社交网络的传播测量与时间序列聚类分析

周雪峰[1],徐 恪[1],张蓝珊[2],张 赛[1]

[1](清华大学 计算机科学与技术系,北京 100084)
[2](北京邮电大学 数字媒体与设计艺术学院,北京 100876)
E-mail: zhouxf11@mails.tsinghua.edu.cn

摘 要: 对社交网络的信息传播进行时间序列聚类是研究其规律非常有效的方法. 目前,相关的工作特别是针对国内社交网络的时间序列聚类研究,还不够深入. 对时间序列聚类算法 K-SC 算法进行了针对性的改进,提出的 T-SC 算法借鉴了凝聚层次聚类的思想解决了聚类个数设置的难题. 对人人网、腾讯微博和百度贴吧三个国内非常有代表性的社交网络进行了大量的测量和分析工作,并运用 T-SC 算法对测量数据进行了聚类分析. 研究发现了不同社交网络典型而又互不相同的传播模式: 人人网的视频分享呈现明显的周期性,每个周期内的分享传播存在一个主流的模式,该模式与一天之中不同时段人人网的在线人数变化趋势非常相近; 腾讯微博的转发传播呈现爆发性,绝大多数的转发出现在微博发出后的 48 小时之内,其主流的传播模式是微博发出后大量传播并迅速消失; 百度贴吧帖子的生命期很长,但是没有一个占主导地位的传播模式. 本文创新性的将聚类分析的结果应用于信息传播的预测,根据已知的传播时间序列,得到未来信息传播行为在聚类层面的预测,为解决传播预测的难题提供了新的思路.

关 键 词: 社交网络; 测量; 时间序列; 聚类分析; 传播预测
中图分类号: TP393 **文献标识码:** A **文 章 编 号:** 1000-1220(2015)07-1545-08

Propagation Measurement and Cluster Analysis of Time Series in Social Networks

ZHOU Xue-feng[1], XU Ke[1], ZHANG Lan-shan[2], ZHANG Sai[1]
[1](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)
[2](School of Digital Media & Design Arts, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: Cluster analysis of time series is a very effective method to study the law of information spreading in the social networks field. It is very significative to proceed a comprehensive study on domestic social networks by means of cluster analysis of time series, but there is still less work. Firstly, this article improves the K-SC algorithm, which is used to proceed the cluster analysis of time series. Learning from the thinking of hierarchical agglomerative clustering algorithm, we propose the T-SC algorithm, which avoids the difficulty to set the number of clusters. Then, we measure and analyze three sites: Renren、Tencent Weibo and Baidu Post Bar, all of which are very representative social network sites in China, and finish the cluster analysis on the measurement data of them by utilizing the T-SC algorithm. We arrive at typical and distinct conclusions about information spreading in these three sites: video sharing of Renren shows obviously cyclical and in each cycle there is a mainstream mode of transmission, which is very similar to the trend of number of online people at different times in a day; Tencent Weibo shows explosively spread and the vast majority of forwarding appears in the 48 hours just after being published. The mainstream mode of transmission is spreading massively after being published and then disappearing rapidly. Posts of Baidu Post Bar have a very long age and there is no one dominant mode of transmission. At last, we provide a new method to challenge the difficult task of propagation prediction of social networks, by innovatively applying the result of cluster analysis. Just based on the known time serie, we can get prediction of future spreading on the level of clustering.

Key words: social networks; measurement; time series; cluster analysis; propagation prediction

1 引 言

社交网络本质上是一个在互联网上人与人相互连接构成的网络,在这个网络上用户可以发布信息、获取其他用户发布的信息,典型的代表有 Facebook、Twitter、人人网、新浪微博等. 经历了最近 10 年的飞速发展,社交网络极大的改变了我们的生活,已经成为人们不可或缺的娱乐方式和交流渠道. Facebook 的月活跃用户数已经达到 12 亿,每天上传的新照片数达到 3.5 亿张,成长为访问量比肩 Google、市值上千亿的互联网巨头; 在国内,截至 2012 年底,新浪微博的注册用户已经

收稿日期: 2014-04-13 收修改稿日期: 2014-06-12 基金项目: 国家科技重大专项项目(2012ZX03005001)资助; 国家自然科学基金项目(61170292、61161140454)资助; 国家"八六三"高技术研究发展计划项目(2013AA013302)资助; 国家"九七三"重点基础研究发展计划项目(2012CB315803)资助; 清华信息科学与技术国家实验室(筹)学科交叉基金资助. 作者简介: 周雪峰,男,1987 年生,硕士研究生,主要研究方向为在线社交网络; 徐 恪,男,1974 年生,博士,教授,博士生导师,主要研究方向为新一代互联网体系结构、高性能路由器、P2P 与应用层网络、Overlay 网络、物联网; 张蓝姗,女,1981 年生,博士,讲师,主要研究方向为在线社交网络、影视艺术、电视节目综合评估体系; 张 赛,男,1988 年生,博士研究生,主要研究方向为在线社交网络.

超过 5 亿,日活跃用户达到 4600 万.

伴随而来的是学术界对社交网络研究的愈发重视. 不止是计算机科学领域,作为真实社会网络在互联网上的虚拟映射,社交网络用户数目巨大、数据获取容易的特点使它成为社会学研究的对象;作为一种典型的复杂网络,网络科学、物理学领域对社交网络研究也颇多涉及.

信息传播是社交网络研究的重点. 相较于传统的 web 以信息为中心的传播,社交网络上的信息传播以人为中心,人与人之间的连接关系对信息的传播影响巨大,因此二者存在根本的不同. 而且,不同类型的社交网络具有不同的连接类型,Facebook、人人网用户之间的连接是双向的较强连接,Twitter、新浪微博用户之间则是单向关注的较弱连接. 那么,各类社交网络具有怎样的信息传播特性? 不同社交网络的信息传播是否存在共性或者区别其他的特殊规律? 解答这些问题具有重要的理论和应用价值,也是本文的主要目的.

社交网络的信息传播具有典型的时序特征,可以通过时间序列得到很好的描述. 我们通过多种数据获取手段对人人网、腾讯微博、百度贴吧三个不同类型的社交网络进行了测量,收集、分析了大量的时间序列数据. 应用聚类分析[1] 这一数据挖掘中常用的方法,找出了时间序列集合中的簇,并试图挖掘这些簇背后隐藏的传播意义.

信息传播研究的一个理想目标是能够精确地预测一条信息未来的传播趋势,但由于其复杂性这个目标并不容易实现. 时间序列聚类分析的结果一定程度上揭示了社交网络信息传播的典型模式,这启发我们:能否利用聚类分析的结果在聚类的层面进行传播预测? 本文对这一思路也进行了尝试.

本文的主要贡献有:

1. 提出 T-SC 算法,借鉴凝聚层次聚类的思想解决了聚类个数设置的难题.

2. 对人人网视频分享、腾讯微博转发和百度贴吧回复三类社交网络中的信息传播进行了测量和分析;对传播的时间序列进行了聚类和分析.

3. 给出了将聚类分析结果应用于信息传播预测的思路和算法.

本文的组织结构如下:第 2 节是相关工作的介绍;第 3 节介绍文中使用的聚类算法;第 4 节是对三个社交网络的传播测量和时间序列聚类分析;在第 5 节利用聚类分析的结果进行传播预测;第 6 节是对本文工作的总结和展望.

2 相关工作

社交网络的研究主题非常丰富. Xu 等人[2] 从测量的角度对社交网络拓扑结构、用户行为、网络演化、信息传播等多个方向的研究工作进行了调研和综述. 此外,社团发现、个性化推荐、隐私安全等也是社交网络研究的热点. 限于篇幅,这部分我们重点在信息传播及其预测、聚类分析两个与本文内容有关的主题展开介绍.

2.1 信息传播及其预测

作为复杂网络的一种,社交网络的信息传播研究能够很好地借鉴已有成熟工作的复杂网络传播机制,比如已经较为完备的传染病模型[3]. 经典的传染病模型有 SI 模型、SIR 模型和 SIS 模型. Gruhl 等人[4] 基于关键词研究了博客空间里的信息传播,他们正是利用了传统的传染病模型刻画博客空间里的信息传播;Adar 和 Adamic[5] 进一步扩展了传染病模型在社交网络信息传播领域的应用.

社交网络兴起后数据逐渐丰富,利用测量手段、数据驱动的信息传播研究大量出现. Adam[6] 利用 Facebook 的数据研究了情绪在社交网络上的传播;Wu 等人[7] 对 Twitter 上信息的产生、流动和消亡进行了研究,发现 Twitter 信息高度集中于少量精英用户,且用户呈现明显的类别同质性;Bakshy 等人[8] 研究了 Second Life 社交网络里的信息传播,发现社交网络对内容的接受起着重要作用;Zhang 等人[9] 的文章对新浪微博的信息传播进行了测量和分析,发现新浪微博存在很强的“名人效应”. 另外,影响力最大化[10]、节点传播影响力分析[11] 也是社交网络信息传播研究的重要方向.

信息传播研究的一个重要目标是实现对传播的预测. Cha 等人[12] 发现 YouTube 视频早期浏览量和晚期浏览量之间是线性相关的;Szabo 等人[13] 进一步的提出了 3 种模型来预测信息的流行度;Li 等人发表于 CIKM 的文章[14] 基于视频的固有吸引力,结合传播结构的影响,提出了新的视频流行度预测方法 SoVP,取得了很好的预测效果.

2.2 聚类分析

聚类是指在数据集中查找以某种方式集合在一起的数据点集合的过程. 目前已经存在大量的聚类算法[15],有基于划分、层次和密度等多个分类. 当前聚类分析研究的主要关注点包括聚类方法的可伸缩性、高维聚类分析技术以及大数据库中混合数据和分类数据的聚类方法等[1].

2.2.1 k-means 聚类算法及其改进

k-means 算法因其简单高效的特点成为学术界应用最多、最为经典的聚类分析算法,研究人员对 k-means 算法本身进行了很多的改进工作[16-18],也经常按照自己特定的需求对它进行针对性的改进. 例如,Zhou 等人[19] 的文章对 k-means 算法聚类个数设置的问题进行了研究,提出了一种新的确定最佳聚类个数的方法;文章[20] 提出 k-means + + 算法,重点对初始聚类中心的选择方法做出了改进;本文介绍的 K-SC 算法、T-SC 算法也都源于在社交网络时间序列的特定应用场景下对 k-means 算法的改进.

2.2.2 时间序列聚类分析

时间序列作为一种数据形式,广泛存在于商业、医学、工程和社会科学等大型数据库中,形成了规模庞大的时间序列数据库. 与传统的点数据聚类相比,时间序列因其巨大的维数特征导致聚类方法更加复杂.

文章[21] 对时间序列聚类的相关研究进行了综述. 而在社交网络信息传播的特定场景下,需要我们对传统的时间序列聚类算法进行针对性的改进. Yang 等人[22] 提出了对 Twitter 等社交网络信息传播的时间序列进行聚类的 K-SC 算法,算法的重点是针对社交网络特定场景定义了时间序列之间的距离(即相似性度量);Han 等人[23] 针对 K-SC 算法高时间复杂性等问题,结合小波变换技术提出了 WKSC 算法,显著减低了时间复杂度,同时改进了聚类效果.

3　聚类方法

3.1　问题建模

给定 N 个信息,对每个信息 i 我们有一个长度为 L 的踪迹集合$(s_j,t_j)_i$,其中 $i=1\cdots N,j=1\cdots L$,表示用户截至 s_j 时刻共 t_j 次提到(即分享、转发或回复)了信息 i. 对于这 N 个踪迹集合,我们通过计算在单位时间 t 内提到信息 i 的次数从而构造时间序列 $x_i(t)$,即构造在时间 t 内提到信息 i 的次数的序列.

显然 $x_i(t)$ 描述了信息 i 传播的速率,代表了 i 的流行度随时间的变化情况,如图 1.

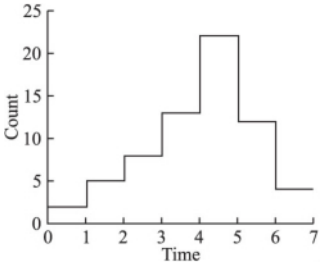

图 1　时间序列 2 5 8 13 22 12 4,$t=1$,$L=7$

Fig. 1　Time serie 2 5 8 13 22 12 4,$t=1$,$L=7$

3.2　K-SC 算法介绍

K-SC 算法中时间序列 x 和 y 的距离 $\hat{d}(x,y)$ 定义为

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

其中,$y_{(q)}$ 是把时间序列 y 平移 q 个时间单位后的结果,$\|\cdot\|$ 是 l_2 范数. 式中找出了使 x 和 y 相匹配的最优平移 q 和缩放系数 α. K-SC 算法的具体描述见算法 1.

K-SC 算法一次迭代的复杂度为 $O(\max(NL^2,KL^3))$,L^3 的复杂度还是很大的. 与 k-means 算法一样,K-SC 算法需要设置聚类的个数 K 和初始聚类分配,其聚类结果对 K 较敏感,而且初始聚类分配的选取会影响 K-SC 算法的收敛速度.

3.3　K-SC 算法的聚类个数问题

在对数据进行聚类分析之前通常并不知道最终的聚类个数,这也是应用 k-means 算法及其衍生算法存在的主要问题,它们都需要在算法的一开始设置聚类的个数. 对此研究人员提出了多种解决方案,其中最为常见的是通过描述聚类有效性的函数指标(如 Calinski-Harabasz[24]、Silhouette[25] 等)来对一个区间内的聚类个数进行遍历,选择使相关的函数指标最优的聚类个数. K-SC 算法从 k-means 算法遗传了设置初始聚类个数的难题,其解决方案是通过综合考察 Hartigan 和 Silhouette 两个指标来确定最佳的聚类个数. 这种综合考察函数指标的方案一般都需要人的参与来主观确定特定聚类个数下的聚类效果,难以做到无人值守; 另一方面,一个区间内的多次聚类、每次聚类计算多个函数指标不可避免的会带来较大的计算开销.

算法 1. K-SC 算法 $K\text{-}SC(x,K,C)$.

输入: 时间序列 $x=\{x_1,\cdots,x_N\}$,聚类个数 K,初始聚类分配 $C=\{C_1,\cdots,C_K\}$

WHILE

　$\hat{C}\leftarrow C$

　FOR $j=1$ TO K DO

　　$M\leftarrow\sum_{i\in C_j}\left(I-\frac{x_i x_i^T}{\|x_i\|^2}\right)$

　　$\mu_j\leftarrow M$ 最小的特征向量

　　$C_j\leftarrow\varnothing$

　ENDFOR

　FOR $i=1$ TO N DO

　　$j^*\leftarrow\mathrm{argmin}_{j=1,\cdots,K}\hat{d}(x_i,\mu_j)$

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

　ENDFOR

UNTIL $\hat{C}=C$

输出: 聚类中心 $\mu=\{\mu_1,\cdots,\mu_K\}$,聚类分配结果 $C=\{C_1,\cdots,C_K\}$

3.4　T-SC 算法

基于以上分析,我们设计了 T-SC 算法,为解决聚类个数设置的问题提供了一个新的思路. 该算法借鉴层次聚类中自底向上的凝聚思想,首先设定一个聚类个数的上限,之后迭代的进行聚类,将聚类结果中聚类中心过于相近的类合并为一类,直到聚类中心之间的距离均超过一定的阈值后结束凝聚,得到最终的聚类结果. 算法 2 是 T-SC 算法的具体描述,其算法复杂度 T 倍于 K-SC 算法.

算法 2. T-SC 算法 $T-SC(x,T,C)$.

输入: 时间序列 $x=\{x_1,\cdots,x_N\}$,聚类个数上限 T,初始聚类分配 $C=\{C_1,\cdots,C_T\}$

$\mu',C'\leftarrow K-SC(x,T,C)$

$D\leftarrow\mu'$ 两两之间距离构成的方阵

$\alpha\leftarrow 0.5$

FOR $i=1$ TO T DO

　FOR $j=i+1$ TO T DO

　　IF $D(i,j)/mean(D)<\alpha$ DO

　　$C'\leftarrow$舍弃 μ'_j 并调整后的聚类分配

　　$T\leftarrow T-1$

　　$\mu',C'\leftarrow T-SC(x,T,C'')$

　　RETURN

　　ENDIF

　ENDFOR

ENDFOR

输出: 聚类个数 T,聚类中心 $\mu'=\{\mu'_1,\cdots,\mu'_T\}$,聚类分配结果 $C'=(C'_1,\cdots,C'_T)$

Yang 等人使用 K-SC 算法对数据集 Memetracker phrases 进行了聚类分析,本文同样选择这一组数据利用 T-SC 算法进行聚类分析,以对比观察 T-SC 算法的聚类效果.

聚类结果如下页图 2 所示,其中(a)、(b)是 K-SC 算法的聚类结果和 Silhouette 指标;(c)、(d)是 T-SC 算法的聚类结果和 Silhouette 指标. (c)中 T-SC 算法聚类得到了 5 个簇,这 5 个簇的聚类中心与(a)中 K-SC 算法的聚类中心有非常一致的对应关系,不同的是将(a)中明显接近的 3、6 合并为(c)中的 1,得到了更合理的聚类结果. Silhouette 指标方面,T-SC 算法的均值 0.2139 也略优于 K-SC 算法的 0.2010.

实验表明,在对算法性能要求不高的场景下,T-SC 算法无需设置初始的聚类个数,且能得到与 K-SC 算法相当甚至

更优的聚类结果.

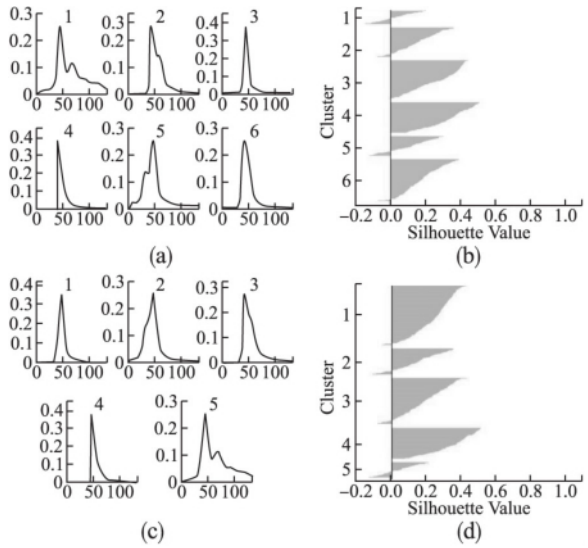

图 2 Memetracker phrases 的聚类结果

Fig. 2 Clustering results of Memetracker phrases

4 聚类分析

本节应用 T-SC 算法对测量得到的真实社交网络数据进行聚类分析. 研究对象选取的是国内三个非常典型又存在巨大差异的社交网站: 人人网、腾讯微博和百度贴吧. 人人网基于同学、同事关系构建关系网络,是现实世界社会关系的互联网映射; 腾讯微博则不局限于认识的好友,其用户多关注明星、名人等陌生人, "名人效应"是其典型特征; 百度贴吧的用户围绕某一主题展开讨论,以话题为中心,用户之间的好友关系较为薄弱. 数据的多样性有利于研究分析各类社交网络信息传播规律的异同,得出更加全面的结论. 具体的数据集描述见表 1.

表 1 数据集

Table 1 Data sets

	序列总数	选取个数	选取规则	获取方式
人人	335283	1000	选取分享最多的视频	官方日志
腾讯微博	1265848	1000	选取转发最多的微博	开放平台编程
百度贴吧	2807618	1000	选取回复较多的帖子	网络爬虫

针对不同社交网络的特点,本文运用了多种不同的数据收集方法,包括官方日志、开放平台 API 编程和网络爬虫等. 特别的,网络爬虫需要长时间、海量的互联网访问,不可避免的会出现网络超时、网页数据错误等异常导致爬虫退出. 为此我们为爬虫的异常处理模块添加了向指定邮箱地址发送邮件的功能,通过智能终端接收邮件即可立即发现爬虫异常的发生,不需要人工值守、检查爬虫的运行情况.

4.1 人人网

4.1.1 数据及预处理

人人网是一个类似于 Facebook 的基于真实个人信息和现实好友关系的社交网站,拥有 2 亿活跃用户. 人人网的用户可以分享自己原创或好友分享的日志、视频、图片等内容. 我们主要关注视频分享在人人网中传播的时间序列. 因为获得了人人网的官方支持,所以我们获取了人人网的部分日志数据. 这些数据无法通过开放平台 API 编程或者网络爬虫的方法获取,是研究人人网较为原始且全面准确的资料,有很高的研究价值.

表 2 分享/转发/回复次数

Table 2 Number of sharing/forwarding/replying

	最小值	最大值	平均值	中位值
人人网分享	1	154955	22.9	1
腾讯微博转发	0	22593	4.2	0
百度贴吧回复	0	1581713	64.5	7

人人网用户的视频分享行为会以日志的形式记录下来,经过整理共得到了 335283 个视频分享的时间序列,统计数据见表 2. 其中,分享最多的视频被分享了 154955 次,仅被分享 1 次(即只有发布者的分享)的视频占总数的 57.4%,分享 10 次以下的视频占总数的 90.1%,14.9% 的视频占了 94.0% 的分享次数. 可见绝大多数视频的分享次数很少,而少量视频占了绝大多数的分享次数.

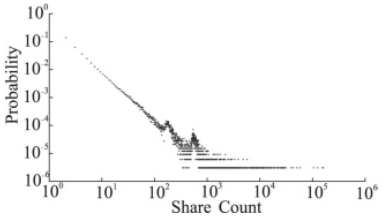

图 3 分享次数的概率分布

Fig. 3 Probability distribution of sharing number

在图 3 的对数坐标下观察视频分享次数的概率分布,分享次数较小(小于 150)时近似为一条直线,是非常明显的幂律分布. 我们选取前 1000 个分享次数最多的时间序列作为待聚类集合,其分享次数不小于 861.

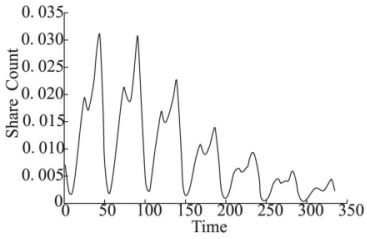

图 4 一条典型的视频分享时间序列曲线

Fig. 4 A typical curve of time serie of video sharing

把上述 1000 个时刻序列转换为时间-速度序列(以下简称时间序列),如图 4 所示,横坐标为时间 t,纵坐标为时段 $[t-0.5h,t)$ 内视频的分享次数 v_t. 为了减小曲线的震荡,我们使用了高斯平滑滤波器对时间序列进行了平滑处理. 图 4 中我们看到有很多呈周期性分布的尖峰. 当我们关注其最高峰附近传播的规律时,其他尖峰会带入噪声和干扰. 因此这里考虑较长时间的时间序列是不合理的,需要截断长序列以关

注最活跃的部分.

为了截断时间序列,我们首先测量最高峰的时间跨度. 令 T_p 为时间序列达到最高峰 v_p 的时刻. 对于阈值 xv_p ($0 < x < 1$) ,记 $T_1(x)$ 为 T_p 之前的最后一个速度小于 xv_p 的时刻,$T_2(x)$ 为 T_p 之后的第一个速度小于 xv_p 的时刻. 图 5 显示了 x 取不同值时所有时间序列 $T_p - T_1(x)$ 、$T_2(x) - T_p$ 和 $T_2(x) - T_1(x)$ 的中值.

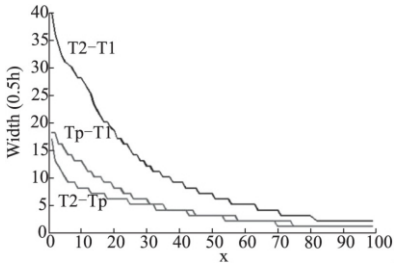

图 5　不同阈值下的时间跨度

Fig. 5　Time span under different thresholds

从图 5 可以看出,大多数的视频分享时间序列维持在峰值 v_p 某一阈值范围内的持续时间很短. 例如,序列从不到最高峰的 1% 上升到最高峰,继而下降到不到最高峰的 1% 只需 20 个小时,这一点也反映了视频分享时间序列的动态变化以天为周期的特征,与图 4 的直观观察相符. 另外,当 x 较小时,$T_p - T_1(x)$ 和 $T_2(x) - T_p$ 曲线的差异较大,并且 $T_2(x) - T_p < T_p - T_1(x)$,说明一般情况下,时间序列较缓慢地达到其最高峰,然后快速下降.

根据上面的分析,我们以 $L = 48$(即 24 小时) 为长度,按照将最高峰 v_p 定位到每个截断时间序列的 2/3 位置(即 $t = 32$) 的方法来截断原始时间序列.

4.1.2　聚类结果

对上述时间序列运用 T-SC 算法,得到如图 6 所示的聚类结果及其 Silhouette 指标图,各个簇所占个数依次为 153、85、762. 可以看到,人人网视频分享在一个周期内存在一个主

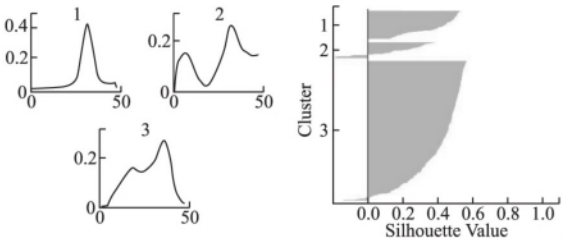

图 6　T-SC 算法对人人网视频分享时间序列聚类的结果

Fig. 6　Result of T-SC

流的传播模式(模式 3) ,符合该模式的时间序列占到总数的 76.2%. 该模式的特点是一个小高峰后略有下降,然后一个大的高峰后迅速下降,这与一天之中不同时段人人网活跃用户数的变化有关. 本课题组在人人网的教育网出口处进行抓包实验,对一天中不同时段活跃 IP 数进行了统计,其归一化之后的变化曲线见图 7,与模式 3 非常一致.

4.2　腾讯微博

4.2.1　数据及预处理

腾讯微博是一个提供微型博客服务的类 Twitter 网站,是中国最大的微博网站之一,目前已拥有 5 亿用户. 腾讯微博用户可以转发或评论其他用户的微博,使其得以传播. 我们使用腾讯微博的开放平台API获取腾讯微博的转发数据. 开放平

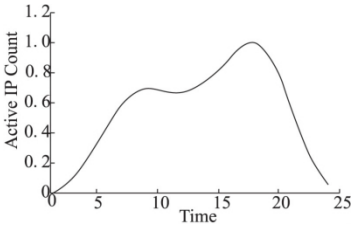

图 7　活跃 IP 数

Fig. 7　Active IP count

台 API 获取的数据相比网络爬虫抓取的数据更加简单、规范,数据获取的效率更高. 具体来说,我们使用接口 public_timeline 获取当前最新发表的微博,该接口能够获取广播大厅的最新微博,通过这种方式我们得到了 1265848 条微博的 id 集合; 使用接口 re_count 获取每条微博的转发数,排序后选取被转发最多的 1000 条微博,其转发次数不小于 500 次; 最后使用接口 re_list 获取这些微博中每一条微博的转发时刻序列,以 $0.5h$ 为单位制作时间序列.

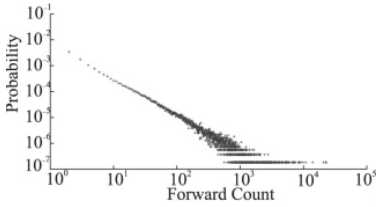

图 8　转发次数的概率分布

Fig. 8　Probability distribution of forwarding number

1265848 条微博中转发最多的微博被转发了 22593 次,未被转发的微博占了总数的 88.5% ,97% 的微博转发次数不超过 10 次,5% 的微博占了 98% 的转发次数,可见绝大多数微博未被转发或只被转发了几次,而少数微博占了绝大多数的转发次数. 其转发次数概率分布如图 8,呈近似的幂律分布,转发统计数据见上页表 2.

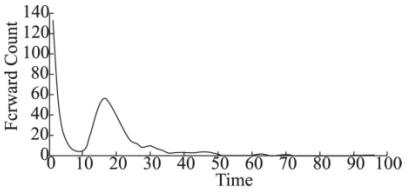

图 9　一条典型的微博转发时间序列曲线

Fig. 9　A typical curve of time serie of retweeting

微博类社交网络的特点是信息发出后的传播很快,并快速衰减. 图 9 展示的是一条典型的腾讯微博时间序列曲线,横轴是时间 t,纵坐标是时段 $[t - 0.5h, t)$ 内微博的转发次数 v_t. 可以看见,微博在 48 小时后明显趋于消失.

定义 $\varphi = Number_{48}/Number_{all}$,即 48 小时内的转发数占总转发数的比例. 下页图 10 是 φ 的累积分布,选取的 1000 条微

博中 90% 的微博 48 小时内的转发次数占据了该微博总转发次数 95% 以上的比例. 针对微博的这一特性,我们重点关注一条微博在发出之后 48 小时内的传播特性. 以 0.5h 为单位,得到 96 个单位长度上的时间序列.

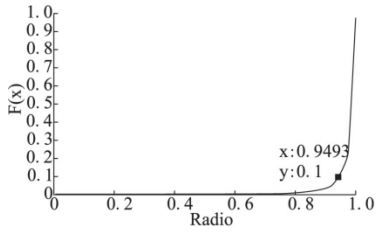

图 10 φ 的累积分布

Fig. 10 Cumulative distribution of φ

4.2.2 聚类结果

对上述腾讯微博转发时间序列运用 T-SC 算法,我们得到如图 11 所示的聚类结果及其 Silhouette 指标图,各个簇所占个数依次为 360、81、19、440、23、77. 六个簇中,簇 1 和簇 4 占绝大多数,而且二者有着几乎相同的传播模式: 在微博发出后很短的时间内爆发,然后迅速衰减,可见这是腾讯微博主流的传播模式; 只有少数微博(簇 2、3、5) 是发出一段时间之后达到转发的高峰; 另外一小部分微博(簇 6) 会在快速衰减之后有一个小的反弹.

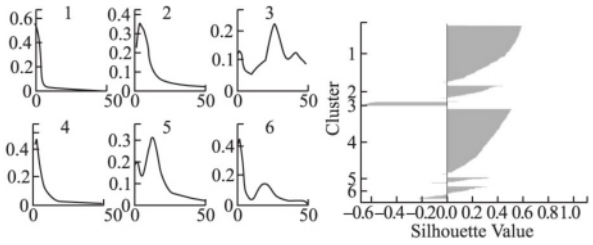

图 11 T-SC 算法对腾讯微博转发时间序列进行聚类的结果

Fig. 11 Result of T-SC

4.3 百度贴吧

4.3.1 数据及预处理

百度贴吧是全球最大的中文社区,是一种基于关键词的主题交流社区. 它与搜索紧密结合,让那些对同一个话题感兴趣的人们聚集在一起,方便地展开交流和互相帮助. 百度贴吧在国内有着巨大的用户基数和影响力,遗憾的是目前学术界对百度贴吧的研究还很少涉及.

百度贴吧没有开放平台 API 可供调用,所以我们通过 python 编写的网络爬虫来获取贴吧帖子的回复数据,共获取了 2807618 个帖子的回复数据,从中筛选了回复数较多的 1000 个帖子,回复次数在 946 次以上. 我们并没有选择回复数最多的帖子,原因是这些帖子的回复数多在一万次以上,部分帖子的回复数甚至达到了百万量级. 网络爬虫方式的低效难以处理如此多的回复数据,且太多次的访问也会引起贴吧官方检查出异常访问并禁止扫描 IP 的贴吧访问.

在 2807618 个帖子中,回复最多的帖子被回复了 1581713 次,未被回复的帖子占总数的 16.8%,这个数据明显低于人人网和腾讯微博,说明大多数的帖子都有回复,贴吧帖子更容易得到吧友的回应; 回复 10 次以下的帖子占总数的 57.0%,5% 的帖子占了 70% 的回复次数. 其回复次数概率分布如图 12,呈近似的幂律分布,统计数据见上页表 2.

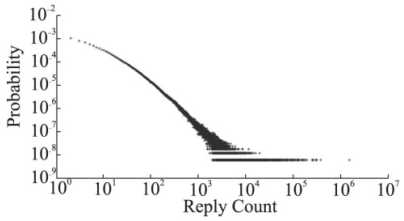

图 12 回复次数概率分布

Fig. 12 Probability distribution of replying number

百度贴吧的信息传播与人人网、腾讯微博有很大的不同,其最大的特点是很多帖子生命期很长,且能保持相当的活跃度. 图 13 是所选取的 1000 条帖子截止到数据获取时的生命长度的统计,单位是天. 图中,近 20% 的帖子的生命期达到了 500 天以上. 生命期后 1/3 时间段内的回复数占总回复数比例的均值也达到了 13%,说明帖子到生命期的后段依然保持着一定的活跃度.

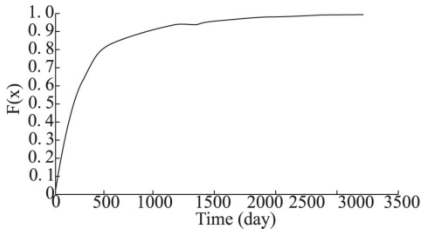

图 13 帖子生命长度的累积分布

Fig. 13 Cumulative distribution of age of post

针对贴吧帖子的这一特性,我们考察贴吧帖子在整个生命期内的传播特性. 首先我们爬取贴吧帖子的所有回复时间,包括对回复进行回复的时间,排序之后把帖子的生命长度分为 30 段,统计每一段内的回复数,得到时间序列.

4.3.2 聚类结果

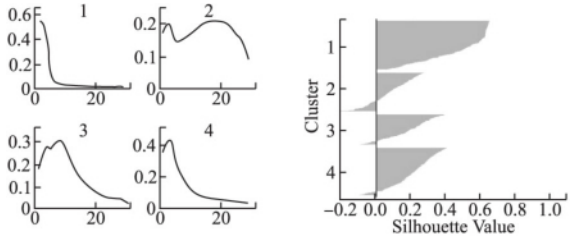

图 14 T-SC 算法对百度贴吧回复时间序列聚类的结果

Fig. 14 Result of T-SC

对上述贴吧回复时间序列运用 T-SC 算法,我们得到如图 14 所示的聚类结果及其 Silhouette 指标图,各个类所占个数依次为 298、203、197、302. 聚类结果中的四个簇所占比例大致相当,没有一种占统治地位的模式. 模式 1 中的帖子在发布一段时期内有较多的回复,但之后很长时期只有零星的回复出现; 模式 2 中的帖子一直都有很高的活跃度,即使中间经历小的下降之后又会有很长时间的稳定回复; 模式 3 中的帖子发布后活跃度逐渐上升,达到回复活跃度的顶峰后缓慢下降; 模式

4 与模式 3 比较相似,不同的是其回复活跃度的上升和下降都更快.

5　传播预测

时间序列的聚类分析能够发掘信息传播的规律,利用聚类分析的结果我们尝试着对信息的传播进行预测. 一条信息在某时刻 t 之后的传播行为与 t 之前的传播行为存在某种联系,即使不能建立精确的预测模型和预测公式,我们也能实验性的对未来的传播行为做出某种程度的预测.

以腾讯微博的聚类分析为例进行说明,我们首先得到长度为 48 小时的时间序列的集合 U,并对 U 做聚类分析,得到聚类中心的集合 $\mu_1,\cdots,\mu_K$. 然后基于某一条信息 x 的前 12 个小时的时间序列 $x_{(1,12)}$,判断该序列与聚类中心集合 $\mu_1,\cdots,\mu_K$ 中哪一个聚类中心的前 12 小时时间序列距离最近,记该序列为 y,我们把聚类中心 y 作为对 x 传播的近似描述,y 序列的后 36 小时的序列 $y_{(13,48)}$ 可做为对 x 序列接下来 36 小时传播序列 $x_{(13,48)}$ 的预测.

算法 3 是对上述想法的具体实现. 性能方面,在已知聚类中心的情况下算法的复杂度较低,其最坏情况运行时间是聚类个数 K 的线性函数.

算法 3. 基于聚类分析的传播预测算法.

输入: 时间序列 $x_{(1,12)}$,聚类个数 K,聚类中心 $\mu_1,\cdots,\mu_K$.

FOR $i=1$ TO K DO

$$d_i \leftarrow \hat{d}(x_{(1,12)},\mu_{i(1,12)})$$

ENDFOR

$$i^* \leftarrow \text{argmin}_{i=1,\cdots,K} d_i$$

$$y \leftarrow \mu_{i^*}$$

RETURN $y, y_{(13,48)}$

输出: 预测结果 y 和 $y_{(13,48)}$

我们利用腾讯微博的数据对上述算法进行了实验验证. 在已通过聚类分析得到聚类中心的基础上,对 100 条未参与聚类分析的腾讯微博进行传播预测. 首先对"正确预测"进行定义:

对 x 整个 48 小时上的传播,正确预测定义为预测得到的 y 即为真实 x 的聚类中心,即

$$y = \text{argmin}_{\mu=\mu_1,\cdots,\mu_K} \hat{d}(x,\mu) \quad (2)$$

对 x 未来 36 小时上的传播,正确预测定义为预测得到的 $y_{(13,48)}$ 与真实的 $x_{(13,48)}$ 属于同一个聚类中心,即

$$\mu^* \leftarrow \text{argmin}_{\mu=\mu_1,\cdots\mu_K} \hat{d}(x_{(13,48)},\mu_{(13,48)}) \quad (3)$$

$$y_{(13,48)} = \mu^*_{(13,48)} \quad (4)$$

进一步的,把以上实验中已知时间序列长度12小时作为变量,观察预测正确率随已知时间序列长度变化的规律,结果如图 15 所示,上图是的 x 预测正确率曲线,下图是 $x_{(13,48)}$ 的预测正确率曲线,横轴的时间单位都是 0.5h.

表 3　预测结果

Table 3　Result of prediction

预测对象	正确次数	错误次数	正确率
x	98	2	98%
$x_{(13,48)}$	73	27	73%

我们看到,算法 3 对 x 的整体预测呈现很高的正确率,即我们只需考察微博发出之后很短时间内的传播趋势,即可对其整体上的传播模式有一个较为准确的估计. 对 $x_{(13,48)}$ 的预测在已知时间序列长度处于 $12h\sim22h$ 区间时也有较高的正确率. 违反直觉的是,已知时间长度超过一定值时,预测准确率反而下降,这是因为根据我们对预测成功的定义,当已知时间长度过长时,未知时间序列即待预测时间序列过短,这样也会影响预测结果的判断.

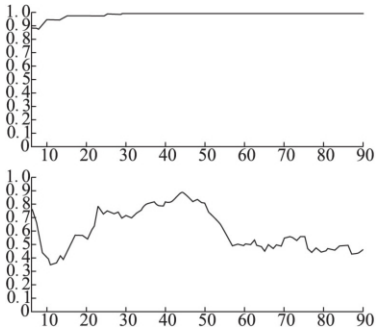

图 15　预测正确率曲线

Fig. 15　Curve of prediction accuracy

虽然这里的预测只是聚类意义上的,无法实现对信息传播真正意义上的精确预测,但对于我们从无到有地了解一个信息未来的传播走势还是很有参考意义的,也为致力于传播预测难题的研究人员提出了一种可以继续尝试的新思路.

6　总结和展望

本文改进了用于时间序列聚类分析的 K-SC 算法,提出的 T-SC 算法借鉴凝聚层次聚类的思想解决了聚类个数设置的难题. 我们选取了在中国非常有代表性的三个社交网站: 人人网、腾讯微博和百度贴吧,进行了大量测量和分析工作,利用对信息传播时间序列的聚类分析,重点观察三个社交网站典型的传播模式,并加以对比. 研究发现,每个社交网站都有其区别其他的典型传播特征,比如人人网视频分享有明显的周期特征,腾讯微博呈现很强的爆发性,百度贴吧有很长的生命周期. 最后,利用聚类分析的结果我们也创新性的进行了信息传播预测的探索.

解决聚类个数问题的另一种思路是采用不需要设置聚类个数的算法,比如 AP 聚类算法. 如何用 AP 聚类算法实现社交网络时间序列的聚类,进而与 T-SC 算法的聚类结果进行比较和分析是我们下一步的工作. 此外,如何进一步优化文中聚类层面的传播预测结果,并将其应用于真正的传播预测也是值得思考的研究方向.

References:

[1] Han J, Kamber M, Pei J. Data mining: concepts and techniques [M]. Morgan Kaufmann, 2006.

[2] Xu Ke, Zhang Sai, Chen Hao, et al. Measurement and analysis of online social networks [J]. Chinese Journal of Computers, 2014, 37(1): 165-188.

[3] Anderson R M, May R M, Anderson B. Infectious diseases of humans: dynamics and control [M]. Oxford: Oxford University Press, 1992.

[4] Gruhl D, Guha R, Liben-Nowell D, et al. Information diffusion through blogspace [C]. WWW 2004, New York, NY, USA: ACM: 1-58113-844-X/04/0005, 2004.

[5] Adar E, Adamic L A. Tracking information epidemics in blogspace [C]. The 2005 IEEE/WIC/ACM International Conference on Web Intelligence, 2005. doi: 10.1109/WI.2005.151.

[6] Kramer A D I. The spread of emotion via facebook [C]. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2012: 767-770.

[7] Wu S, Hofman J M, Mason W A, et al. Who says what to whom on twitter [C]. Proceedings of the 20th International Conference on World Wide Web, ACM, 2011: 705-714.

[8] Bakshy E, Karrer B, Adamic L A. Social influence and the diffusion of user-created content [C]. Proceedings 10th ACM Conference on Electronic Commerce (EC-2009), Stanford, USA, 2009. doi: 10.1145/1566374.1566421.

[9] Zhang Sai, Xu Ke, Li Hai-tao. Measurement and analysis of information propagation in online social networks like microblog [J]. Journal of Xi'an Jiaotong University, 2013, 47(2): 124-130.

[10] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network [C]. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2003: 137-146.

[11] Ver Steeg G, Galstyan A. Information transfer in social media [C]. Proceedings of the 21st International Conference on World Wide Web, ACM, 2012: 509-518.

[12] Cha M, Kwak H, Rodriguez P, et al. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system [C]. Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, ACM, 2007: 1-14.

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

[14] Li H, Ma X, Wang F, et al. On popularity prediction of videos shared in online social networks [C]. Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management. ACM, 2013: 169-178.

[15] Kaufman L, Rousseeuw P J. Finding groups in data: an introduction to cluster analysis [M]. Wiley. com, 2009.

[16] Hamerly G, Elkan C. Alternatives to the k-means algorithm that find better clusterings [C]. Proceedings of the Eleventh International Conference on Information and Knowledge Management, ACM, 2002: 600-607.

[17] Mishra B K, Rath A, Nayak N R, et al. Far efficient K-means clustering algorithm [C]. Proceedings of the International Conference on Advances in Computing, Communications and Informatics. ACM, 2012: 106-110.

[18] Frahling G, Sohler C. A fast k-means implementation using coresets [J]. International Journal of Computational Geometry & Applications, 2008, 18(6): 605-625.

[19] Zhou Shi-bing, Xu Zhen-yuan, Tang Xu-qing. New method for determining optimal number of clusters in K-means clustering algorithm [J]. Computer Engineering and Applications, 2010, 46(16): 27-31.

[20] Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding [C]. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, 2007: 1027-1035.

[21] Warren Liao T. Clustering of time series data-a survey [J]. Pattern Recognition, 2005, 38(11): 1857-1874.

[22] 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. ACM, 2011: 177-186.

[23] Han Zhong-ming, Chen Ni, Le Jia-jin, et al. An efficient and effective clustering algorithm for time series of hot topics [J]. Chinese Journal of Computers, 2012, 35(11): 2337-2347.

[24] Caliński T, Harabasz J. A dendrite method for cluster analysis [J]. Communications in Statistics-theory and Methods, 1974, 3(1): 1-27.

[25] Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset [J]. Genome Biology, 2002, 3(7): 36.

附中文参考文献:

[2] 徐 恪,张 赛,陈 昊,等. 在线社会网络的测量与分析 [J]. 计算机学报,2014,37(1):165-188.

[9] 张 赛,徐 恪,李海涛. 微博类社交网络中信息传播的测量与分析 [J]. 西安交通大学学报,2013,47(2):124-130.

[19] 周世兵,徐振源,唐旭清. 新的 K-均值算法最佳聚类数确定方法 [J]. 计算机工程与应用,2010,46(16):27-31.

[23] 韩忠明,陈 妮,乐嘉锦,等. 面向热点话题时间序列的有效聚类算法研究 [J]. 计算机学报,2012,35(11):2337-2347.