不同的时间序列形状。簇 1 中的团体项目事件都只有一个明显较高的峰值,而簇 3 中的个人项目(男子或女子非团体项目)事件都显示出两个明显的峰值。实验结果表明,聚类结果中同一类事件具有明显的共性特征,并且显示出类似的流量时序模式,从而验证了 K-SC 算法对于基于流量时间序列的社交网络事件聚类的有效性。

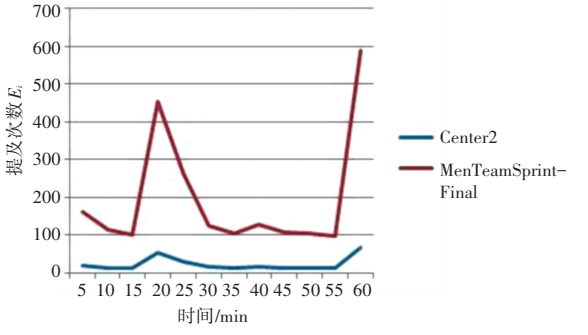
表 2 基于流量时间序列的场地自行车比赛事件聚类结果

Tab. 2 Clustering results of track cycling events based on traffic time series

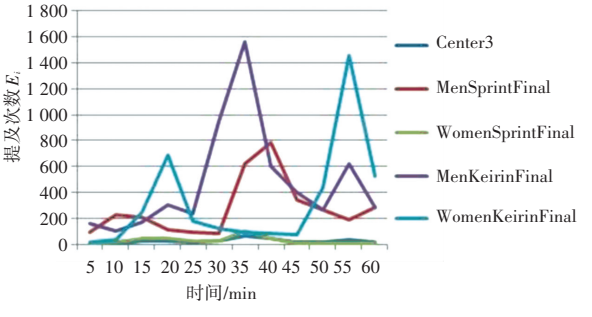
簇序号	比赛事件名称
簇 1	Women Team Sprint Final(女子团体竞速赛决赛)
	Men Team Pursuit Final(男子团体追逐赛决赛)
	Women Team Pursuit Final(女子团体追逐赛决赛)
簇 2	Men Team Sprint Final(男子团体竞速赛决赛)
簇 3	Men Sprint Final(男子竞速赛决赛)
	Women Sprint Final(女子竞速赛决赛)
	Men Keirin Final(男子凯琳赛决赛)
	Women Keirin Final(女子凯琳赛决赛)



（a）簇 1 中事件的流量时序折线图



（b）簇 2 中事件的流量时序折线图



（c）簇 3 中事件的流量时序折线图

图 2　每个簇中比赛事件的流量时序折线图

Fig. 2　Line graphs of the traffic time series for events within each cluster

3　结束语

社交网络上用户实时生成的在线数据表现出丰富的时间动态性，但一些具有共性特征的事件可能会呈现出相似的流量模式。本文根据事件的流量时间序列对事件进行聚类，找到事件的共性特征。首先，利用皮尔逊相关系数来确定各事件的主题标签；然后，利用各事件的主题标签获得各事件的流量时间序列；最后，利用 K-SC 聚类算法对多个事件的流量时间序列进行聚类，发现同类事件的共性特征。利用推特上 2020 东京奥运会期间场地自行车比赛事件的推文数据验证了本文方法可以对基于流量时间序列的社交网络事件有效聚类，从而发现同类事件的共性特征。

参考文献

[1] RATKIEWICZ J, FLAMMINI A, MENCZER F. Traffic in social media I: paths through information networks[C]//2010 IEEE second international conference on social computing. IEEE, 2010: 452-458.

[2] DING S, LAI K, WANG D. A study on the characteristics of the data traffic of online social networks[C]//2011 IEEE International Conference on Communications (ICC). IEEE, 2011: 1-5.

[3] 张艳丰. 在线用户评论行为时间序列关联特征规律研究[D]. 长春：吉林大学，2018.

[4] XIAO Q, BERTINO E. Detecting deceptive engagement in social media by temporal pattern analysis of user behaviors: a survey[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2017, 7(5): e1210.

[5] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C]//Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1967: 281-297.

[6] YANG J, LESKOVEC J. Patterns of temporal variation in online media[C]//Proceedings of the fourth ACM international conference on Web search and data mining. 2011: 177-186.