

面向网络舆情的无监督演化趋势评估方法

秦涛<sup>1</sup>, 王熙凤<sup>1</sup>, 沈壮<sup>1</sup>, 陈周国<sup>2</sup>, 丁建伟<sup>2</sup>

(1. 西安交通大学智能网络与网络安全教育部重点实验室, 710049, 西安;

2. 中国电子科技集团第三十研究所保密通信国家重点实验室, 610093, 成都)

**摘要:** 针对网络舆情突发性强、标注数据较少且管控资源有限等问题,提出了一种网络舆情演化趋势评估无监督学习算法,筛选演化趋势重要的舆情事件进行优先管控,提升网络监管的工作效率。针对舆情事件并发性强的特点,利用多指标排序算法对舆情演化趋势重要性评估问题进行形式化描述;针对舆情突发性强、难以获得大量标注数据的问题,利用主曲线排序算法对舆情演化趋势重要性评估问题进行建模,采用 3 阶贝塞尔曲线进行模型求解,充分利用评估指标中的顺序关系和数值关系;结合典型公开数据集和自主构建的舆情数据集对所提算法进行了验证分析,实验结果表明,所提算法可在无标注数据的情况下实现舆情事件演化态势重要性的评估,为资源有限情况下的舆情事件管控提供决策支撑。

**关键词:** 舆情事件管控;资源有限;无监督学习;贝塞尔曲线

**中图分类号:** TP393 **文献标志码:** A

**DOI:** 10.7652/xjtuxb202011014 **文章编号:** 0253-987X(2020)11-0113-08



OSID 码

Research on Unsupervised Method for Changing Trend Evaluation of Internet Public Opinion Events

QIN Tao<sup>1</sup>, WANG Xifeng<sup>1</sup>, SHEN Zhuang<sup>1</sup>, CHEN Zhouguo<sup>2</sup>, DING Jianwei<sup>2</sup>

(1. MOE Key Lab for Intelligent and Network Security, Xi'an Jiaotong University, Xi'an 710049, China;

2. Science and Technology on Communication Security Lab, 30th Research Institute of China Electronics Technology Group Corporation, Chengdu 610093, China)

**Abstract:** Aiming at the abruptness posed by public opinion emergencies, limited labeled data and management resources, we proposed an unsupervised algorithm for evaluating the importance of public opinion changing trend and controlling the important public opinion events within limited time and reducing their harmful influence to the society. Firstly, the importance of changing trend evaluation problem is transformed into a multi-index ranking based on the management experience. Secondly, to solve the problems posed by limited labeled data, we use the principal curve algorithm to formulate the problem of changing trend evaluation, and use the third-order Bessel curve to obtain the final ranking results. The designed model can fully capture the structural and value characteristics of the original data. Finally, we employ the typical public data set and the self-built public opinion event data set to verify the proposed method. The experimental results verify that the developed method has high efficiency in public opinion event changing trend ranking without prior knowledge, and provides a decision support for public opinion event management with limited resources.

**Keywords:** public opinion management; limited resources; unsupervised learning; Bessel curve

由于网络的开放性和匿名性,各大社交平台已成为网络舆情事件滋生和传播的温床,给国家和社会稳定带来了严重的危害,舆情管控已成为迫切需要重视的问题<sup>[1]</sup>。受到人力、物力等资源的限制,无法实现对同时发生的所有舆情事件进行管理。本文设计了一种网络舆情演化趋势重要性评估方法,筛选重要舆情事件作为管控对象,在管理人员时间和精力有限的情况下降低网络舆情事件的危害。

近年来,网络舆情事件管理和控制受到了大量研究者的关注,国内外学者围绕舆情事件重要性评估做了大量研究。高承实等采用德尔菲法和层次分析法来评估舆情事件安全演化趋势<sup>[2]</sup>;戴媛等在层次分析法的基础上结合模糊数学理论建立安全评估模型<sup>[3]</sup>;程华等采用 DS 证据理论实现舆情重要性评估中的多指标融合算法<sup>[4]</sup>,但是上述算法多依赖于专家知识和人工经验,可扩展性和泛化性较差,无法处理无标注、无先验知识情况下的舆情事件重要性评估。

网络舆情演化态势的评估也可以利用舆情事件的主题、参与者和所构建的事件传播网络进行。黄星等利用加速遗传算法对舆情风险等级评价<sup>[5]</sup>;杨静等利用动态贝叶斯网络建立舆情风险等级划分模型,并根据分级结果对舆情事件采取相应的措施<sup>[6]</sup>,但是这些风险等级划分算法往往只关注舆情事件的某一方面,未能全面考虑造成舆情态势恶化的多个因素。在事件参与者方面,Gong 等提出了一种基于结构洞理论的网络舆情控制方法,利用意见领袖和网民之间的连接特征控制舆情的传播<sup>[7]</sup>;Gu 等综合考虑节点之间的共同好友和交互关系构建了舆情态势评估模型实现舆情态势评估<sup>[8]</sup>,但是这些传播网络和好友关系往往只适合于特定事件,事件的不同将导致传播网络的变化。在演化趋势方面,Qiu 等利用时间序列预测模型对事件变化态势进行预测<sup>[9]</sup>,但是这些预测方法往往需要较多的先验知识。

多指标排序模型可以应用于多舆情事件态势评估中,亦即根据一定的评估指标,对事件的演化态势进行排序。Aslam 等利用 Borda Count 元搜索模型,并结合贝叶斯推理设计了更为高效的排序算法<sup>[10]</sup>;Young 等设计了近似最优化规则的 Condorcet 方法实现数据排序<sup>[11]</sup>。这些排序方法可以应用于舆情重要性评估中,但是这些排序方法仅考虑各指标内的次序关系,而忽略数值之间的隐式关系,在

应对大规模、突发性强的舆情演化趋势评估需求时存在一定的不足。

针对上述问题,本文将趋势重要性评估问题转化为多指标排序问题,设计了一种无监督的网络舆情演化趋势重要性评估方法,在无标注数据的情况下实现舆情演化趋势重要性评估。针对舆情事件的标注数据较少的特点,本文设计了一种基于主曲线理论的排序学习算法,采用 3 阶贝塞尔曲线对数据进行排序,综合考虑数据特征之间的结构关系和数值隐式关系,最大限度地求取数据集的近似全序。采用包括 QS 2019 世界大学排行榜(QS2019)<sup>[12]</sup>和 Quality of Life Country 2019 国家生活质量排行榜(QLC2019)<sup>[20]</sup>的有标注公开数据集作为基准评估所提算法的性能,采用由 10 个典型舆情事件构成的舆情数据集评估所提算法的实用性。实验结果表明,本文所提算法效果较好且具有实用性,可为资源有限情况下的舆情管控提供决策支撑。

## 1 多指标舆情演化趋势评估建模

### 1.1 问题描述

网络舆情事件的重要性评估可以转化为多指标排序问题,即根据所设计的指标体系,对当前发生的舆情事件进行重要性排序。

在多指标排序模型中,如果指标体系具有严格的单调性,则可以根据指标的量化结果直接进行排序,但是实际环境中根据不同的指标得到的排序结果往往是不同的。例如,在 2020 年 QS 世界大学排行榜中<sup>[12]</sup>,根据教师/学生比例(Faculty/Student Ratio)指标,斯坦福大学得分为 98.6 分,哈佛大学得分为 99.6 分,按照该指标进行排序,斯坦福大学排在哈佛大学之后;而根据国际教师比率(International Faculty Ratio)指标,斯坦福大学得分为 98.4 分,哈佛大学得分为 97.4 分,按照该指标进行排序,斯坦福大学排在哈佛大学之前。由不同指标得到的排序结果相互矛盾,无法得到唯一的排序结果,这也是多指标排序面临的难点问题之一。

### 1.2 问题建模

多指标排序结果相互矛盾的根源是数据集存在噪声,亦即不存在全序关系。我们用  $\epsilon$  表示数据噪声。对于带有噪声  $\epsilon$  的数据集  $\mathbf{D}$ ,无法直接提取其中存在的全序关系,但是可以挖掘与数据分布最接近且含有全序关系的数据集  $\mathbf{D}'$ ,即  $\mathbf{D}' = \mathbf{D} - \epsilon$ 。对

数据集  $D$  进行排序就可以转化为求取数据集  $D'$  的全序关系,以获得最优的符合数据集  $D$  特征的排序结果。

设排序函数为  $\varphi$ ,排序评分为  $s$ ,则  $s=\varphi(D',\theta)=\varphi(D-\varepsilon,\theta)$ 。由于排序函数需满足单调性,即  $\varphi$  存在逆函数  $f(s,\theta)$ ,满足

$$\varepsilon = D - D' = D - f(s,\theta)$$

(1)

在几何上, $f(s,\theta)$ 可以视作数据集  $D$  在一维流形空间并满足单调性要求的主曲线。为使所得到的排序尽可能地反映原始数据集的特征,则需最小化  $D$  与  $D'$  间的噪声  $\varepsilon$ ,据此舆情演化态势重要性排序问题可以形式化描述为

$$\left. \begin{aligned} \min J(\theta) &= \sum_{i=1}^n \|x_i - f(s_i, \theta)\|^2 \\ \text{s. t. } &\left(\frac{\partial f(s_i, \theta)}{\partial s_i}\right)^T (x_i - f(s_i, \theta)) = \mathbf{0}, \\ &i = 1, 2, \dots, n \end{aligned} \right\} \quad (2)$$

其中  $f(s_i,\theta)$ 是评估指标  $x_i$  在主曲线上的投影,优化目标为所有数据  $x_i$  与投影点  $f(s_i,\theta)$ 间距离的均方误差之和最小。上述优化目标是排序主曲线算法的一般形式,但是从任意函数族中选取符合条件的主曲线显然是一个 NP 难问题<sup>[13]</sup>。

1.3 方法框架

为实现无标注数据下的舆情态势评估,结合上述分析,本文所提出的无监督舆情演化态势重要性评估方法整体框架如图 1 所示,包括以下 3 个步骤。

步骤 1:原始特征提取。结合舆情态势管控需求,从原始舆情信息中抽取包含人、事、势 3 方面的评估指标,并对指标进行归一化处理,克服不同量纲带来的差异。

步骤 2:根据所确定的评估指标,采用主曲线排序算法对舆情数据集进行排序,充分利用数据点之

间的隐含关系和位置关系,通过不断迭代优化,获得最终排序结果。

步骤 3:舆情事件管控。根据输出的评估值,获得最终的重要性评估结果,结合实际需求,实现重要舆情事件的管控,降低舆情事件的危害。

2 层次化评估指标体系构建

构建指标体系是多指标排序任务的基础,也是舆情演化态势评估结果可靠性的根本。国内研究者围绕舆情评估指标体系进行了初步的研究,例如冯江平等从政府、网媒、网民 3 个网络舆情主体出发,构建了包括政府舆情应对能力、网媒舆情热度、网民心理特征在内的网络舆情评价指标体系<sup>[14]</sup>;张一文等针对突发舆情事件的评估需求,构建了包括舆情产生导火索、舆情产生主体、舆情产生载体、舆情调控主体的网络舆情热度评价指标体系<sup>[15]</sup>,但是上述指标体系往往聚焦于舆情事件的某个方面,无法全面反映舆情事件的整体状态。

本文从舆情事件的整体出发,由舆情事件的人、事、势 3 个要素构建态势评估指标体系。其中人是指舆情事件的参与者,是舆情事件形成的主要因素,这部分指标主要度量舆情事件中的参与者的特征,例如参与者的年龄分布、地域分布等;事是指舆情事件的主题和具体的分面,这是判断事件是否为舆情事件的根本,这部分指标主要关注舆情事件的基本特征和规律,例如事件主题的敏感程度、事件出现的时间段等;势是指舆情事件传播和演化的态势,主要反映舆情事件在时间维度的动态变化和演变特征,是判断具体舆情事件是否成为管控对象的根本,例如舆情事件在时间维度呈现的传播飙升趋势、话题敏感程度变化趋势等。

从以上 3 个因素出发,本文构建了涵盖 14 个具体指标的评估指标体系,既涵盖了舆情事件的参与

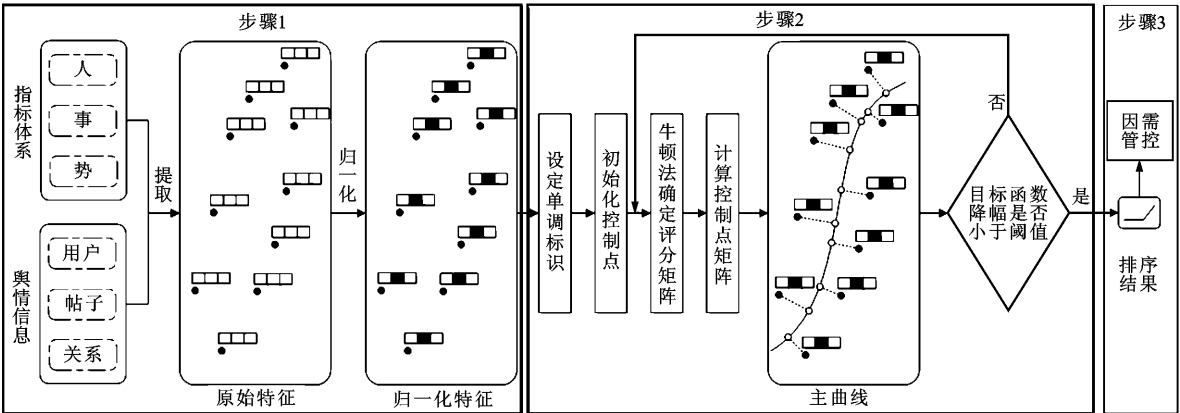


图 1 无监督舆情演化态势重要性评估方法整体框架图

者,也涵盖了舆情事件的主题及其在时间维度的演化特征,可以从整体上刻画舆情事件的特征,为评估结果的可靠性奠定基础。

同时,为了克服指标量纲不同带来的影响,对抽取的特征进行了归一化处理,表达式如下

$$d'_i = \frac{d_i - d_i^{\min}}{d_i^{\max} - d_i^{\min}} \quad (3)$$

式中: $d_i$  是舆情某个评估指标的特征值; $d'_i$  是归一化后的特征值; $d_i^{\min}$  和  $d_i^{\max}$  分别是该评估特征的最小值和最大值。这样就可以在序关系不变的情况下把原始数据的特征归一化到 $[0,1]$ 范围内。

### 3 无监督排序模型构建

#### 3.1 贝塞尔曲线

舆情态势评估可以转换为如式(2)中所示的优化问题,采用贝塞尔曲线对上述问题进行建模求解。贝塞尔曲线<sup>[16]</sup>是一种可以参数化表示的光滑曲线,具有凸包性、参数放缩不变性、可微性、仿射变换不变性等特性,在计算机图形的矢量画图中发挥了重要的作用。

一条  $k$  阶的贝塞尔曲线的参数化表示形式为

$$\left. \begin{aligned} f(s_i) &= \sum_{r=0}^k B_r^k(s_i) p_r, \quad s_i \in [0,1] \\ B_r^k(s_i) &= \binom{k}{r} (1-s_i)^{k-r} s_i^r \\ i &= 1, 2, \dots, n \\ \binom{k}{r} &= \begin{cases} \frac{k!}{r!(k-r)!}, & 0 \leq r \leq k \\ 0, & \text{其他} \end{cases} \end{aligned} \right\} \quad (4)$$

式中: $p_r \in R_d$ ,  $p_r$  是贝塞尔曲线上的控制点,  $r \in [0, k]$ , 决定了贝塞尔曲线的基本形状,  $p_0$  和  $p_k$  分别是贝塞尔曲线的起点和终点;  $s_i$  是在  $[0,1]$  范围内的实数; 插值函数  $B_r^k(s_i)$  是伯恩斯坦多项式。由此,贝塞尔曲线可以抽象为一条开始于  $p_0$  ( $s_i=0$ ) 并结束于  $p_k$  ( $s_i=1$ ) 的非线性插值曲线, 曲线上各点的坐标是由  $s$  作为自变量计算得到的伯恩斯坦值与控制点矩阵相乘的结果。

#### 3.2 无监督排序学习算法

本文利用 3 阶贝塞尔曲线 ( $k=3$ ) 构建无监督排序学习模型。根据贝塞尔曲线的性质<sup>[17]</sup>, 3 阶贝塞尔曲线可以在评估指标归一化处理后不破坏原始数据集中的序关系; Li 等也已证明对于任意多维特征向量的数据集, 可以表征其结构的 3 阶贝塞尔曲线一定存在<sup>[18]</sup>, 因此由 3 阶贝塞尔曲线构建无监督排

序算法是可行的。

3 阶贝塞尔曲线对应的非线性约束优化问题可以形式化为

$$\begin{aligned} \min C(\mathbf{P}, \mathbf{s}) &= \sum_{i=1}^n \|x_i - \mathbf{PBS}_i\|^2 \\ \text{s. t.} \quad &\left(\frac{\partial \mathbf{PBS}}{\partial s_i}\right)^T (x_i - \mathbf{PBS}) = \mathbf{0} \end{aligned} \quad (5)$$

式中:  $\mathbf{s} = ((s_1, s_2, \dots, s_n))$ ,  $\mathbf{P} = (p_0, p_1, p_2, p_3)$ ,  $\mathbf{S}_i = (1, s_i, s_i^2, s_i^3)^T$ ,  $s_i \in [0, 1]$ ,  $\mathbf{B} = \begin{bmatrix} 1 & -3 & 3 & -1 \\ 0 & 3 & -6 & 3 \\ 0 & 0 & -3 & 3 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ ,  $\mathbf{S} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ s_1 & s_2 & \dots & s_n \\ s_1^2 & s_2^2 & \dots & s_n^2 \\ s_1^3 & s_2^3 & \dots & s_n^3 \end{bmatrix}$ ,  $i=1, 2, \dots, n$ 。

通过期望最大化(EM)算法<sup>[19]</sup>可以求解上述优化问题稳定收敛后的近似解。首先对控制点进行初始化, 其中起点和终点分别为  $p_0, p_3$ ,  $p_0 = 1/2(1-\alpha)$ ,  $p_3 = 1/2(1+\alpha)$ , 另外两个控制点在数据分布内随机选取, 单调标识  $\alpha$  根据数据集各维度特征的物理含义设定。由式(5)得到第  $t$  步迭代时

$$\left(\frac{\partial \mathbf{P}^{(t)} \mathbf{BS}^{(t)}}{\partial s_i^{(t)}}\right)^T (x_i - \mathbf{P}^{(t)} \mathbf{BS}^{(t)}) = 0 \quad (6)$$

将  $\mathbf{P}^{(t)}$  代入式(6)可得到关于评分向量  $\mathbf{S}^{(t)}$  的一元五次方程, 利用牛顿法可以求取  $\mathbf{S}^{(t)}$  的近似解, 以此更新  $\mathbf{S}$ 。

第  $t+1$  步迭代时,  $\mathbf{P}^{(t+1)}$  为式(5)的最优解, 即

$$\mathbf{P}^{(t+1)} = \operatorname{argmin} \sum_{i=1}^n \|x_i - \mathbf{P}^{(t)} \mathbf{BS}_i^{(t)}\|^2 \quad (7)$$

排序学习算法的优化目标是最小化目标函数  $C(\mathbf{P}, \mathbf{s})$ , 迭代停止条件为某步迭代中  $C(\mathbf{P}, \mathbf{s})$  不再下降或者下降幅度小于某一阈值  $\zeta$ , 表达式如下

$$C(\mathbf{P}^{(t+1)}, \mathbf{s}^{(t+1)}) - C(\mathbf{P}^{(t)}, \mathbf{s}^{(t)}) < \zeta \quad (8)$$

由贝塞尔曲线的存在性可得该优化过程将在有限时间内收敛, 计算得到的  $(\mathbf{P}^{(t+1)}, \mathbf{s}^{(t+1)})$  即使目标函数  $C$  达到最小时的控制点和相应的排序评分序列。

综合上述分析, 基于 3 阶贝塞尔曲线的排序算法的求解过程的伪代码描述如下: 首先将数据特征向量的每个分量都归一化到  $[0,1]$  上, 并依据设定的数据集的单调标识对控制点  $\mathbf{P}^{(0)}$  进行初始化; 随后, 遵循期望最大化算法进行迭代求解, 在第  $t$  步迭代时, 利用牛顿法由  $\mathbf{P}^{(t)}$  确定相关性评分向量  $\mathbf{S}^{(t)}$ , 将  $\mathbf{S}^{(t)}$  代入优化函数从而得到  $\mathbf{P}^{(t+1)}$ , 在此过程中控制点的位置需要约束在  $m$  维空间的超立方体  $[0,1]^m$

内<sup>[17]</sup>;最后,当某步迭代的目标函数值下降幅度小于设定的阈值  $\zeta$  时,迭代结束,获得模型的解。

## 4 模型初始化与性能评估

### 4.1 实验数据

本文利用两种数据集对所提算法进行验证分析,首先利用公开有标注的数据集作为基准,对算法的性能进行验证分析;随后利用自主构建的舆情数据集对所提算法的实用性进行验证。

公开数据集包括 QS 2019 世界大学排行榜和 QLC 2019 国家生活质量排行榜。其中 QS 2019 世界大学排行榜是对全球 500 余所大学进行综合排名,所采用的 6 个指标为学术声誉、雇主声誉、每个教师的论文引用数量、教师/学生比例、国际教师比率以及国际学生比率。QLC2019 是根据对超过 120 000 位来自美洲、亚洲、欧洲和非洲的公民的调查结果,对 70 个国家生活质量进行排名的数据集,所采用的 8 个评价指标为房价收入比、购买力指数、通勤时间指数、污染指数、气候指数、人身安全指数、医疗保健服务指数和生活成本指数。

舆情数据集由采集于新浪微博中典型舆情事件的帖子构成,结合微博热搜榜和爬取时间段,选择特定主题舆情事件,通过设置关键词仅收集此主题下特定时间范围内的帖子。在 2019 年 6 月 7 日~2019 年 6 月 14 日期间,共采集了 10 个典型舆情事件的帖子 43 042 条,事件主题和事件概述如表 1 所示。所选择的事件主题是一周时间内发生的典型舆情事件,涵盖了政治、民生、娱乐等多种类型的舆情事件,具有一定的代表性。

表 1 舆情事件数据集

| 序号 | 事件主题          | 事件概述                    | 帖子数    |
|----|---------------|-------------------------|--------|
| 1  | 中美贸易战中<br>的华为 | 华为公司在中美贸易战<br>中被美国制裁、打压 | 10 758 |
| 2  | 赵志勇强奸<br>杀人案  | “特大强迫卖淫案”强奸犯<br>赵志勇被处死刑 | 3 857  |
| 3  | 高考公平性         | 高考泄题、漏题等假信息             | 4 428  |
| 4  | 林志玲结婚         | 娱乐明星林志玲结婚               | 4 038  |
| 5  | 王源学盛饭         | 娱乐明星王源学盛饭               | 4 611  |
| 6  | 孙宇晨炒作         | 孙宇晨炒作比特币                | 2 510  |
| 7  | NBA 决赛        | NBA 总决赛相关信息             | 4 084  |
| 8  | 俄部署导弹         | 俄部署导弹疑似针对我国             | 1 734  |
| 9  | 扫黑除恶          | 打击黑恶势力的专项活动             | 3 487  |
| 10 | 百度            | 百度人事变动,股价大跌             | 3 535  |

由表 1 可以看出,关于中美贸易战话题的帖子数最多,关于俄部署导弹的帖子数最少。如按照帖子数多少进行管控优先级排序,则王源学盛饭、NBA 总决赛等娱乐事件将具有较高的管控优先级,这显然和舆情实际管控需求不相符,为此必须研究舆情事件重要性评估方法。

首先利用时间片划分方法,将舆情事件数据以天为单位进行划分,从每天的数据中抽取第 2 节所设计的 14 个评估指标,据此对所收集的 10 个舆情事件重要性进行排序,动态量化一周时间内舆情事件的重要性和管控需求。

### 4.2 算法性能评价标准

本文利用通用的归一化折损累计增益( $N_k$ )来评价算法的性能<sup>[21]</sup>,其定义如下

$$N_k = \frac{G_k}{I_k} = \frac{\sum_{i=1}^k (2^{r_i} - 1) / \lg(i + 1)}{\sum_{i=1}^{|R|} (2^{r_i} - 1) / \lg(i + 1)} \quad (9)$$

式中: $r_i$  代表排序结果中第  $i$  个位置上数据的相关度; $G_k$  代表排序结果的折损累计增益值,其中引入  $\frac{1}{\lg(i+1)}$  作为位置信息对总价值带来的折损,越靠后的数据的价值越低; $I_k$  表示排序的最佳结果的折损累计增益。利用  $I_k$  对  $G_k$  进行归一化以实现不同长度的数据列表间的统一比较。

### 4.3 实验参数设置

贝塞尔曲线的初始化参数  $\alpha$  需要根据数据集排序指标物理意义与排名的正负相关性确定,排序指标与排名正相关,则单调性标志对应维度设为 1,反之则为 -1。例如 QLC2019 的评价指标中,国家的房价收入比与国家排名负相关,则对应维度设为 -1,购买力指数与排名正相关,则对应维度设为 1。

针对 QS2019 数据集,根据数据集中待排名的高校数  $N$ ,确定相关度分级  $K$ ,  $K = N/10$ ,此时各相关度等级上平均高校数约为 10 个。根据数据集特征各维的物理含义,设  $\alpha = [1, 1, 1, 1, 1, 1]$ ,则贝塞尔曲线的起点  $p_0 = \frac{1-\alpha}{2} = [0, 0, 0, 0, 0, 0]$ ,终点  $p_3 = \frac{1+\alpha}{2} = [1, 1, 1, 1, 1, 1]$ 。

针对 QLC2019 数据集,根据数据集中待排名的国家数,确定相关度分级  $K$ ,  $K = N/5$ ,此时各相关度等级上平均国家数约为 5 个。根据数据集各评估指标的物理含义,设单调性标志  $\alpha = [-1, 1, -1, -1, 1, 1, 1, -1]$ ,则起点  $p_0 = \frac{1-\alpha}{2} = [1, 0, 1, 1, 0,$

$0,0,1]$ , 终点  $p_3=\frac{1+\alpha}{2}=[0,1,0,0,1,1,1,0]$ 。

针对舆情数据集,根据数据集各评估指标的物理含义,设  $\alpha=[1,1,1,1,1,1,1,1,1,1,1,1,1]$ , 则起点  $p_0=\frac{1-\alpha}{2}=[0,0,0,0,0,0,0,0,0,0,0,0,0]$ , 终点  $p_3=\frac{1+\alpha}{2}=[1,1,1,1,1,1,1,1,1,1,1,1,1]$ 。

### 5 模型测试与结果分析

本文利用 Borda<sup>[10]</sup> 和 Condocret<sup>[11]</sup> 作为对比基准算法,衡量所提算法的性能。Borda 算法是一种无监督多序列融合的排序方法,它依据各特征分别计算数据集的排序,在每个特征的排序序列中,以数据的位次进行评分,所有的排序位次评分之和作为总的排序评分。Condocret 算法也是一种无监督多序列融合的排序方法,与 Borda 算法不同的是,它不再以数据在每个特征上的排序位次进行评分,而是采用两两比较的形式,统计一对数据在各排序序列中的位次先后关系,拥有更多优先关系的数据最终排序靠前。

#### 5.1 公开数据集结果分析

QS2019 大学排行数据集上的实验结果如表 2 所示,实验结果表明数据集规模  $q$  在 50~500 之间,本文算法均可以在  $N_k$  指标上取得更好的表现,比 Borda 算法有 4.43%~15.96% 的提升,比 Condocret 算法有 0.69%~11.39% 的提升。结合 QLC 2019 数据集上的实验结果如表 3 所示,实验结果表明数据集规模在 10~70 之间,所提方法均可以在  $N_k$  指标上取得更好的表现,比 Borda 算法有 4.43%~15.96% 的提升,比 Condocret 算法有 6.19%~44.56% 的提升。

表 2 QS2019 大学排行数据集上的实验结果

| 算法        | $N_k$  |         |         |         |         |
|-----------|--------|---------|---------|---------|---------|
|           | $q=50$ | $q=100$ | $q=150$ | $q=200$ | $q=250$ |
| 本文算法      | 0.99   | 0.95    | 0.98    | 0.98    | 0.99    |
| Borda     | 0.93   | 0.90    | 0.87    | 0.85    | 0.94    |
| Condocret | 0.94   | 0.95    | 0.89    | 0.87    | 0.95    |

| 算法        | $N_k$   |         |         |         |         |
|-----------|---------|---------|---------|---------|---------|
|           | $q=300$ | $q=350$ | $q=400$ | $q=450$ | $q=500$ |
| 本文算法      | 0.98    | 0.99    | 0.99    | 0.99    | 0.99    |
| Borda     | 0.82    | 0.90    | 0.95    | 0.94    | 0.95    |
| Condocret | 0.97    | 0.95    | 0.95    | 0.97    | 0.98    |

相比 Borda 和 Condocret 算法,本文所提算法在多指标排序上取得较好的结果。Borda 和 Condocret 算法在求取数据排序结果时,都只考虑了数据各维度指标的排序序列,实质上是各特征排序结果的融合算法,但是并没有考虑到不同指标间和同一指标内的绝对数值关系。本文所提算法从数据结构出发,利用主曲线的性质,在优化迭代时可以自适应地学习到各指标的数值关系,在保证全序性的条件下,最大限度地求取数据集的近似全序,以此作为最终的排序结果。

表 3 QLC2019 数据集上的实验结果

| 算法        | $N_k$  |        |        |        |
|-----------|--------|--------|--------|--------|
|           | $q=10$ | $q=20$ | $q=30$ | $q=40$ |
| 本文算法      | 0.81   | 0.86   | 0.94   | 0.88   |
| Borda     | 0.69   | 0.81   | 0.88   | 0.84   |
| Condocret | 0.36   | 0.79   | 0.88   | 0.79   |

| 算法        | $N_k$  |        |        |
|-----------|--------|--------|--------|
|           | $q=50$ | $q=60$ | $q=70$ |
| 本文算法      | 0.93   | 0.88   | 0.88   |
| Borda     | 0.87   | 0.78   | 0.79   |
| Condocret | 0.87   | 0.78   | 0.78   |

#### 5.2 舆情数据集结果分析

将本文所提算法应用于自主构建的舆情数据集,图 2 给出了 2019 年 6 月 7 日~2019 年 6 月 13 日间典型舆情事件演化趋势重要性的演变趋势评估结果。

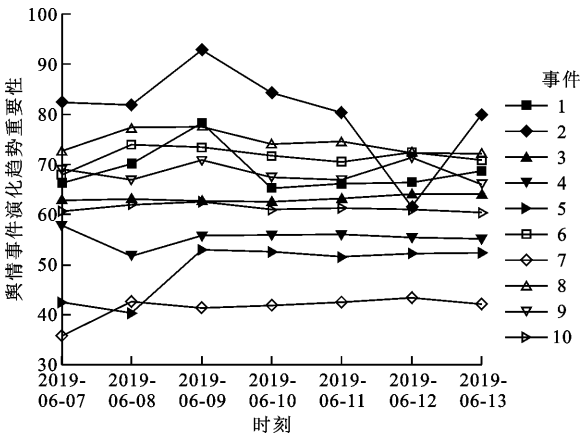


图 2 舆情演化趋势重要性的演变趋势评估结果

由图 2 可以看出,以赵志勇强奸杀人、中美贸易战中的华为等为主题的事件演化趋势较为严重,事件主题较为敏感,需要管控。从事件本身来看,赵志勇强奸杀人事件中因为赵志勇恶劣的行径被判处死

刑,而华为公司因在贸易战中被美国封杀,这两个事件都激起了网民强烈的反应,网络言论较为激烈且热度较高,导致事件演化趋势评估分较高,需要进行适当的管控,分析结果符合实际舆情管控需求。演化趋势重要性评估较低的事件中,例如以林志玲结婚、王源学盛饭等为主题的事件是娱乐事件,虽然参与人很多,但是网民情绪普遍冷静和中立,事件的发展速度也不快,管控需求并不大,也符合实际舆情管控需求。

在公开数据集上的分析结果验证了本文所提算法的效果较好,表明本文所提算法可以在无专家知识或标注数据的情况下,根据一定的评估指标体系实现研究对象的排序评估。在舆情数据集上的分析结果验证了本文所提算法的实际应用可行性,表明本文所提算法可以从多个并发的舆情事件中筛选急需管控的舆情事件,为资源有限的多舆情事件管控提供决策支持。

## 6 结 论

针对舆情事件管控资源受限且难以获取大量有标签数据的难题,本文设计了一种基于排序主曲线模型的舆情演化趋势无监督评估算法,利用 3 阶贝塞尔曲线进行建模求解。在评估指标设计方面,本文从人、事、势舆情 3 个要素出发,构建了全面反映舆情演化的指标体系;在评估模型设计方面,所提算法能够综合利用数据各评估指标的数值隐式关系和结构关系,最大限度的求取数据集的近似全序。结合典型公开数据集和自主构建舆情数据集的实验结果表明,所提的算法相比于其他算法具有较好的性能,能够满足真实网络环境下的舆情管控需求。

## 参考文献:

- [1] 中国互联网络信息中心. 第 45 次中国互联网络发展状况统计报告 [R/OL]. (2020-04-28) [2020-05-01]. <http://www.cnnic.net.cn/hlwfzyj/hlwzxbg/hlwjtjb/202004/P020200428596599037028.pdf>.
- [2] 高承实, 荣星, 陈越. 微博舆情监测指标体系研究 [J]. 情报杂志, 2011, 30(9): 66-70.  
GAO Chengshi, RONG Xing, CHEN Yue. Research on public opinion monitoring index-system in micro-blogging [J]. Journal of Intelligence, 2011, 30(9): 66-70.
- [3] 戴媛, 郝晓伟, 郭岩, 等. 我国网络舆情安全评估指标体系的构建研究 [J]. 信息安全, 2010, 4: 14-17.
- DAI Yuan, HAO Xiaowei, GUO Yan, et al. Research on the construction of the evaluation index system of network public opinion security in China [J]. Netinfo Security, 2010, 4: 14-17.
- [4] 程华, 杜思伟, 徐萃华, 等. 基于 DS 证据的信息融合算法多指标融合 [J]. 华东理工大学学报(自然科学版), 2011, 4(18): 100-103.  
CHENG Hua, DU Siwei, XU Cuihua, et al. A DS-based multi-index fusion of information fusion algorithm [J]. Journal of East China University of Science and Technology, 2011, 4(18): 100-103.
- [5] 黄星, 刘樑. 突发事件网络舆情风险评价方法及应用 [J]. 情报科学, 2018, 36(4): 3-9.  
HUANG Xing, LIU Liang. The evaluation method and application of unexpected events network public opinion [J]. Information Science, 2018, 36(4): 3-9.
- [6] 杨静, 邹梅, 黄薇. 基于动态贝叶斯网络的网络舆情危机等级预测模型 [J]. 情报科学, 2019, 37(5): 92-97.  
YANG Jing, ZOU Mei, HUANG Wei. The prediction model of network public opinion crisis level based on dynamic Bayes network [J]. Information Science, 2019, 37(5): 92-97.
- [7] GONG Cheng, DU Yajun, CHEN Fei. A structural hole-based approach to control public opinion [C] // Proceedings of the 2019 5th IEEE International Conference on Computer and Communications. Piscataway, NJ, USA: IEEE, 2019: 1423-1429.
- [8] GU Yiran, ZHANG Bingyan. Social network public opinion evolution model based on node intimacy [C] // Proceedings of the 2019 Chinese Control and Decision Conference. Piscataway, NJ, USA: IEEE, 2019: 4496-4500.
- [9] QIU Qizhi, HE Jinchao, CHEN Hui, et al. Research on the evolution law of emergency network public opinion [C] // Proceedings of the 2019 12th International Symposium on Computational Intelligence and Design. Piscataway, NJ, USA: IEEE, 2019: 157-161.
- [10] ASLAM J A, MONTAGUE M. Models for meta-search [C] // Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2001: 276-284.
- [11] YOUNG H P. Condorcet's theory of voting [J]. American Political Science Review, 1988, 82(4): 1231-1244.
- [12] QUACQUARELLI S. QS world university rankings

- [EB/OL]. (2019-02-27) [2020-06-07]. <https://www.qschina.cn/university-rankings/world-university-rankings/2019>.
- [13] 张军平, 王珏. 主曲线研究综述 [J]. 计算机学报, 2003, 26(2): 1-18.  
ZHANG Junping, WANG Jue. An overview of principal curves [J]. Chinese Journal of Computers, 2003, 26(2): 1-18.
- [14] 冯江平, 张月, 赵舒贞, 等. 网络舆情评价指标体系的构建与应用 [J]. 云南师范大学学报(哲学社会科学版), 2014, 46(2): 75-84.  
FENG Jiangping, ZHANG Yue, ZHAO Shuzhen, et al. Construction and application of the evaluation system of internet public opinion [J]. Journal of Yunnan Normal University (Humanities and Social Sciences), 2014, 46(2): 75-84.
- [15] 张一文, 齐佳音, 方滨兴, 等. 基于贝叶斯网络建模的非常规危机事件网络舆情预警研究 [J]. 图书情报工作, 2012, 56(2): 76-81.  
ZHANG Yiwen, QI Jiayin, FANG Binxing, et al. Online public opinion risk warning based on Bayesian network modeling [J]. Library and Information Service, 2012, 56(2): 76-81.
- [16] FARIN G. Curves and surfaces for computer-aided geometric design: a practical guide [M]. San Francisco, USA: Morgan Kaufmann Publishers, 1988: 1-334.
- [17] HU B, MANN G, GOSINE R. Control curve design for nonlinear (or fuzzy) proportional actions using spline-based functions [J]. Automatica, 1998, 34(9): 1125-1133.
- [18] LI Chunguo, MEI Xing, HU Baogang. Unsupervised ranking of multi-attribute objects based on principal curves [J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(12): 3404-3416.
- [19] MCLACHLAN G J, KRISHNAN T. The EM algorithm and extensions [M]. Hoboken, NJ, USA: John Wiley and Sons, 2008: 1-400.
- [20] US News and World Report L. P. Quality of life country 2019 [EB/OL]. (2019-01-23) [2020-03-07]. <https://www.usnews.com/news/best-countries/quality-of-life-rankings>.
- [21] JÄRVELIN K, KEKÄLÄINEN J. IR evaluation methods for retrieving highly relevant documents [C] // Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2000: 41-48.

(编辑 武红江)

## (上接第 105 页)

- WANG Lijun, LI Qing, TONG Chaonan, et al. Overview of active disturbance rejection control for systems with time-delay [J]. Control Theory & Applications, 2013, 30(12): 1521-1533.
- [16] 王永帅, 陈增强, 孙明玮, 等. 一阶大时滞对象降阶自抗扰控制的鲁棒稳定性 [J]. 中国科学技术大学学报, 2019, 49(1): 55-62.  
WANG Yongshuai, CHEN Zengqiang, SUN Mingwei, et al. Robust stability of reduced-order linear ADRC for first-order plants with large time-delay [J]. Journal of University of Science and Technology of China, 2019, 49(1): 55-62.
- [17] HOFFMANN G, HUANG Haomiao, WASLANDER S, et al. Quadrotor helicopter flight dynamics and control: theory and experiment [C] // Proceedings of the 2007 AIAA Guidance, Navigation and Control Conference and Exhibit. Reston, VA, USA: AIAA, 2007: 1670-1689.
- [18] BHAT S P, BERNSTEIN D S. Finite-time stability of continuous autonomous systems [J]. SIAM Journal on Control and Optimization, 2000, 38(3): 751-766.
- [19] WANG Xinhua, LIN Hai. Design and frequency analysis of continuous finite-time-convergent differentiator [J]. Aerospace Science and Technology, 2012, 18(1): 69-78.
- [20] 王永帅, 陈增强, 孙明玮, 等. 一阶惯性大时滞系统 Smith 预估自抗扰控制 [J]. 智能系统学报, 2018, 13(4): 500-508.  
WANG Yongshuai, CHEN Zengqiang, SUN Mingwei, et al. Smith prediction and active disturbance rejection control for first-order inertial systems with long time-delay [J]. CAAI Transactions on Intelligent Systems, 2018, 13(4): 500-508.

(编辑 陶晴)