面向热点话题时间序列的有效聚类算法研究

韩忠明[1)] 陈 妮[1)] 乐嘉锦[2)] 段大高[1)] 孙践知[1)]

[1)](北京工商大学计算机与信息工程学院 北京 100048)

[2)](东华大学计算机科学与技术学院 上海 200051)

摘 要 聚类热度时间序列是揭示和建模网络热点话题形成与发展的重要过程. Leskovec 等人在 2010 年提出面向话题时间序列的 K_SC 聚类算法,其精确度较高且能较好地刻画话题内在发展趋势特征. 但 K_SC 算法具有对初始类矩阵中心高度敏感、高时间复杂度等特性,使其难以在实际高维大数据集上应用. 文中结合小波变换技术,提出一个新的迭代式聚类算法 WKSC,主要提出两个创新:(1)用 Haar 小波变换将原始时间序列进行压缩,降低原始时间序列的维度,从而降低了算法的时间复杂度;(2)在 Haar 反小波变换中,将低维聚类返回得到的矩阵中心作为高维聚类的初始矩阵中心,在迭代聚类过程中优化了对初始矩阵中心高敏感性的问题,提高了聚类的效果. 文中分别采用国内外 3 个数据集作为测试样本,进行了大量的实验. 实验结果表明 WKSC 算法能显著降低聚类的时间复杂度,同时改进聚类效果. WKSC 算法可很好的应用于大量高维热点话题的模式分析.

关键词 聚类; 时间序列; 热点话题; 小波

中图法分类号 TP391 **DOI 号**: 10.3724/SP.J.1016.2012.02337

An Efficient and Effective Clustering Algorithm for Time Series of Hot Topics

HAN Zhong-Ming[1)] CHEN Ni[1)] LE Jia-Jin[2)] DUAN Da-Gao[1)] SUN Jian-Zhi[1)]

[1)](*School of Computer Science and Information Engineering, Beijing Technology and Business University, Beijing* 100048)

[2)](*School of Computer Science, Donghua University, Shanghai* 200051)

Abstract Hot degree time series clustering is very important for revealing and modeling development process of hot topics in Web sites. In 2010, Leskovec and his colleagues proposed a K-Spectral Centroid (K_SC) time series clustering algorithm, which has higher accuracy and can be used to better describe the trend of hot topics. But K_SC algorithm is sensitive to the initialization of cluster centers and has high time complexity. Therefore, it is difficult to directly apply K_SC to high dimensional data. Based on wavelet transform technology, a new iteration clustering algorithm—WKSC is proposed in this paper, which has two improvements: (1) the original time series are compressed by Haar wavelets transform to lower dimensions of the original time series. WKSC algorithm groups topics based on lower dimensions time series and the time complexity is reduced; (2) the clustering results from previous iteration of K_SC are used as the initial assignment at the high level, then the high sensitivity to cluster centers is solved. Three datasets from different sources were selected and comprehensive experiments were conducted. Experimental results show that WKSC algorithm can significantly reduce time complexity, and improve the quality of clustering result, which means WKSC algorithm can be used on massive and high dimension hot topics.

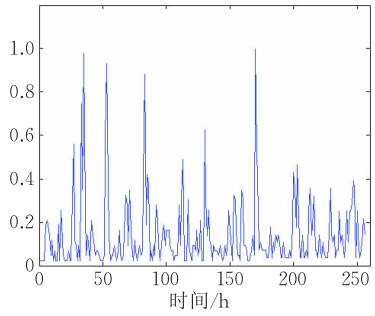
Keywords clustering; time series; hot topics; wavelet

收稿日期:2012-06-05;最终修改稿收到日期:2012-08-31. 本课题得到国家自然科学基金(61170112)、北京市属高等学校科学技术与研究生教育创新工程建设项目(PXM2012_014213_000037)资助.**韩忠明**,男,1972 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为互联网数据分析与挖掘、海量数据处理等. E-mail: hanzm@th.btbu.edu.cn. **陈 妮**,女,1987 年生,硕士研究生,主要研究方向为互联网数据分析与挖掘等. **乐嘉锦**,男,1951 年生,教授,博士生导师,主要研究领域为数据仓库与数据挖掘等. **段大高**,男,1976 年生,博士,副教授,主要研究方向为多媒体数据挖掘等. **孙践知**,男,1967 年生,副教授,主要研究方向为复杂网络分析与挖掘等.

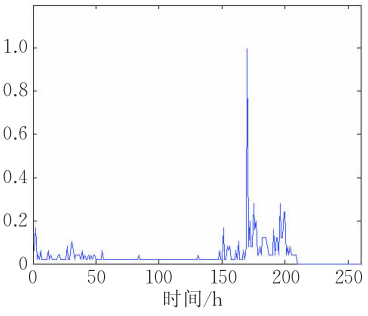
1 引 言

交互式网络(WEB 2.0)上的热点话题不仅极大地影响着虚拟网络社会中各种事件的形成与发展,同时也影响着真实人类社会中人们对于事件的看法和判断,甚至于影响着政府与司法机构对事件的判决.热点话题中的参与数量(如发帖数、转发数、评论数)是衡量话题热度的重要度量,这些度量随着时间而变化,呈现为一个时间序列.

不同来源、不同媒介的网络热点话题的热度随着时间的发展呈现一定规律的变化[1-4],不同类型话题的发展趋势具有不同的规律性,如图 1 所示.其中图 1(a)是关于劫持事件话题的热度序列图,有数个高峰,衰减速度比较缓慢;图 1(b)是关于艺人大婚话题的热度序列图,只有一个高峰,人们对该事件的关注度很快下降,具有"高挥发性"的特点.由此可见网上的话题呈现出丰富的时态信息.



(a) 劫持事件话题的热度序列图



(b) 艺人大婚话题的热度序列图

图 1 两个热点事件的热度时间序列

对热点话题的热度时间序列进行聚类并建模分析[5-6],从而了解话题的传播模式,这个研究方向已经引起了大量学者的高度兴趣.对热点话题聚类通常采用的方法有两种:(1)基于内容对热点话题聚类;(2)基于时序特征对热点话题聚类.基于内容的话题聚类技术可以有效地识别和跟踪具有相似内容的话题,被广泛地应用在 TDT 和文本分析领域.但是这类方法无法刻画话题的发展趋势,难以满足对话题建模的需要.而基于时序特征对热点话题聚类则将焦点定位于话题的发展趋势,较少考虑话题内容.这类方法的聚类结果可以刻画话题的发展趋势,为话题建模和预测奠定基础.

Yang 等人[7]致力于基于时序特征对热点话题进行聚类,提出了基于话题热度趋势的 K_SC 聚类算法.K_SC 算法为了刻画两个话题发展趋势的内在规律特征,提出了新的时间序列差异度公式,保证任意两个时间序列的相似性只与它们的趋势走向有关,而和它们的峰值数值以及在何时达到峰值无关.实验表明该算法对话题聚类具有很好的效果,但其对初始类矩阵中心的高度敏感、高时间复杂度使其难以在大数据集上应用.

本文针对 K_SC 的高时间复杂度、对初始类别矩阵中心高度敏感特性进行了一系列改进.在其基础上融合了小波变换技术的思想[8],采用了降维、逐层更新各个类的矩阵中心的方法,提出新的聚类算法 WKSC(Wavelet-Based K_SC Algorithm),并采集实际数据进行了大量的实验分析.实验结果表明 WKSC 算法在与 K_SC 算法同一实验条件,同一数据集的前提下能有效降低聚类的时间复杂度并提高聚类的质量.

本文第 2 节回顾相关研究工作;第 3 节深入描述和分析 WKSC 时间序列聚类算法;第 4 节进行两个算法的复杂度理论分析;第 5 节描述本文的实验设置与实验结果分析;第 6 节进行总结并探讨下一步可能的研究方向.

2 相关研究

2.1 时间序列聚类相关研究

时间序列聚类是一种完全根据数据自身所提供的信息进行分类的一种方法,因而要求面对数据挖掘的聚类算法应具有一定的自适应性.文献[9]提出了分割时间序列的方法,对分割后的子序列进行聚类、分类、异常检测、时间序列建模;文献[10]提出了用 OLS 算法实现对在线时间序列的分割,OLS 算法能够有效地在线检测出数据挖掘应用中感兴趣的关键变化点,而且"过拟合"程度低;文献[11]提出了将时间序列进行线性划分的方法,利用线段来近似表示时间序列从而获取时间序列的变化趋势,这一研究工作让后来的研究者产生了如何通过降维的时间序列最大程度地保留原时间序列的信息的想法;文献[12]对文献[11]提出线性划分方法进行了详细的评述并予以扩充;文献[13]提出了基于斜率提取

边缘点的时间序列分割.

2.2　热点话题建模与聚类相关研究

近年来，大量研究者对网络上的话题与行为进行了深入研究，这些研究表明用户的行为(话题)可以被建模和理解，但首先需要对话题进行聚类. 文献[6]开启了人们对互联网上人类行为动力学的研究，指出人类行为的特性不能用传统的泊松过程进行描述，可能存在复杂的动力学机制. 随后学者们展开了大量的相关性研究和考证. 文献[14-15]研究了人们在博客系统和评价网络中的级联行为，结果都非常符合文献[14]提出的长尾理论；文献[16]打破了前人用局部特征表征整个序列的方法，提出了基于全局相似性的 GSClu 算法，用来进行序列的聚类. 但是要对互联网上人们的行为进行建模却是一件很困难的事，因为隐藏在这些背后的行为是高度不可预测的[5-7].

从海量的数据中挖掘有价值的信息是数据挖掘研究的目标，数据聚类是数据挖掘中最常用、最有效的手段之一. 层次聚类算法和基于划分的 K-Means 聚类算法是最主要的两种聚类方法. 层次聚类算法通用性强，但只能被应用在小数据集上；基于划分的 K-Means 聚类算法简单、快速而且能有效地处理大数据集，但需要事先定义类的数量，距离矩阵公式不合理且时间复杂度高. 针对 K-Means 算法应用在时间序列问题上的主要缺点，文献[7]提出了 K_SC 算法，定义了新的时间序列差异度公式和更新矩阵中心(简称距心)的公式，但并没有解决高时间复杂度的问题.

3　WKSC 时间序列聚类算法

3.1　相关定义

热点话题. 对于一条消息(帖子、微博)，如果该消息在其所在的网站被标注为热点话题，或者评论(转发、回复)超过 5000，我们称之为热点话题.

热度. 对于一个热点话题，在时间间隔 Δt 内被关注(用户发表评论和对该热点话题的报导)的次数称为该热点话题在时间间隔 Δt 内的热度.

热度序列. 通过记录一定时间范围内的热度值能得到关于该话题热度的时间序列，称为热度时间序列，简称热度序列. 根据话题热度序列，可以画出热度时间序列图，表示该热点话题的关注度发展趋势，即该话题受到的关注度是怎么随着时间的推移而发生改变.

中心曲线. 聚类结果的每一个类别中所有时间序列成员共同形成的矩阵中心曲线称为中心曲线，每一个类的矩阵中心表示该类成员的共同模式特征.

我们旨在用比 K_SC 算法更短的执行时间将具有相同发展趋势特征的时间序列聚在同一个类里，不同趋势特征的时间序列聚在不同类，其中第 i 个类用 C_i 来表示.

3.2　K_SC 算法分析

文献[7]为了刻画两个话题趋势的内在规律特征，提出了新的时间序列差异度公式，保证任意两个时间序列的相似性只与它们的趋势走向有关，而与它们的峰值数值以及在何时达到峰值无关.

K_SC 算法采用了基于划分的聚类方法，先随机地把所有的时间序列进行分类，根据矩阵中心计算公式计算出每个类的矩阵中心. 再根据算法定义的差异度矩阵，把所有的时间序列归类到和它差异度最小的类中，最后更新该类的矩阵中心. K_SC 算法是一个迭代过程，其停止条件是每类中的成员不再变化或达到预定义的迭代次数. 最后形成的聚类成为最优分类.

(1) 差异度公式

$$\hat{d}(\boldsymbol{x},\boldsymbol{y})=\min_{\alpha,q}\frac{\|\boldsymbol{x}-\alpha\boldsymbol{y}_{(q)}\|}{\|\boldsymbol{x}\|} \tag{1}$$

其中，$\boldsymbol{y}_{(q)}$ 表示时间序列 $\boldsymbol{y}$ 经过 q 个时间单位的平移后所形成的时间序列，并且 $\boldsymbol{y}_{(q)}$ 和时间序列 $\boldsymbol{x}$ 的峰值处于同一时间点上；α 为比例系数，即将 $\boldsymbol{x}$ 和 $\boldsymbol{y}_{(q)}$ 置于同一时间轴时，峰值的比例系数，y 轴的最大值量化为 1. 式(1)表明任意两个时间序列的相似性只与它们的趋势走向有关，而和它们的峰值数值以及在何时达到峰值无关.

(2) 更新矩阵中心公式

$$\begin{aligned}\mu_k^* &=\arg\min_{\boldsymbol{\mu}}\sum_{\boldsymbol{x}_i\in C_k}\hat{d}(\boldsymbol{x}_i,\boldsymbol{\mu})^2\\ &=\arg\min_{\boldsymbol{\mu}}\sum_{\boldsymbol{x}_i\in C_K}\min_{\alpha_i,q_i}\frac{\|\alpha_i\boldsymbol{x}_{i(q_i)}-\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2}\\ &=\arg\min_{\boldsymbol{\mu}}\frac{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|^2}\end{aligned} \tag{2}$$

其中 $\boldsymbol{M}=\sum_{\boldsymbol{x}_i\in C_K}\left(\boldsymbol{I}-\frac{\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}}{\|\boldsymbol{x}_i\|^2}\right)$，$\mu_k^*$ 表示完成一次聚类之后的矩阵中心，$\boldsymbol{x}_i$ 表示第 i 类的各个成员. 式(2)本质在于找到该类中的新矩阵中心，使其和类中所有成员的平方和最小，降低了类中离异值的影响.

3.3　WKSC 算法

K_SC 算法的时间复杂度很高，对于 100 个具

有 128 个时间点的序列，算法迭代过程中每次都需要 $100\times128^3=2097152000$ 次的差异度计算. 此外，K_SC 算法对初始类的选择很敏感，如果初始类选择较差，则聚类的收敛过程非常缓慢. 如何在聚类过程中进行降维，提高算法对初始数据的有效选择以致降低聚类算法的时间复杂度是本文算法研究的出发点.

本文在小波变换[8]的基础上提出了改进的 K_SC 算法，称为 WKSC(Wavelet-Based K_SC Algorithm)算法. WKSC 可以分为两步：(1) 小波分解；(2) 还原高维并聚类. WKSC 算法利用小波技术对高维数据进行分解，可以实现高维数据的降维，在低维数据上进行聚类，具有很高的效率，再把低维聚类的结果作为迭代的基础就能有效解决 K_SC 算法对初始类别的敏感性问题.

我们采用 Haar 小波实现高维数据的降维[8]，Haar 小波是通过把维度为 N 的时间序列的两个相邻值取平均值的方法，得到一个平滑的、$N/2$ 维度的新时间序列，并记录这两个相邻值的差异值，用作反小波变换的参数. WKSC 算法通过将所有的热度时间序列都进行完全小波变换，即经过 $\log_2^N$(N 指时间序列的维度)层的变换，最终一个时间序列的维度为 1，如图 2 所示. 然后从低维序列开始进行聚类，聚类算法采用 K_SC 的核心. 由于数据维度很低，所以聚类将很快完成，但是低维数据无法刻画原始序列的趋势和特征，聚类得到的成员和中心曲线效果都可能较差，所以我们根据反小波变换，将序列逐步进行高层次的还原，对高层次的序列进行聚类，并采用低维聚类结果作为高层聚类的初始矩阵中心.

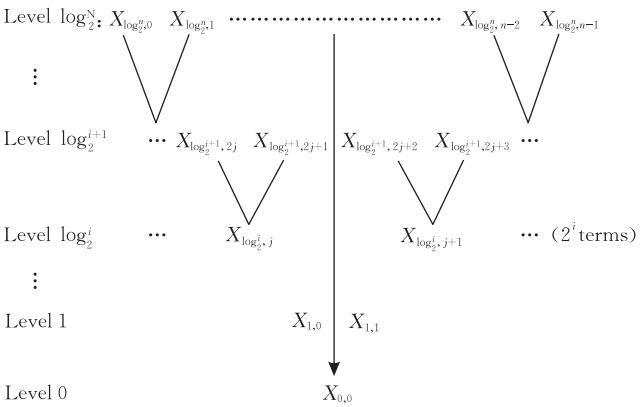


图 2 热度时间序列的 Haar 小波变换过程示意图

算法在迭代过程中，采用两种结束条件：

(1) 如果低层时间序列的聚类情况在高层聚类时没有发生改变则跳出整个循环，迭代结束，聚类完成.

(2) 指定算法在反小波变化到指定层次时，迭代结束，聚类完成.

具体算法描述如算法 1 所示. 算法第 1 行到第 3 行进行不同层次的离散 Harr 小波变换，得到每次变化的结果，存储形式为向量. 第 4 行到第 10 行对不同层次的时间序列进行聚类. 首先对最高阶的变换结果进行反小波变换，得到压缩比例最高的序列，对其进行 K_SC 算法聚类，其值作为下一次聚类的初始值，然后依次循环，直到算法结束条件为真.

算法 1. WKSC 算法描述.

输入：N 个维度为 L 的时间序列，K 个随机类 $C=\{C_1,\cdots,C_K\}$

输出：K 个类的矩阵中心

定义：起始层用 S 表示

1. for $i=1$ to N do
2. $\quad \boldsymbol{z}_i \leftarrow$ Discrete Haar Wavelet Transform ($\boldsymbol{x}_i$);
3. end for
4. for $j=S$ to $\log_2(L)$ do
5. $\quad$ for $i=1$ to N do
6. $\quad\quad y_i \leftarrow$ Inverse Discrete Haar Wavelet Transform ($\boldsymbol{z}_i(1:2^j)$){$\boldsymbol{z}_i(1:n)$ means the first n elements of $\boldsymbol{z}_i$};
7. $\quad$ end for
8. $\quad (C,\mu_1,\cdots,\mu_k) \leftarrow$ K_SC($\boldsymbol{y},C,K$);
9. $\quad$ if (finish(C)) break;
10. end for
11. return $C,\mu_1,\cdots,\mu_k$.

4 复杂度分析

在评价 K_SC 与 WKSC 聚类算法的时间复杂度时，我们均做如下假设：有 N 个时间序列，每个时间序列的维度为 L，初始定义类的个数为 K.

4.1 K_SC 算法复杂度分析

在更新每个类的矩阵中心步骤时，计算每个时间序列的矩阵 $\boldsymbol{M}$(见式(2))及其特征值的时间复度为 $O(L^3)$，那么计算全部类的矩阵中心的复杂度为 $O(\max(NL^2,KL^3))$. 挑选出每个类所属成员需要花费 $O(KNL)$，所以执行一次 K_SC 算法的聚类时间复杂度为 $O(\max(NL^2,KL^3))$. 由于 K_SC 算法对初始矩阵中心的高度敏感性，通过调用 K_SC 算法多次，反复更新矩阵中心才能达到聚类最优，最终聚类完成. 假设需要迭代的次数为 P，则其时间复杂度是 $O(P\times\max(NL^2,KL^3))$.

4.2 WKSC 算法复杂度分析

在 WKSC 算法中，计算每个时间序列的完全 Haar 小波变换需要耗费 $O(L)$的时间，设 $L=2^n$. 当

对处于 i 层的时间序列(维度为 2^i)进行 WKSC 算法聚类时，同样需要计算每个时间序列的矩阵 $\boldsymbol{M}$ 和 $\boldsymbol{M}$ 的特征值，其复杂度为 $O((2^i)^3)$.

设 $L_i=2^i$，所以 $L_i=L/2^{n-i}$，那么第 i 层计算每个时间序列的矩阵 $\boldsymbol{M}$ 和其特征值的复杂度为

$$O(L_i)=O((L/2^{n-i})^3)=O(L^3/2^{3\times(n-i)}),$$

所以执行一次 WKSC 算法的聚类时间复杂度为

$$O\left(\max\left(N\frac{L^2}{2^{2\times(n-i)}},K\frac{L^3}{2^{3\times(n-i)}}\right)\right).$$

相对于 K_SC 算法的聚类时间复杂度而言，由于 WKSC 算法中参与聚类的时间序列的维度比原始时间序列的维度缩减了很多，从而降低了复杂性的阶数.

WKSC 算法和 K_SC 算法初始的矩阵中心都是随机的，但不同之处在于 WKSC 算法将低维聚类矩阵中心作为高维聚类初始矩阵中心，这就让矩阵中心在低维时间序列的聚类中逐渐趋于最优值，在 WKSC 算法运行到高维时间序列时，调用聚类算法的次数也将显著减少.

5　实验比较与分析

5.1　实验设置

实验共使用 3 个数据集，前两个数据集均来自 Stanford 大学①. MemePhr 数据集选自博客和网站上的 1000 个热门帖子和新闻，以每小时的评论数作为热度，维度为 128；Twhtag 数据集选自 twitter 上的 1000 个热门帖子，以每小时该话题被提到的次数作为热度，维度为 128；第 3 个数据集来自我们从天涯和百度贴吧上采集的 314 个热门话题(简称为 ChinDt)，以每小时的评论数作为热度，记录热度时间序列，维度为 256. 本文的实验均在同一平台之下进行，我们采用 matlab 实现 WKSC 聚类算法②.

MemePhr 数据集的大小为 125 KB，Twhtag 数据集为 120 KB. ChinDt 原始数据集大小为11.8 MB，经过预处理后的结果为 771 KB.

5.2　实验结果与分析

我们对 K_SC 和 WKSC 算法分别进行了不同粒度上的效率和效果评价实验. 文献[7]的实验中，类别个数设为 6，为了客观比较，本文也选择聚类的类别个数是 6.

5.2.1　K_SC 算法和 WKSC 算法效率比较

因为 K_SC 算法无法对维度进行分层处理，所以需要将 2 个算法在同一个维度下进行实验. 数据集 ChinDt 计算到维度 256，其它两个数据集计算到维度 128. 对 K_SC 和 WKSC 算法在 3 个数据集下进行聚类所消耗时间结果如图 3 所示，其中时间单位为 s.

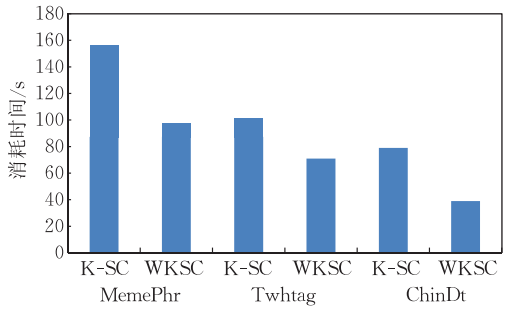


图 3　2 个算法在 3 个数据集下的消耗时间结果

从图 3 上可以看出两个结论：(1) 相对于 K_SC 算法而言，WKSC 算法所消耗时间对不同的数据集都有了较大的改进，至少减少了 30% 的时间；(2) K_SC 算法对序列的维度和发展变化特性的敏感性高于对序列个数的敏感性. 如 Twhtag 数据集中的序列个数要比 ChinDt 数据集中序列个数多 2.18 倍，但是 K_SC 算法所需时间并没有显著减少. 其原因是 ChinDt 数据集中序列变化趋势比 Twhtag 数据集中的序列变化趋势复杂、维度高. 由于 WKSC 算法采用了降维和提高初始类比精度的策略，所以虽然 ChinDt 数据集的维度较高，但是也能在短时间内进行聚类.

话题热度序列聚类的目的是探索其发展趋势，可以用中心曲线来直观地展示聚类的效果，表示出每个类中成员的共同模式特征. 图 4～图 9 给出了两个算法对不同数据集聚类得出的中心曲线.

中心曲线结果图上每一个曲线代表一个类. 从图 4～图 9 上可以看出，MemePhr 数据集下 K_SC 和 WKSC 算法的类别对应关系分别为(a)-(f)，(b)-(b)，(c)-(e)，(d)-(a)，(e)-(c)，(f)-(d). Twhtag 数据集下类别对应关系为(a)-(f)，(b)-(a)，(c)-(e)，(d)-(b)，(e)-(d)，(f)-(c). ChinDt 数据集下类别对应关系分别为(a)-(e)，(b)-(c)，(c)-(b)，(d)-(f)，(e)-(d)，(f)-(a). 说明：对应关系中前者为 K_SC 算法对应的类，后者为 WKSC 算法对应的类.

从中心曲线上分析，K_SC 和 WKSC 算法得出的类别趋势基本相同，尤其是对于 MemePhr 和 Twhtag 数据集. 对于 ChinDt 数据集，存在部分差异，但是每个类的发展趋势相同，差异在于趋势过程中的一些波动，但并没有造成序列趋势的改变.

WKSC 算法在反小波变换的过程中可以在任意维

① http://snap.stanford.edu/data/volumeseries.html
② 获取算法程序，请联系作者邮箱.

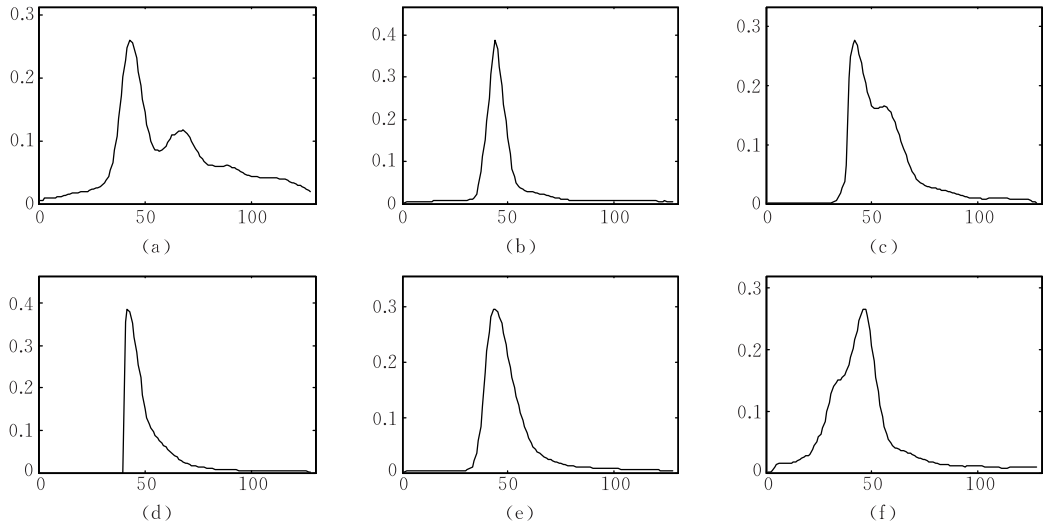


图 4 K_SC 算法在 MemePhr 数据集下的中心曲线

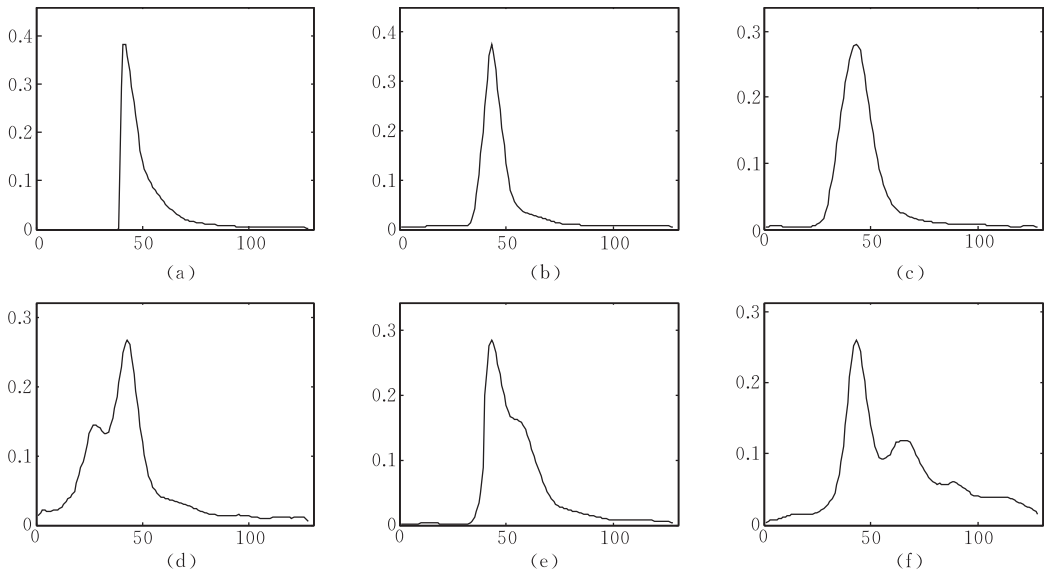


图 5 WKSC 算法在 MemePhr 数据集下的中心曲线

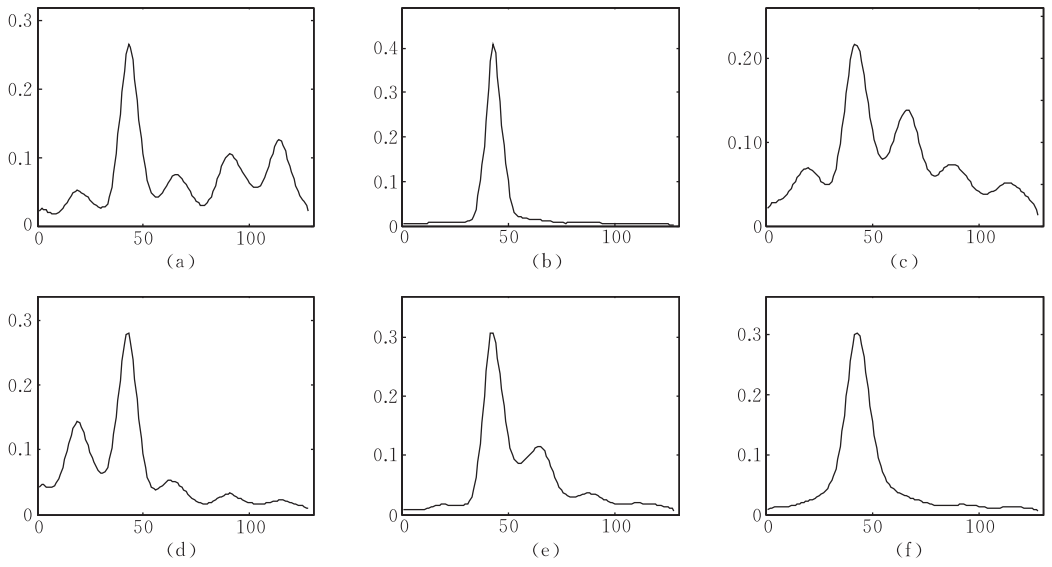


图 6 K_SC 算法在 Twhtag 数据集下的中心曲线

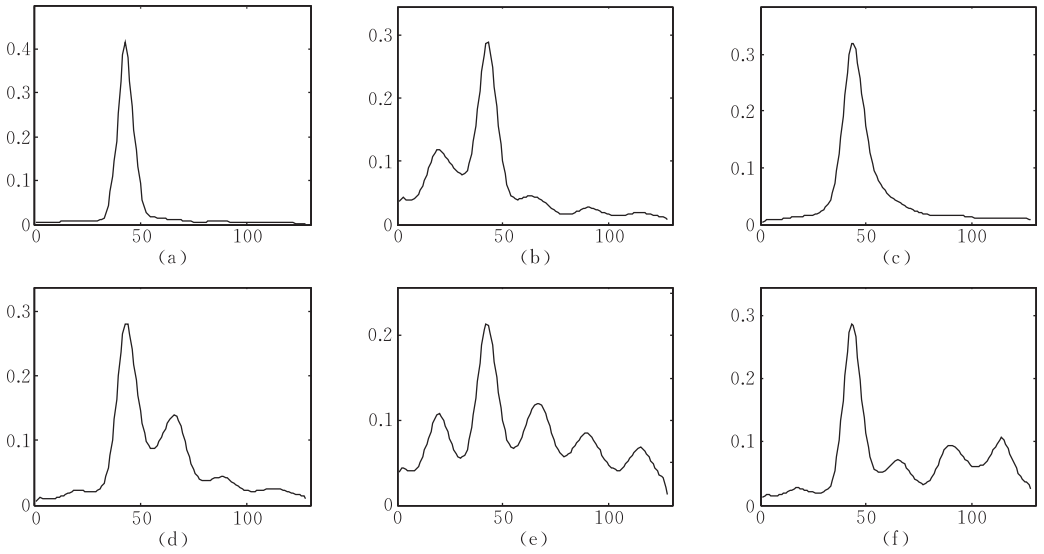


图 7　WKSC 算法在 Twhtag 数据集下的中心曲线

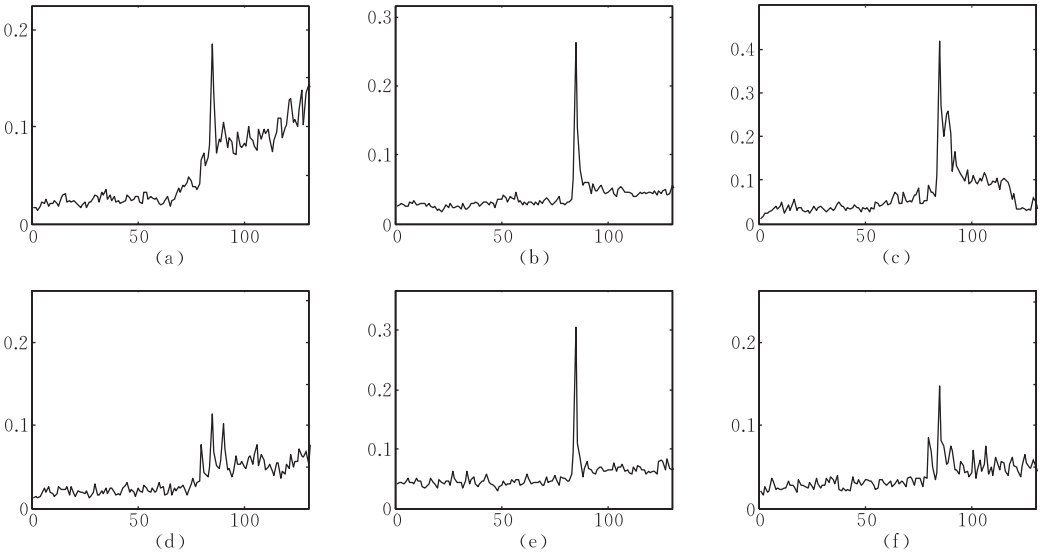


图 8　K_SC 算法在 ChinDt 数据集下的中心曲线

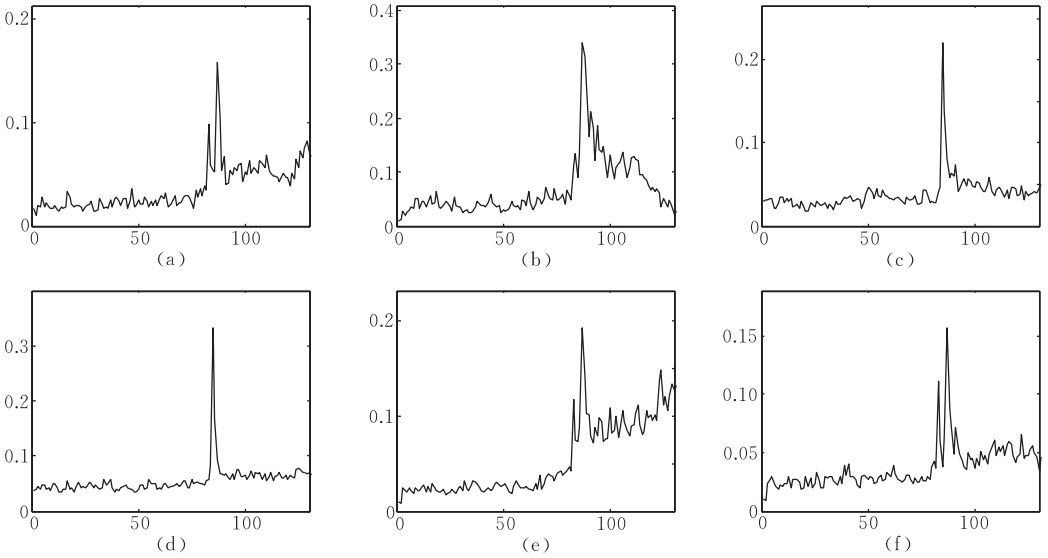


图 9　WKSC 算法在 ChinDt 数据集下的中心曲线

度层次进行停止，我们计算了 WKSC 算法在不同维度层次下所消耗的时间，如图 10 所示. 图 10 中 x 轴表示不同数据集的维度，y 轴表示运行时间，单位为 s.

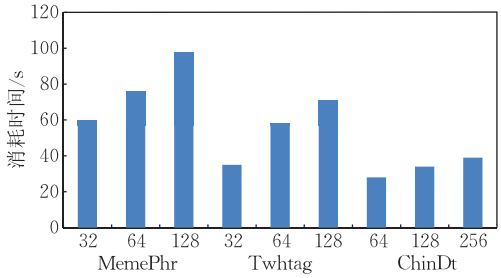


图 10 WKSC 算法在 3 个数据集、不同维度下的消耗时间

从图 10 上可以看出 WKSC 算法对于低维数据可以在很短的时间之内聚类完成，对于 MemePhr 和 Twhtag 数据集，维度为 64 时，WKSC 所消耗的时间基本为 K_SC 算法的 50%. 对于 ChinDt 数据集，维度为 256 时，也比 K_SC 算法减少 50%的时间. 另外，当维度增大时，WKSC 算法基本成线性比例增长. 这说明 WKSC 算法对于高维数据具有较好的处理能力，可以在大量的实际高维话题聚类中使用.

WKSC 算法在低维层次聚类时具有很高的效率. 那么，能否在低维层次就取得与 K_SC 算法相似的中心曲线，从而刻画话题趋势？我们选择 MemePhr 和 Twhtag 数据集在维度为 64、ChinDt 数据集在维度为 128 的情况下，算法得到的中心曲线，分别显示在图 11～图 13 上.

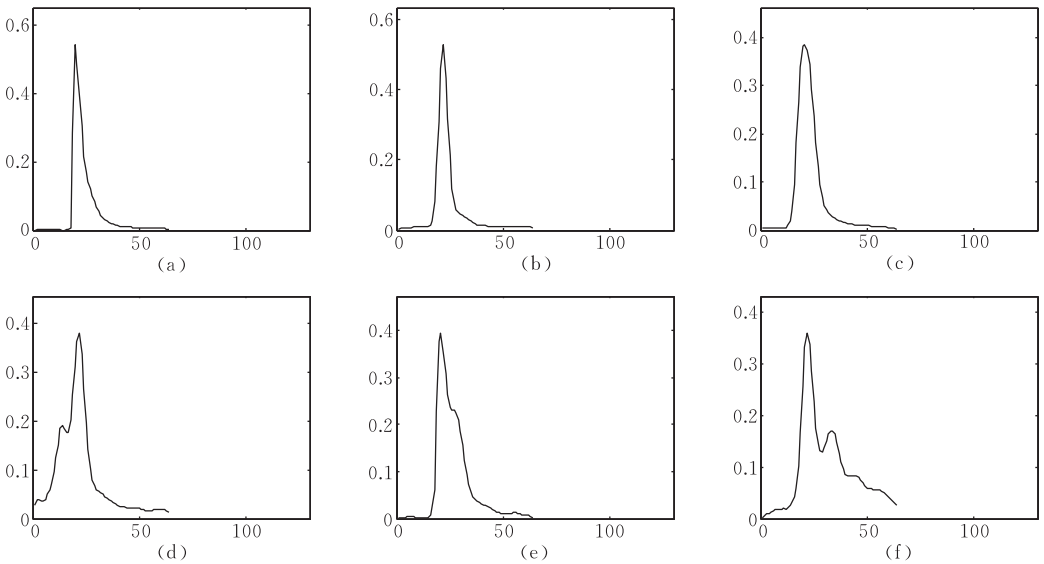


图 11 低维 MemePhr 的中心曲线

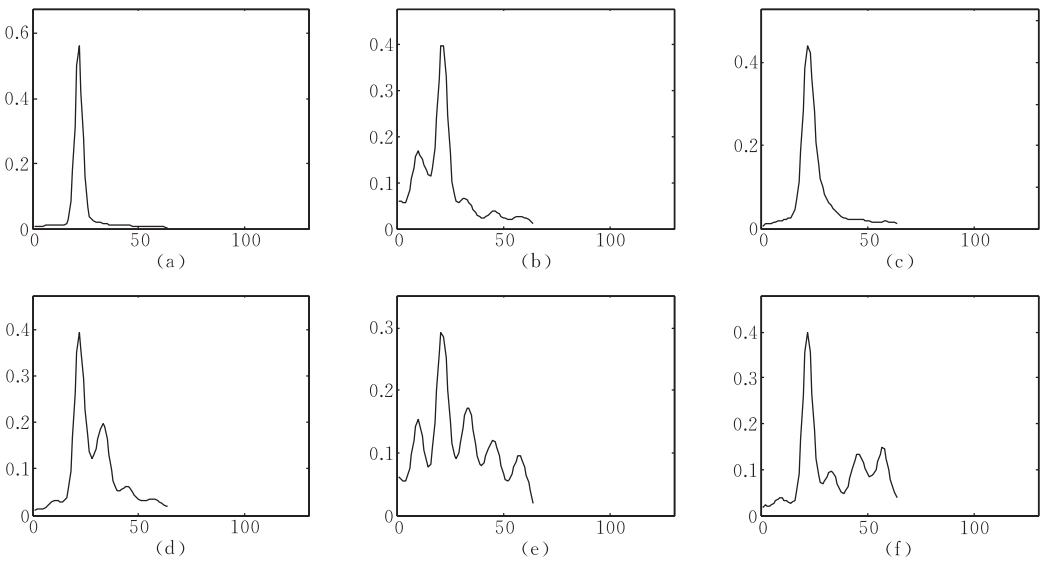


图 12 低维 Twhtag 的中心曲线

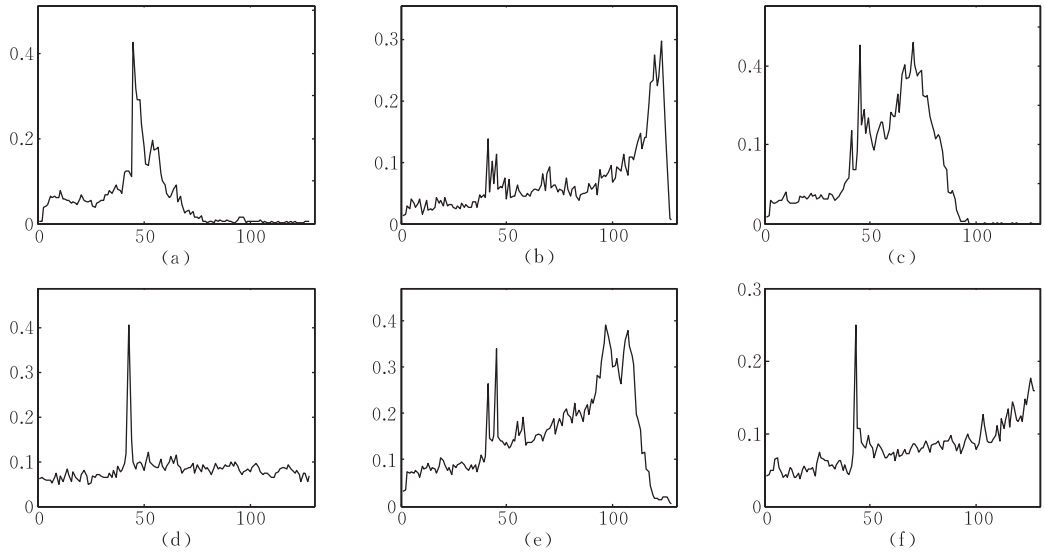


图 13　低维 ChinDt 的中心曲线

从图 11～图 13 可以看出，WKSC 算法在低维层次进行聚类时，得出的类别趋势与相同数据集高维层次下的结果一致，话题的发展趋势与波动均能清晰表达. 这表明 WKSC 算法在低维层次聚类能够满足分类话题、发现话题趋势的目标. 这个特性对于高维话题聚类具有重要意义，当话题的维度和数量很大时，可以用低维数据代替原始数据进行聚类，得出的中心曲线趋势能够刻画话题的发展趋势.

5.2.2　K_SC 算法和 WKSC 算法聚类效果分析和比较

作为一个聚类算法，不仅需要考虑到聚类的效率，还需要考虑类中的成员是否合理，也就是聚类结果的合理性问题. 我们用 2 个指标分别从类内和类间评价了两种算法聚类结果的合理性.

(1) F-Value(F 值). F-value 的计算方法如式(1). F-value 刻画了每个类内部成员的紧凑度，F 值越小表示聚类效果越好.

(2) D-Value(D 值). $D=\sum \hat{d}(\mu_i,\mu_j)^2$，其中 μ_i 是类 i 的矩阵中心，μ_j 是类 j 的矩阵中心. D-Value 代表类与类之间的差异性，类与类的矩阵中心 D 值越大表示聚类效果越好.

表 1 给出了 K_SC 和 WKSC 算法在 3 个数据集下最高维度层次聚类的 F 值和 D 值. 从表 1 中可以看出，无论类中成员的紧凑度还是类与类之间的差异度，WKSC 算法的值都优于 K-SC 算法. 这表明 WKSC 算法在降低时间复杂度的同时还改进了聚类效果，其原因在于 WKSC 算法改进了聚类的初始矩阵中心.

表 1　两个算法在不同数据集下的 F 值和 D 值

MemePhr 数据集			Twhtag 数据集			ChinDt 数据集		
聚类算法	F 值	D 值	聚类算法	F 值	D 值	聚类算法	F 值	D 值
K_SC	83.8	3.47	K_SC	64.9	3.68	K_SC	36.0	10.1
WKSC	72.6	3.94	WKSC	56.2	4.36	WKSC	28.2	10.6

如 5.2.1 节分析，WKSC 算法在低维层次聚类时，效率和中心曲线都具有较好的表现. 现在分析 WKSC 算法在不同维度层次的聚类效果，我们将 3 个数据集、不同维度层次下聚类算法的 F 值和 D 值列在表 2 中.

表 2　WKSC 算法在不同维度层次下的 F 值和 D 值

MemePhr 数据集			Twhtag 数据集			ChinDt 数据集		
维度	F 值	D 值	维度	F 值	D 值	维度	F 值	D 值
32	109.7	2.02	32	60.3	3.89	64	61.9	8.23
64	75.8	3.50	64	57.3	4.03	128	43.4	10.16
128	72.6	3.94	128	56.2	4.36	256	28.2	10.6

通过表 2 的结果，我们可以看出随着 Haar 小波反变换到越高的层次，F 值越小，说明各类成员的紧凑度越高；同样类与类之间的差异值越大，表示聚类划分的界限越清晰. 在最高维度(也就是原始的时间序列)时，WKSC 算法效果最好，但是时间复杂度也最高.

从表 2 的数据还能看出，WKSC 算法对 MemePhr 和 Twhtag 数据集在 64/128 维度之间、ChinDt 数据集在 128/256 维度之间的聚类差异并不显著，说明了 WKSC 算法在较低维度时已经取得了较好的聚类精度，却可以降低较多的运行时间. 这个实验结果同样表明了对于高维时间序列，WKSC 聚类算法可以通过降维来获得好的效果和效率.

6 结论与总结

分析和建模交互式网络上的热点话题是一个具有很大挑战性的研究问题,而对话题热度进行聚类则为建模提供了一个有效的手段,本文针对热点话题的热度时间序列聚类开展了一系列研究与实验.首先总结了已有的聚类算法以及目前的应用研究热点,分析了 K_SC 算法的优点以及高时间复杂度等特点,基于小波变换提出了 WKSC 算法.

我们在 3 个各具代表性的话题上进行了大量的实验,从不同角度分析和对比了 K-SC 算法和 WKSC 算法的性能.实验结果表明 WKSC 聚类算法可以有效降低聚类时间复杂度,平均意义下可以节省 50%的消耗时间.WKSC 算法可以满足高维大数据集的聚类需求,在实际使用时能取得很好的效果.

利用 WKSC 算法对互联网上的海量热点话题进行聚类,从而发现更科学合理的话题类型、应用话题的中心曲线进行建模分析都将是未来值得研究的问题.在大规模数据上应用 WKSC 算法时,如何根据话题数据特征,自动设定与调整 K 值,也是值得进一步研究的问题.

参 考 文 献

[1] Szabo G, Huberman B A. Predicting the popularity of online content. Communications of the ACM, 2010, 53(8): 80-88

[2] Kumar R, Novak J, Raghavan P et al. On the bursty evolution of blogspace//Proceedings of the World Wide Web Conference on Internet and Web Information Systems. Washington, USA, 2005: 159-178

[3] Mei Q, Liu C, Su H et al. A probabilistic approach to spatiotemporal theme pattern mining on weblogs//Proceedings of the 15th International Conference on World Wide Web. New York, USA, 2006: 533-542

[4] Crane R, Sornette D. Robust dynamic classes revealed by measuring the response function of a social system//Proceedings of the National Academy of Sciences of the United States of America. Washington, USA, 2008: 15649-15653

[5] Malmgren R D, Stouffer D B, Motter A E et al. A Poissonian explanation for heavy tails in email communication//Proceedings of the National Academy of Sciences of the United States of America. New York, USA, 2008: 18153-18158

[6] Barabási A L. The origin of bursts and heavy tails in human dynamics. Nature, 2005, 435(7039): 207-211

[7] Yang J, Leskovec J. Patterns of temporal variation in online media//Proceedings of the 4th ACM International Conference on Web Search and Data Mining. New York, USA, 2011: 177-186

[8] Chan F K P, Fu A W C, Yu C. Haar wavelets for efficient similarity search of time-series: With and without time warping. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(3): 686-705

[9] Li Bin, Tan Li-Xiang, Zhang Jing-Song. Time series symbolic methods facing data mining. Journal of Circuit and Systems, 2000, 5(2): 9-14(in Chinese)
(李斌, 谭立湘, 章劲松. 面向数据挖掘的时间序列符号化方法研究. 电路与系统学报, 2000, 5(2): 9-14)

[10] Li Ai-Guo, Qin Zheng. On-line segmentation of time-series data. Journal of Software, 2004, 15(11): 1671-1679(in Chinese)
(李爱国, 覃征. 在线分割时间序列数据. 软件学报, 2004, 15(11): 1671-1679)

[11] Keogh E, Kasetty S. On the need for time series data mining benchmarks: A survey and empirical demonstration. Data Mining and Knowledge Discovery, 2003, 7(4): 349-371

[12] Tewari G, Snyder J, Sander P V. Signal-speciallized parameterization for piecewise linear reconstruction//Proceedings of the Eurographics Symposium on Geometry Processing. New York, USA, 2004: 55-64

[13] Zhan Yan-Yan, Xu Rong-Cong, Chen Xiao-Yun. Time series piecewise linear representation based on slope extract edge point. Computer Science, 2006, 33(11): 139-142(in Chinese)
(詹艳艳, 徐荣聪, 陈晓云. 基于斜率提取边缘点的时间序列分段线性表示方法. 计算机科学, 2006, 33(11): 139-142)

[14] Leskovec J, McGlohon M, Faloutsons C, Glance N, Hurst M. Cascading behavior in large blog graphs//Proceedings of the 7th SIAM International Conference on Data Mining. Pittsburgh, USA, 2007: 551-556

[15] Leskovec J, Singh A, Kleinberg J. Patterns of influence in a recommendation network//Proceedings of the PAKDD on Knowledge Discovery and Data Mining. Berlin, Germany, 2006: 380-389

[16] Dai Dong-Bo, Tang Chun-Lei, Xiong Yun. Sequence clustering algorithms based on global and local similarity. Journal of Software, 2010, 21(4): 702-717(in Chinese)
(戴东波, 汤春蕾, 熊赟. 基于整体和局部相似性的序列聚类方法. 软件学报, 2010, 21(4): 702-717)

[17] Yang Yi-Ming, Pan Rong, Pan Jia-Lin, Yang Qiang, Li Lei. A comparative study on time series classification. Chinese Journal of Computers, 2007, 30(8): 1259-1266(in Chinese)
(杨一鸣, 潘嵘, 潘嘉林, 杨强, 李磊. 时间序列分类问题的算法比较. 计算机学报, 2007, 30(8): 1259-1266)

HAN Zhong-Ming, born in 1972, Ph. D., associate professor. His research interests include Web data analysis and mining, massive data analysis and mining, etc.

CHEN Ni, born in 1987, M. S. candidate. Her research interests include Internet analysis and mining.

LE Jia-Jin, born in 1951, professor, Ph. D. supervisor. His research interests include data warehouse and data mining, etc.

DUAN Da-Gao, born in 1976, Ph. D., associate professor. His research interests include multimedia information retrieval and data mining, etc.

SUN Jian-Zhi, born in 1967, associate professor. His research interests include complex network analysis and data mining, etc.

Background

Hot topics on interactive web such as BBS, Blogger and Microblogging system enormously affect not only the development of the various events in virtual world, but also people's perspectives and judgments in real world. It may also affect the attitude and verdict of government and judicial authorities towards those issues. Modeling and predicting development of hot topics is very important, but it is very difficult to model hot topics because different hot topics share different trend patterns. Therefore, clustering hot topics is very helpful to find essential patterns inside hot topics.

Quantity of participance, including post quantity, forward quantity, and comment quantity, is one of the dominant measurements for popularity of online topics, which change over time and present as a time-series. Nowadays, there are two types of clustering methods for hot topics. Contents-based clustering methods classify topics according to content similarity. These methods can be used to detect topics but are poor to depict trend pattern of topics. Time-series based clustering methods take a metric such as quantity of participance as a time-series to classify topics. Leskovec and his colleagues developed a K-Spectral Centroid (K-SC) time series clustering algorithm. It was testified by a large number of experiments that K-SC showed high efficiency in clustering online topics, but its shortcoming of being highly sensitive to the initialization of cluster centers. Moreover, K-SC algorithm has high time complexity and prevents its application on high dimension data.

In this paper, we improved K-SC algorithm by combining Haar wavelets transform and K-SC algorithm. A new iteration Wavelet-Based K-SC algorithm (WKSC for short) is proposed. In WKSC algorithm, the original time series are compressed by Haar wavelets transform to lower the dimensions of original time series. Then, we classify topics based on the lower dimensions time series using K_SC algorithm and the clustering results are used as the initial assignment at the high level clustering process.

Comprehensive experiments are conducted on three representative datasets. We analyzed and compared the performance of K-SC and WKSC from different respects. Experimental results show that WKSC clustering algorithm can significantly reduce clustering time complexity and save 50% cost of time averagely, which means that WKSC algorithm can be used on massive and high dimension hot topics. WKSC algorithm also can improve the quality of clustering result.

This work is supported by the National Natural Science Foundation of China under Grant No. 61170112 and Funding Project for Innovation on Science, Technology and Graduate Education in Institutions of Higher Learning under the Jurisdiction of Beijing Municipality (PXM2012_014213_000037).