基于地理标签的推文话题时空演变的可视分析方法

孙国道 周志秀 李 思 刘义鹏 梁荣华
(浙江工业大学信息工程学院 杭州 310023)

摘 要 社交媒体中,用户所发布的推文内容记录了与用户相关的各种信息。文字信息中涵盖了推文中包含的各种话题,以及时间和空间信息,从这些信息中分析出话题的时空演变情况具有十分重要的研究意义。针对推文数据,设计了一套可视分析流程来挖掘推文信息,通过用户交互的方式多角度地展示了推文话题的时空演变过程。首先,基于部分历史推文数据,通过DBSCAN(Density-Based Spatial Clustering of Applications with Noise)聚类算法,结合泰森多边形对全球地理空间进行区域划分;然后,针对用户查询搜索的兴趣话题,索引找到所有相关的推文内容,并将信息与聚类中心绑定;最后,通过设计的多个结合时序聚类算法和自适应算法的可视化视图来展示话题的时空演变过程。通过推特官网提供的API抓取存储的推文数据,并进行实验和分析,结果表明:改进的可视化视图自适应布局算法有效地解决了图形遮挡问题,完整展现了推文的时空演变模式;地理区域的划分以及可视化组件能够有效帮助研究人员分析推文的时空演变以及全球关注的热点话题分布。

关键词 推文话题,可视化分析流程,自适应布局算法,聚类,时空演变

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.08.007

Spatio-Temporal Evolution of Geographical Topics

SUN Guo-dao ZHOU Zhi-xiu LI Si LIU Yi-peng LIANG Rong-hua
(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract The tweets posted by users in social media record a wide variety of user information. The text information includes various topics contained in the tweet. It is very important to analyze the temporal and spatial evolution of topics from these messages. Based on the tweet data, this paper designed a set of visual analysis process to mine the tweet information and display the spatiotemporal evolution process of the tweet topic through user interaction. Specifically, based on the partial historical tweet data, the global geographic space is divided by the DBSCAN clustering algorithm combined with the Tyson polygon. For the user to query the search topic of interest, the index finds all relevant tweet content and binds the information to the cluster center. Finally, the temporal and spatial evolution of the topic is demonstrated by the design of multiple combined time series clustering algorithms and visualization components of the adaptive algorithm. Through the experiment and analysis of the tweet data stored in the API provided by Twitter official website, the improved visual view adaptive layout algorithm effectively solves the problem of graphic occlusion and fully displays the temporal and spatial evolution mode of the tweet. The division of geographic regions and visualization components can effectively help researchers analyze the temporal and spatial evolution of tweets, as well as the distribution of hot topics of global concern.

Keywords Tweet topic, Visual analysis process, Adaptive layout algorithm, Clustering, Spatio-Temporal evolution

1 引言

社交网络服务通过诸如网站在线平台的建立来反映人们之间的联系,提供了一个用户通过英特网来分享想法、活动、事件和兴趣的平台,例如推特和微博。社交媒体是基于用户关系的内容生产与交换的平台,每天都在全球范围内产生大量的信息,包含政治新闻、突发事件和名人活动等。推特作为当前社交媒体平台最流行的软件之一,允许人们产生、传播和交换信息。这些信息包含了各种用户的签到信息,每天大约产生300万条用户签到信息,包含了经度-纬度坐标。

目前,社交媒体上的话题标签主要有两种功能。1)话题标注:在社交媒体中,用户会使用一个带有井号(#)的具有代

到稿日期:2018-11-26 返修日期:2019-01-18 本文受国家自然科学基金(61602409)资助。

孙国道(1988-),男,博士,讲师,CCF会员,主要研究方向为信息可视化,E-mail:guodao@zjut.edu.cn;**周志秀**(1994-),女,硕士生,主要研究方向为信息可视化;**李 思**(1992-),男,硕士生,主要研究方向为信息可视化;**刘义鹏**(1987-),男,博士,讲师,CCF会员,主要研究方向为医学成像和图像分析;**梁荣华**(1974-),男,博士,教授,博士生导师,CCF高级会员,主要研究方向为信息可视化和计算机图形图像处理,E-mail:rhliang@zjut.edu.cn(通信作者)。

表性的话题词来标注自己所发推文的主题，因此话题标签具有话题标注的功能。2）话题参与：话题标签的另一主要目的是发挥“话题参与”的功能，即将同一个话题下的信息汇聚起来，以提高信息传播和组织的效率[1-3]。

用户发布的包含话题标签的推文，也在一定程度上反映了网民所关注的话题。比如，某个用户发送的推文中包含标签＃euro2016，据此可以洞察该用户发布的推文是关于 2016 年法国欧洲杯的信息。具体来说，本文把研究重点放在推特传播的话题标签上，一个话题标签是一个简单的用户生成的注释。在推特上话题标签提供了许多用途，最重要的功能是将推文内容与特定事件的标签联系在一起，其次话题标签之间可以有包含关系（例如欧洲杯热门国家的话题标签＃italiangermania和＃gerita 都可以被大标签欧洲杯＃euro2016包含）。

因此，本文采用不同的话题标签作为推文的分类标准，在可视分析流程中筛选带有地理签到信息的推文话题标签作为研究内容。针对处理好的带有地理标签的推文数据，我们的目标是探索以下几个问题。

（1）话题时空演变研究：用户对话题的关注程度随着时间的推移在全球空间上有着怎样的传播轨迹？相同话题随时间的热度变化在空间上是否呈现一定的相似性？

（2）话题相关性分析：相关话题之间是否存在某种联系？相关话题在时空的演变模式上是否存在相似性？

针对上述问题，构建数据挖掘模型，改进区域相似度聚类算法，设计有效的可视化组件来展示这些话题的时空演变模式和话题相关性，具体的可视化分析流程如图 1 所示。对此，本文开发了可视分析原型系统（如图 2 所示，其中 A 和 B 为话题时空演变模式可视化分析，C 为话题区域相似聚类分析，D 为话题相关性树图分析，E 为相关话题时空演变比较分析），通过对抓取真实推特数据的可视化分析，表明本文设计的可视分析流程可以有效地帮助研究人员分析推文内容的时空演变模式。在可视化展示分析过程中，本文设计的启发式的自适应布局算法通过节点移动和多边形约束，来进一步减少图形遮挡。

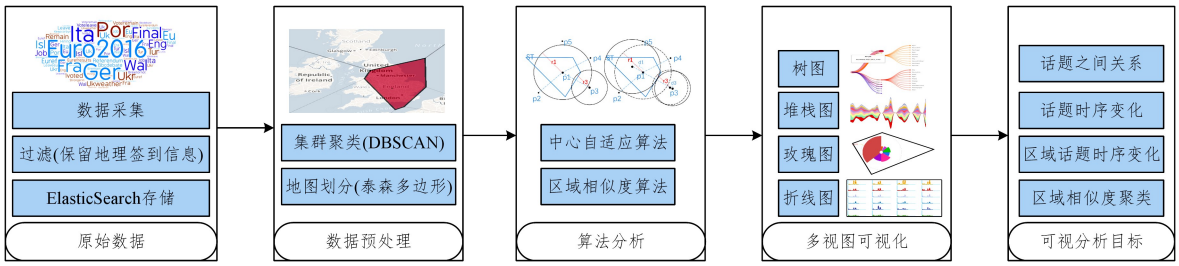


图 1　推文数据可视化分析流程

Fig. 1　Tweet data visualization analysis process

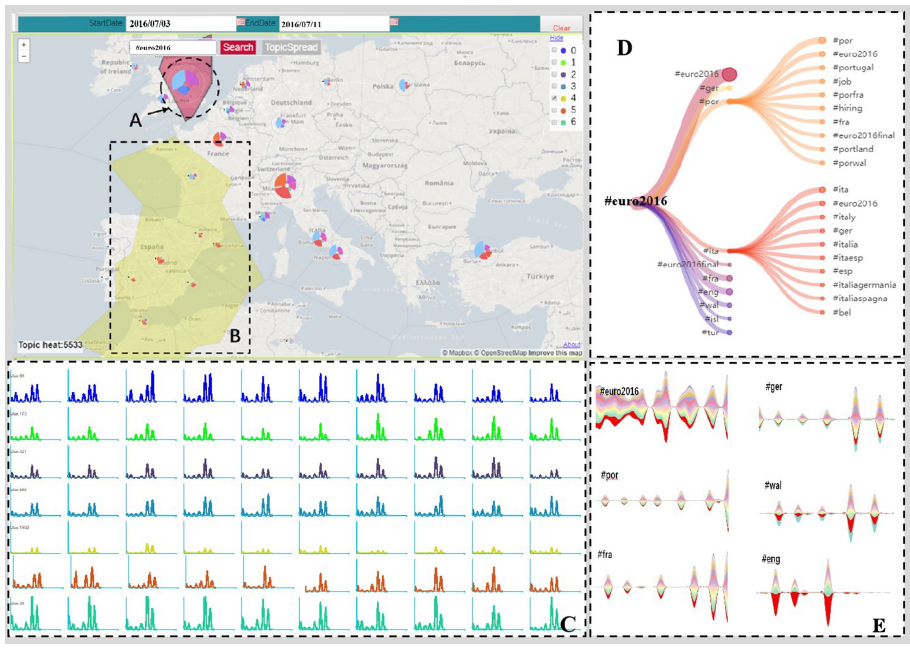


图 2　推文数据可视化分析系统（电子版为彩色）

Fig. 2　Tweet data visualization analysis system

2　相关工作

2.1　社交媒体数据的可视分析

早期，一些学者研究了推特作为社交网络的一般属性，并分析了其在社交网络上的信息反馈[4-7]。在该方向上，与推文地理标签相关的大多数工作都将精力集中于解释、探索焦点话题在网络中的传播规律上。例如，Romero 等[8]研究了催化地理标签反应的因素，发现重复暴露的地理标签增加了推文

再次转发的机会，特别是在地理标签有争议的情况下。文献[9]基于语言原理的方法研究了标签创建、使用和传播的属性。在相关研究中，基于线性回归的方法已被用来预测地理标签在给定时间框架中的受欢迎程度[13]。地理标签在传达推文信息方面有多种方式，最近基于标签的情绪检测[8]和话题跟踪在推特流[10]上的研究是比较热门的方向。地理标签的流行度特性包括地理标签在一段时间内被使用的状态，这种状态主要通过地理标签的频次来界定。Kong 等[11]根据地理标签生命周期中的频次变化定义地理标签的 4 种流行度：出现、爆发、平静、沉寂。Ma 等[12]和 Tsur 等[13]也按照地理标签的使用频率划分地理标签的流行度，通过对流行度的划分，可以将地理标签的流行度预测问题转化为分类问题；按照不同的频次等级对地理标签进行类别划分，使用分类器对地理标签进行流行度类别的预测，从而预测出地理标签在未来的使用频次。Cho 等[14]提出的 CrystalBall 利用社交媒体信息寻找未来可能发生的事件，重点研究各事件在各空间中随时间的变化情况。本文通过对推特中带有地理标签的推文进行筛选，设计了话题关系流图来展现不同话题之间的关系，重点研究探索话题事件的热度随时间变化的规律，以及话题在空间上的演变过程。

2.2 时空数据的可视分析

目前很多工作把社交媒体包含的地理空间信息作为研究的重点，对用户发布信息时标记的空间地理信息进行数据处理和可视化展现。Kamath 等[15]通过建模和理解来分析社交媒体信息的全球传播，构建了一个模型来预测未来社交媒体中的热点新闻的流行区域。He 等[16]利用视觉滤波器对时空文本数据进行研究，引入了话题轨迹的概念来描述话题的时空演化。Kamath 等[17]基于推特传播的 7.55 亿个地理标记的标签，开发了一种全球轨迹驱动方式，以确定将在特定位置变得流行的地理标签，并逐步更新最佳的地理空间模型。MacEachren 等[10]专注于微博的显式和隐式地理信息，开发了一套支持概述和细节方法的地点时间话题模型。Hong 等[18]关注于推特，并通过基于局部多样性、地理多样性和用户兴趣分布三大特点来研究推文传播特点。越来越多的研究者通过提出算法来建立多样性模型，用于用户行为推测和事件因果分析[19-21]。然而，对于没有轨迹信息的非方向统计数据，研究者很难提取时空信息以及可视化数据流的模式。

文献[18]提出了一种新颖的流量分析技术，用于提取、表示和分析非定向时空数据，而不伴随轨迹信息。Andrienko 等[22]提出了另一种方法，以促进对长期数据流动空间和时间的抽象探索工作。Marcus 等[23]和 Cao 等[24]针对推文数据在时空中的追踪问题，分别提出了 Twitinfo 和 Whisper。Twitinfo 的重点在于基于时间轴上的显示突出推文的高峰期，并对其进行标记；Whisper 通过用户组转发推文，提取推文中的情绪，并在空间分层布局上跟踪转推路径。

本文同样从社交媒体中常见的地理标签入手，对推文热度进行跟踪，利用改进的分析算法，侧重于帮助研究人员分析不同话题之间存在的关系；设计可视化组件，探索热门话题在时空中的演变模式。本文工作有助于深入研究社交媒体中话题的演变模式和用户的行为特征。

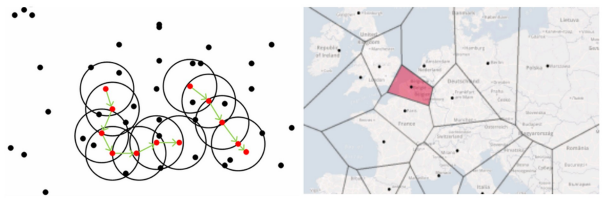
3 数据预处理、聚类与地图划分

3.1 数据的采集和预处理

通过推特官网提供的 API 来抓取推特数据，选取包含签到信息的推文数据作为原始数据，并对数据进行预处理，只保留推文的内容、地理签到信息以及时间信息。本文设计的可视化流程支持用户对话题关键词的搜索以及可视化交互，对数据处理的实时性要求较高，因此使用 ElasticSearch 进行数据存储。ElasticSearch 数据库是一个可拓展的、开源的、全文检索分布式搜索引擎，它提供接近实时的数据查找、索引和挖掘的功能，有助于提高整个可视化分析流程的顺畅性以及用户使用的友好度。

为了分析推文话题在时空范围内的演变情况，选取历史推文数据进行空间聚类，为可视化展示中的地图区域划分提供依据，有利于研究者观察话题在空间上的传播演变。根据过滤后推文数据的时空属性，基于推文数据在空间上的密度，根据推文数据自带的经纬度坐标，采用 DBSCAN 算法对所有数据进行聚类(见图 3(a))。DBSCAN 算法是目前使用得最广泛的一种基于密度的聚类算法，它可以发现任意形状的聚类中心。

经过多次实验和参数调整，最终将半径设置为 50 km，得到了较好的地图空间布局效果。然后利用该聚类算法得出的聚类中心，将全球划分为 226 个区域，聚类的结果如图 3(b)中黑点所示，从结果中发现聚类中心在各个国家的大城市附近，与大城市人口基数大的事实相符。



(a)DBSCAN 算法示意图 (b)对地图进行泰森多边形划分的结果

图 3 数据和地图的预处理(电子版为彩色)

Fig. 3 Data and map preprocessing

3.2 基于泰森多边形的地图划分和数据绑定

在对数据进行 DBSCAN 聚类处理后，保留数据处理后的 226 个聚类中心点坐标。基于聚类中心对全球进行空间划分，使得划分后的每个区域内，用户发布推文内容时所在的地理位置距离该聚类中心最近。本文采用泰森多边形对全球地图进行划分，以为后续自适应算法的南丁格尔玫瑰图可视化组件提供约束条件。泰森多边形的建立过程如下：首先对数据进行 DBSCAN 空间聚类，获得多个聚类中心，以这些聚类中心为基础，找到各个聚类中心的相邻中心点。如图 4 所示，以聚类中心 A 为基础，找到与之相邻的聚类中心 C 和 D。然后分别对上述聚类中心作垂直平分线。通过生成的垂直平分线(如图 4 中虚线所示)来构建泰森多边形，从而使得每个聚类点所围成的区域内的所有点距离该中心最近。确定聚类中心和地图划分被后，在后续的数据查询可视化分析中不会改变，因为全球范围内人口不会在短时间内有大量的迁移过程。预先的聚类和区域划分，加快了后续推文数据搜索处理分析的速度。

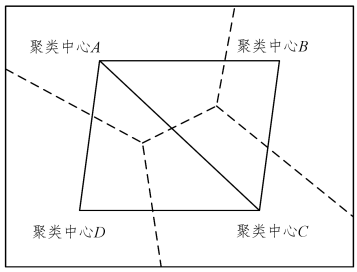


图 4　泰森多边形示意图

Fig. 4　Tyson polygon diagram

后续只需要根据用户查询搜索的兴趣话题，索引找到所有相关的推文内容，然后根据推文数据的经纬度信息与聚类中心绑定，生成可视化组件进行展示分析。在计算用户发布的推文数据距离最近的聚类中心的过程中，本文采用了 k-d tree 算法来加快后台数据的处理速度。k-d tree 是对数据点在 k 维空间中进行分割的一种数据结构，在多维空间数据的索引过程中被广泛应用。经过比较，相比单纯循环计算求出地理距离后再排序的方法，通过 k-d tree 算法优化后的处理效率至少提高了 5 倍，在数据量增大时，处理速度呈几何上升。

4　可视化分析模型与设计

本节将结合改进算法的可视化展示，来研究话题时空演变和话题相关分析。对数据建立了自适应的南丁格尔玫瑰图热度分析模型，用于分析推文话题的时空演变模式；基于相似度聚类模型，对各区域话题的演变模式进行聚类；同时设计了树图来展示相关话题的联系；采用堆栈图对相关话题的热度在时间上的变化做比较分析。

4.1　基于自适应玫瑰图的热度分析模型

通过 DBSCAN 聚类算法将世界地图划分为若干部分后，需要全局展示用户对某个热点话题的关注程度随着时间推移的变化趋势。本文设计了一个可视化组件来展示话题的时空演变情况，采用了南丁格尔玫瑰图来表示一个区域内话题的热度变化情况，用不同颜色来表示不同日期，使用圆环半径的大小来表示不同日期内用户对该话题的关注程度，当用户把鼠标移动到扇形中的某个区域时，界面会显示该扇形所对应时间的话题热度（如图 5 所示，该话题的关注程度在第 8 天时有明显的提高），原型系统界面效果展示图如图 7(b)所示。从中可以直观地看到，话题在哪些区域比较集中，以及话题在各个区域的关注热度在时间维度上的变化情况，同时通过不同的颜色映射可以直观地看出用户对话题的关注热度随时间的变化情况（如图 5 中的红色代表第 8 天的话题热度有明显的提高）。

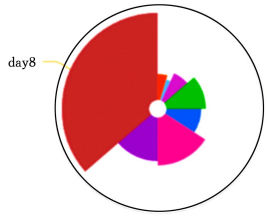
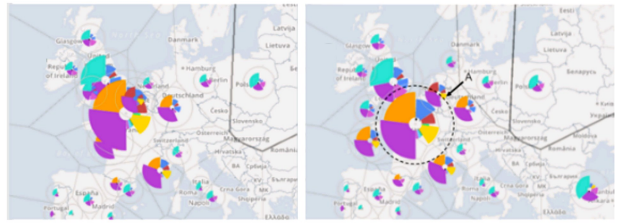


图 5　南丁格尔玫瑰图（电子版为彩色）

Fig. 5　Nightingale rose chart

在地图上各区域使用南丁格尔玫瑰展示话题随着时间变化的情况下，会出现由于某区域的话题量很大而与相邻区域的其他圆环相重叠的问题，如图 6(a)所示。在布局算法中，既要考虑视觉遮挡又要考虑实际地理位置，因此我们设计了一种基于多边形约束的自适应布局的启发式算法来帮助改善此问题。使用地图划分后各区域的泰森多边形作为约束条件，在保证地理相对位置不变的情况下，通过自适应算法移动南丁格尔玫瑰图的圆心，将其约束在所在的多边形内，以避免与相邻区域的重合，从而保证不同区域之间的话题热度比较。



(a)　(b)

图 6　多边形约束的结点布局自适应算法处理前后的效果比较

Fig. 6　Comparison of effects of polygon-constrained node layout adaptive algorithm

多边形约束的结点布局自适应算法如下：

通过 DBSCAN 算法聚类后，将获得的聚类中心的地理位置作为玫瑰图圆心的原始位置$\{p_1, p_2, p_3, p_4, p_5\}$，如图 7(a)所示。以 p_1 点为中心，泰森多边形划分以后形成的多边形为 ST，玫瑰圆面积为 S_1。

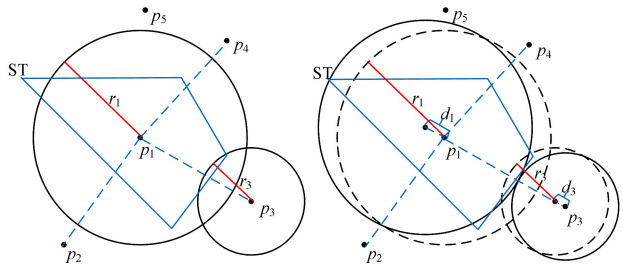
当 p_1 和 p_3 上的玫瑰图重叠时，对于 p_1，反方向移动：

$$d_1 = (r_1 + r_3 - D_{(p_1 - p_3)}) * \frac{r_1}{r_1 + r_3} \tag{1}$$

对于 p_3，反方向移动：

$$d_3 = (r_1 + r_3 - D_{(p_1 - p_3)}) * \frac{r_3}{r_1 + r_3} \tag{2}$$

设 ST 和 S_1 所重叠的面积为 Sc，当 Sc 的大小为 S_1 圆的一半时，圆心位置不再移动，以此作为上限条件，避免了玫瑰图圆心移动到泰森多边形以外。圆心各自移动后的情况如图 7(b)所示。



(a)原始布局　(b)圆心偏移

图 7　圆心点自适应偏移算法示意图

Fig. 7　Schematic diagram of center point adaptive offset

通过自适应算法来避免不同的玫瑰图在可视化展现时出现重叠，图 7(b)验证了自适应算法的有效性。多边形约束的结点布局自适应算法的具体伪代码实现如算法 1 所示。

算法 1　多边形约束的结点布局自适应算法

Inputs：

原始圆心点：

$P_1=(x_1,y_1)$,$P_2=(x_1,y_1)$,$P_3=(x_1,y_1)$,$P_4=(x_1,y_1)$… //核心点空间坐标

List={P_1,P_2,P_3,P_4,P_5,…} //DBSCAN 聚类后的核心点坐标(原始圆心)集合

圆心点所在圆的半径=R_1,R_2,R_3,R_4,R_5…

原始圆心点经过划分后形成的多边形的面积=ST_1,ST_2,ST_3,ST_4,ST_5,…

For P_i in List

  For P_j in List

    If P_i不等于 P_j then

      d_1=Distance(P_i,P_j) //d_1 表示以 P_i 为标准点,P_j 反方向移动的距离

      d_2=Distance(P_j,P_i) //d_2 表示以 P_j 为标准点,P_i 反方向移动的距离

      $P_m=P_i-d_1$//将移动后的点作为新的 P_i

      $P_n=P_j-d_2$//将移动后的点作为新的 P_j

      If ST_i 和以 P_m 为圆心的圆的重叠面积小于圆面积的一半

        //作为移动的限制条件 then 继续循环

      If ST_i 和以 P_m 为圆心的圆的重叠面积小于 ST_i 面积的一半 then 继续循环

      $P_i=P_m$//将移动后的 P_i 作为新的圆心点

      $P_j=P_n$

  End

End

Outputs:经过自适应算法处理后的一组圆心点坐标

在上述多边形约束的结点布局自适应算法中,遍历每个区域(一共 n 个区域)的南丁格尔玫瑰图,针对每个南丁格尔玫瑰图寻找与其重叠的其他玫瑰图,并做自适应调整,因此算法的时间复杂度为 $O(n^2)$。在实际的运算过程中,由于处理的数据量较小,因此复杂度对时间的影响较小。当数据量特别大时,为了提高第二层寻找重叠玫瑰图的效率,可以采用四叉树索引,将地图空间划分为不同层次的树结构。该方法运行在 $O(\log n)$的时间复杂度内。对此,使用四叉树索引可整体降低多边形约束的结点布局自适应算法的时间复杂度,最后复杂度为 $O(n\log n)$。

4.2 基于相似度聚类的区域分析

前文对各区域关于用户感兴趣的话题设计了南丁格尔玫瑰图来展示热度的时序变化情况,但研究者很难在划定的 226 个区域中发现它们关于话题关注度演变的相似性。对此,我们针对各区域的推特话题数据,绘制各个区域关于某个话题的关注程度变化折线图,对具有相似演变模式的区域进行聚类。

这里设计了一种基于 Pearson 相关系数的改进算法来对各区域关于话题的热度变化模式进行聚类。时序聚类算法的公式如式(3)、式(4)所示:

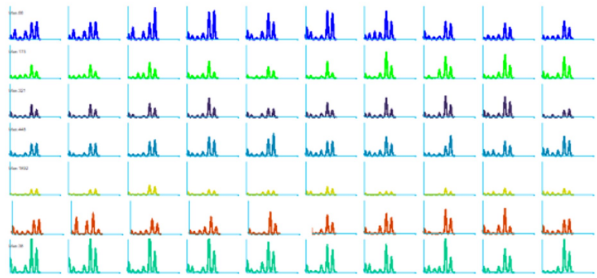
$$s_p=1-\frac{\sum x-\sum y}{\sum x+\sum y},s_p\in(0,1) \tag{3}$$

$$\rho_{x,y}=\frac{\sum xy-\frac{\sum x\sum y}{N}}{\sqrt{(\sum x^2-\frac{(\sum x^2)^2}{N})(\sum y^2-\frac{(\sum y^2)^2}{N})}} \tag{4}$$

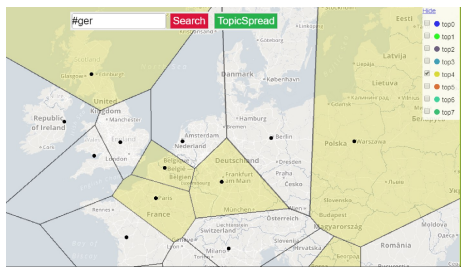
其中,x 和 y 为两个数组,分别表示两个区域关于某个话题在一段时间内的关注程度的变化情况;N 表示数组的数目;$\rho_{x,y}$ 的取值为−1～1,越接近 1 表示两组数据越相似。本文基于 Pearson 相关系数,同时把某个话题在该区域的讨论频率总和作为一个影响因素,增加了参数 S_p 来表示数组 x 和数组 y 的总和是否相近,取值范围为 0～1,如果两组数的总和越相近,则 S_p 的数值越接近于 1。然后,将 $\rho_{x,y}$ 和 S_p 相乘的结果作为衡量两个区域对应两条关于某个话题关注程度曲线的相似程度的标准。

在地图上,根据初步展示的南丁格尔玫瑰图,分析人员首先选择一块感兴趣的区域作为基准点;然后通过该时序聚类算法找到与该基准点有相似变化模式的若干区域,同时把这些区域聚成一类并通过相同的颜色进行标注;接着找到与该基准点相似度最低的一块区域作为下一次遍历的基准点,对剩下的区域进行算法迭代;最后将经过泰森多边形划分后的地图聚成若干类,并通过颜色映射进行可视化展现(如图 8(a)所示,每一行表示一个聚类集合,同一行内的 10 个区域具有话题相似性)。

研究者可以自定义聚类的数目,本文各案例分析定义聚类数目为 7,每类区域折线图的数量为 10,能较好地发现话题的时空演变模式。与此同时,我们将区域聚类结果的话题时空演变模式在地图上进行交互展现,用户可以通过原型系统地图右上方的选择框来选择不同的聚类结果在地图上进行展示。如图 8(b)所示,选择类别 4,使用类别颜色高亮对应区域在地图上的分布情况(如图 8(b)中黄色区域所示),可以发现空间上相近的区域在话题热度的演变模式上具有相似性。



(a)区域话题演变的相似聚类视图



(b)地图上相似区域的选择标注

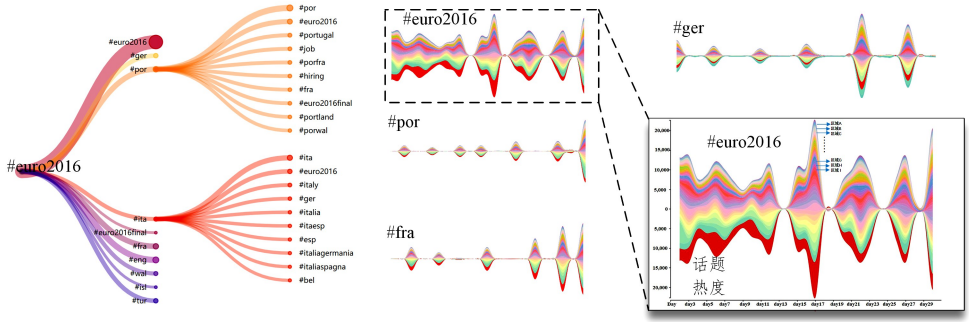
图 8 区域话题热度时序变化的相似性聚类(电子版为彩色)

Fig. 8 Similarity clustering of topic heat variation

4.3 话题相关性分析

4.1 节和 4.2 节解决了话题时空演变模式的探索问题,关于话题相关性分析,本文采用树图来表示不同话题标签之间的关系。首先选择一个大的话题标签作为关键词,然后从用户发表的内容中找到出现这个关键词的所有句子,统计这

些内容中出现的其他分支话题标签的频率，最后进行统计排序。例如，选择＃euro2016 话题标签作为关键词，然后找到与之相关联的分支话题标签进行排序，结果如图 9(a)所示，从中可以明显看出与＃ euro2016 相关性较高的话题是＃por（葡萄牙国家队伍），与 2016 年葡萄牙在欧洲杯取得冠军的事实相符。



(a)话题相关性的树图展示　　(b)相关话题时空演变的比较分析

图 9　话题关系视图(电子版为彩色)

Fig. 9　View of topic relationship

与此同时，我们对相关话题在时空的演变模式上是否存在相似性这一问题进行了进一步的探索，并采用堆栈来展现不同区域内用户对某个话题关注程度的时序变化。如图 9(b)所示，用不同颜色来区分不同的区域，横坐标表示时间，纵坐标每个颜色曲线的宽度表示话题的讨论热度。当研究者在地图上点击某块区域时，该区域在流图中被红色高亮标记(如图 9(b)中各话题堆栈图中的红色区域所示)。通过比较红色高亮区域的宽度变化，可以直观地看到该区域关于某个话题的关注热度在一段时间内的变化情况，以及该区域对某个话题的关注热度在全球范围内所占的比例情况。当研究者点击图 9(a)中的其他分支话题时，原型系统生成对应的话题热度变化图，其有助于比较分析相关话题的热度在时间上的变化情况。从图 9(b)中发现，子话题＃por 和＃fra 与主要话题＃euro2016 具有相似的时空演变模式。

5　案例分析

5.1　不同话题演变模式的分析

本案例选择了＃euro2016 作为感兴趣的话题标签，来探究与其相关性较高的话题，以及它们分别具有的演变模式。如图 2 所示，首先在系统界面的输入框中输入关键词“＃euro2016”，在时间框中选择时间，点击系统中的“Topic-Spread”按钮，图 2 中的 D 区域展示与“＃euro2016”关联度最高的 10 个地理标签。由于我们的数据选择范围是法国欧洲杯最后 7 天内的推特数据，因此与＃euro2016 关联度较高的 10 个地理标签基本上都是参加比赛的国家；将这 10 个地理标签作为关键词，继续寻找与这些关键词关联度较高的 10 个地理标签，在得到这些话题在全球的分布情况后，再分析这些话题之间的联系。

本系统采用树图来表现不同地理话题之间的关系，话题关系视图如图 2 中的区域 D 所示。

可以很清楚地看到，与＃euro2016 相关性较高的为红色分支中的＃por 标签(除了＃euro2016 本身和＃euro2016final 以外)，其表明社交媒体用户对葡萄牙国家队的关注程度高于对其他国家的关注程度(这也与葡萄牙在欧洲杯取得冠军的事实相符)；接着＃por 标签代表的是意大利的＃ita 标签(意大利队在 2016 届欧洲杯中，开局两连胜，成为第 2 支出线球队)。接下来，我们希望从这些子地理标签在一段时间内的演变模式中找出一些与＃euro2016 存在相似性的地方。

在图 2 的 E 区域中可以看出，对话题＃euro2016 的关注程度在这段时间有着明显的波动，这些波动与比赛赛程基本吻合，即对一些热门赛程的关注程度明显增加。另外，我们发现在选择日期的第 4 天左右，＃euro2016 有一个显著的增加，对比其他地理标签发现，意大利(＃ita)和英国(＃eng)的热度在该天也有明显的增加。通过查询赛程发现，这一天这两支队伍刚好有比赛，且在这场比赛中意大利以 2∶1 的比分取得了胜利，从而导致英国(＃eng)这个话题在比赛之后的关注度迅速下降，后来也没有增加。

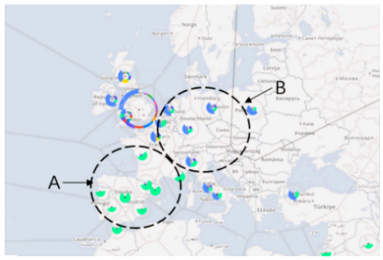
在经过泰森多边形划分的地图上点击英国所在的区域(如图 2 中 A 区域所示)，该区域对＃euro2016 和相关话题的关注程度在图 2 中的 E 区域中通过红色高亮进行展示，从中可以看到英国人民对欧洲杯有着很高的关注程度(通过红色高亮区域所占的比例可以看出)，同时出现的高峰是在赛程进行到一半，也就是英国最后一次对阵意大利的比赛中；另外，英国球迷对下一场意大利(＃ita)和德国(＃ger)的比赛的关注程度仅次于比赛的双方，说明英国人民很希望德国能在这场比赛中战胜意大利，从而实现复仇。

5.2　话题的时空演变可视化展现

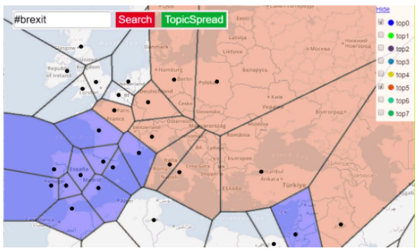
本案例将通过系统来表现热门话题在全球的演变情况。在利用 DBSCAN 算法将全球地理空间划分成 226 个区域后，需要观察某个话题热度在各个区域随着时间变化的情况。本案例选择将英国退出欧盟的关键词“brexit”作为关键词，brexit 是对英国退出欧盟的一种戏谑说法；然后对这期间全球各地关于该地理标签关注热度的变化进行分析，再以英国所在区域为基准点，对具有相似变化模式的区域进行聚类。基于改进的 Pearson 算法，从 226 个区域中选出 7 类，并用不同的颜色来区分不同的类。

从图 10(a)各区域的南丁格尔玫瑰图中可以看出，有些区域对话题的关注度有着明显的变化。案例分析选择热度较

高的顶点作为话题传播的途经点，然后在地图上展示这些点，从中可以推测网民对英国脱欧话题的关注程度有着明显的时空演变过程(从图 10(a)中的 B 区域演变到了 A 区域)，同时对该话题的讨论集中在英国所在的区域。再以英国为基准点，对具有相似变化模式的区域进行聚类，然后通过地理视图上的颜色按钮在地图上选择蓝色和红色，并将这些聚类颜色对应的区域进行高亮标记，如图 10(b)所示。演变模式聚类与图 10(a)发现话题区域变化相互验证，证明了可视化系统模型的有效性。



(a)话题热度变化在地图上的可视化展现



(b)具有类似传播模式的区域在地图上进行高亮

图 10 话题时空演变视图(电子版为彩色)

Fig. 10 Topic space evolution view

6 专家评估与建议

为了评估本文设计的可视化系统，我们邀请了 1 位传播学专业的博士以及 2 位传播学专业的硕士生参与系统实用性和有效性的评估。我们设计了一份在线调查问卷来协助专家进行定量评估，这份问卷调查针对本文工作的 4 个可视化视图和整体系统分别提出了两个问题。专家根据其对系统的案例操作和理解，对每个视图对应的问题分别打分(打分采用 5 分制：0(非常不认可)～5(非常认可))。

从评分结果来看，传播学领域的专家们对本文工作的认可度较高，对整体系统的直观性和实用性均给予了平均 4 分的认可度。4 个视图中，专家们对使用树图来展示话题之间的相关性的评价度最高(平均 4.5 分)。专家们认为树图的展示方法和思路与传播学中的“内容分析法”有异曲同工之妙，可视化的表达比“内容分析法”的单纯数据更加直观，交互设计也很棒。堆栈图也获得了较高的认可，它易于理解且直观，提供了另一种思考的角度，相当于把不同地区的话题热度图放在一起，有利于比较地区之间对同一话题的热度。玫瑰图和折线图获得的认可度分数较低，其对研究话题时空演变的有效帮助性得分都为 3.5 分。这可能是由于两个视图展示的内容具有多样性(时间和空间)，相对而言，用户需要一定的时间来理解和适应。专家们认为研究话题关注度演变的整体思路很好，通过数据有助于找到相关的规律，对于传播学中的公共危机处理、舆论控制都有很好的指导意义。

与此同时，针对本文工作的一些不足，专家们也提出了许多宝贵的建议。比如，在用玫瑰图结合地图展现话题时空演变的方式中采用了过多的颜色表示时间，这导致用户需要花费一段时间才能理解清楚；另外，玫瑰图的时间单位为“天”，不能详细描述趋势，这在一定程度上浪费了已有详细到分或者秒的数据。针对专家们提出的问题，我们将在后续的工作中进行改进，比如增加交互设计，帮助用户理解时空演变的可视化，以及提供不同时间单位的话题热度趋势分析。

结束语 本文基于推特中的话题标签数据，设计了一套可视分析流程，用于交互式可视化分析和理解，以帮助研究推文的时空演变模式。在可视化流程中，设计改进的聚类算法，对话题时序变化分区域聚类。通过启发式布局算法，自适应调整可视化视图的位置，进一步减少了图形遮挡。开发的原型系统能够基于真实数据，有效地推进话题时空演变研究和话题相关性分析。

邀请专家进行评估，验证了本文所设计方法的有效性和实用性。针对领域专家提出的建议，后续还需要增加交互设计，改善颜色映射，提供多种时间单位的话题热度分析。系统目前是基于历史数据进行分析的，未来工作可以对此继续优化，通过采集实时推特数据来分析实时的突发事件模型在时间和空间上的演变。

参 考 文 献

[1] DWYER N, MARSH S. What can the hashtag # trust tell us about how users conceptualise trust? [C]// Twelfth International Conference on Privacy, Security and Trust. New York: IEEE Press, 2014: 398-402.

[2] ZAPPAVIGNA M. Discourse of Twitter and social media: How we use language to create affiliation on the web[M]. A&C Black, 2012.

[3] IVANOVA M. Understanding microblogging hashtags for learning enhancement[J]. Form, 2013, 11(74): 17-23.

[4] HUBERMAN B A, ROMERO D M, WU F. Social networks that matter: Twitter under the microscope[J]. arXiv: 0812.1045, 2008.

[5] KWAK H, LEE C, PARK H, et al. What is Twitter, a social network or a news media? [C]// Proceedings of the 19th International Conference on World Wide Web. New York: ACM Press, 2010: 591-600.

[6] YANG J, LESKOVEC J. Modeling information diffusion in implicit networks[C]// IEEE 10th International Conference on Data Mining (ICDM), 2010. New York: IEEE Press, 2010: 599-608.

[7] LERMAN K, GHOSH R. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks [C]// Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM). Menlo Park, CA: AAAI Press, 2010: 90-97.

[8] ROMERO D M，MEEDER B，KLEINBERG J. Differences in the mechanics of information diffusion across topics：idioms，political hashtags，and complex contagion on twitter[C]// Proceedings of the 20th International Conference on World Wide Web. New York：ACM Press，2011：695-704.

[9] CUNHA E，MAGNO G，COMARELA G，et al. Analyzing the Dynamic Evolution of Hashtags on Twitter：a Language-Based Approach[C]// Workshop on Languages in Social Media. Association for Computational Linguistics，2011.

[10] MACEACHREN A M，JAISWAL A，ROBINSON A C，et al. Senseplace2：Geotwitter analytics support for situational awareness[C]// 2011 IEEE Conference on Visual Analytics Science and Technology（VAST）. New York：IEEE Press，2011：181-190.

[11] KONG S，MEI Q，FENG L，et al. Predicting bursts and popularity of hashtags in real-time[C]// Proceedings of the 37th international ACM SIGIR Conference on Research & Development in Information Retrieval. New York：ACM Press，2014：927-930.

[12] MA Z，SUN A，CONG G. Will this # hashtag be popular tomorrow? [C]// Proceedings of the 35th international ACM SIGIR Conference on Research and Development in Information Retrieval. New York：ACM Press，2012：1173-1174.

[13] TSUR O，RAPPOPORT A. What's in a hashtag?：content based prediction of the spread of ideas in microblogging communities [C]// Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. New York：ACM Press，2012：643-652.

[14] CHO I，WESSLEN R，VOLKOVA S，et al. CrystalBall：A Visual Analytic System for Future Event Discovery and Analysis from Social Media Data[C]// IEEE Conference on Visual Analytics Science and Technology（VAST）. New York：IEEE Press，2017：25-35.

[15] KAMATH K Y，CAVERLEE J，CHENG Z，et al. Spatial influence vs. community influence：modeling the global spread of social media[C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York：ACM Press，2012：962-971.

[16] HE J，CHEN C. Spatiotemporal Analytics of Topic Trajectory [C]// Proceedings of the 9th International Symposium on Visual Information Communication and Interaction. New York：ACM Press，2016：112-116.

[17] KAMATH K Y，CAVERLEE J. Spatio-temporal meme prediction：learning what hashtags will be popular where[C]// Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management. New York：ACM Press，2013：1341-1350.

[18] HONG L，AHMED A，GURUMURTHY S，et al. Discovering geographical topics in the twitter stream[C]// Proceedings of the 21st International Conference on World Wide Web. New York：ACM Press，2012：769-778.

[19] LU Y，WANG H，LANDIS S，et al. A visual analytics framework for identifying topic drivers in media events[J]. IEEE Transactions on Visualization and Computer Graphics，2018，24(9)：2501-2515.

[20] EL-ASSADY M，SPERRLE F，DEUSSEN O，et al. Visual Analytics for Topic Model Optimization based on User-Steerable Speculative Execution[J]. IEEE Transactions on Visualization and Computer Graphics，2019，25(4)：1-20.

[21] WU Y，CHEN Z，SUN G，et al. Streamexplorer：a multi-stage system for visually exploring events in social streams[J]. IEEE Transactions on Visualization and Computer Graphics，2018，24(10)：2758-2772.

[22] ANDRIENKO G，ANDRIENKO N，FUCHS G，et al. Revealing patterns and trends of mass mobility through spatial and temporal abstraction of origin-destination movement data[J]. IEEE Transactions on Visualization & Computer Graphics，IEEE，2017(1)：1.

[23] MARCUS A，BERNSTEIN M S，BADAR O，et al. Twitinfo：aggregating and visualizing microblogs for event exploration[C]// Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York：ACM Press，2011：227-236.

[24] CAO N，LIN Y-R，SUN X，et al. Whisper：Tracing the spatiotemporal process of information diffusion in real time[J]. IEEE Transactions on Visualization and Computer Graphics，2012，18(12)：2649-2658.