计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20200725
57(12):2490-2500,2020

基于排序学习的网络舆情演化趋势评估方法研究

秦 涛[1,2] 沈 壮[1,2] 刘 欢[1,2] 陈周国[3]
[1](智能网络与网络安全教育部重点实验室(西安交通大学) 西安 710049)
[2](西安交通大学电子与信息工程学部 西安 710049)
[3](中国电子科技集团第三十研究所 成都 610093)
(qin.tao@mail.xjtu.edu.cn)

Learning to Rank for Evolution Trend Evaluation of Online Public Opinion Events

Qin Tao[1,2], Shen Zhuang[1,2], Liu Huan[1,2], and Chen Zhouguo[3]
[1](*Key Laboratory for Intelligent Networks and Network Security* (*Xi'an Jiaotong University*), *Ministry of Education*, *Xi'an* 710049)
[2](*School of Electronic and Information Engineering*, *Xi'an Jiaotong University*, *Xi'an* 710049)
[3](*30th Research Institute of China Electronics Technology Group Corporation*, *Chengdu* 610093)

Abstract Public opinion events in social networks have a bearing on social harmony and stability. Analyzing the evolution trend of events so as to manage and control them is able to reduce the impact of malignant online public opinion. However, the lack of labelled data and the limited relevant resources makes the effective management of online public opinion challenging and complicated. To solve those difficulties, we propose a learning-to-rank algorithm for the events evolution trend evaluation, which makes full use of the expert knowledge in the labelled data and the correlation between labelled and unlabelled data to select important public opinion for management, in turn, improves the management efficiency. Firstly, based on the experiences and demands of public opinion management, we design a measurable, accessible and meaningful hierarchical index system, which is derived from the three most important factors of events, for evolution trend evaluation. Secondly, we build an evaluation model for evolution trend evaluation based on the graph convolutional network. Specifically, our model uses the local sensitive Hash algorithm to mine the structural information from the data node's neighborhood and generates the mixed feature of the data node and its neighbor. Finally, we design different loss functions for the labelled and unlabelled data respectively, in order to realize the comprehensive utilization of the expert knowledge in the labelled data and the spatial structure information in the unlabelled data. We verify the efficiency of the proposed model on public datasets MQ 2007-semi and MQ 2008-semi. We also build a real-world public opinion event dataset to verify the practicability and generalization of the proposed algorithm. The experimental results show that the proposed model can evaluate the public opinion event evolution trend with limited expert knowledge, and provide decision support for public opinion event management with limited resources.

收稿日期:2020-09-13;**修回日期**:2020-10-21
基金项目:国家重点研发计划项目(2016YFE0206700);国家自然科学基金创新研究群体(61721002);教育部创新团队(IRT_17R86);国家自然科学基金项目(61772411);陕西省自然科学基金项目(2020JQ-646);中央高校基本科研业务费专项资金(xjh012019026)
This work was supported by the National Key Research and Development Program of China (2016YFE0206700), the Innovative Research Group of the National Natural Science Foundation of China (61721002), the Innovation Research Team of Ministry of Education (IRT_17R86), the National Natural Science Foundation of China (61772411), the Natural Science Foundation of Shaanxi Province (2020JQ-646), and the Fundamental Research Funds for the Central Universities (xjh012019026).

Key words　online public opinion management and control; evolution trend evaluation; limited labelled data; learning to rank; graph convolutional network

摘　要　社交网络中的舆情事件关乎社会的和谐与稳定,分析事件的演化趋势并进行管控能够有效降低恶性舆情事件的影响.但是,高效的舆情管控却面临标注数据少、管控资源有限的难题,采用人机混合增强技术,充分利用少量标注样本中的专家知识,是建立舆情演化态势评估模型的可行思路之一.据此,提出一种基于排序学习的舆情事件演化趋势重要性评估算法,在模型训练过程中,充分利用标注数据中的专家知识以及有标签数据和无标签数据的关联关系,筛选重要舆情事件进行管控,提升管控资源利用效能.首先,结合舆情管控经验和需求,从“人”“事”“势”等三要素出发,构建易获取、可量化、有含义的舆情事件演化态势评估指标体系;其次,基于图卷积神经网络构建舆情演化趋势评估模型,利用局部敏感Hash算法挖掘数据的空间结构信息,并利用图卷积求取数据及其邻域的混合特征;最后,针对有标签数据和无标签数据设计相应的损失函数,实现标注数据中专家知识和无标注数据中空间结构信息的综合利用.在公开数据集MQ2007-semi和MQ2008-semi上验证了算法的有效性,在自主构建的舆情数据集上验证了算法的实用性和泛化性.实验结果表明,所提算法可以根据少量的专家知识或标注数据,实现网络舆情事件演化态势的评估,为资源有限条件下的舆情事件管控提供决策支撑.

关键词　舆情事件管控;演化趋势评估;标签数据缺失;排序学习;图卷积网络

中图法分类号　T393.02

各大社交平台已成为网络舆情事件滋生和传播的温床,给社会和谐稳定带来了严重的负面影响.网络舆情事件的引导和控制是减轻舆情事件负面影响的重要手段之一[1].但是,网络舆情事件涉及的主题复杂多样、并发性强,而受到人力、物力的限制,不能同时对所有的舆情事件进行管理;再者,由于舆情事件突发性强,很难获得大量与之相关的高质量标注数据,导致有监督学习方法训练建立的模型难以泛化,场景改变后效果退化严重.如何采用人机混合增强技术,充分利用少量有标签数据中的专家知识,建立舆情演化态势评估模型是可行思路之一.因此,在分析网络舆情事件特征和监管需求的基础上,本文提出一种神经网络排序模型,针对有标签数据和无标签数据设计相应的损失函数,在模型训练过程充分利用有标签数据和无标签数据之间的关联,提升舆情事件演化态势评估模型的泛化能力,进而提高关键舆情筛选的准确性以及管控资源的利用效能.

首先,本文将舆情演化态势评估问题转化为多指标排序问题,即根据一定的指标体系对网络舆情事件的重要程度进行排序.结合舆情事件特征和管控经验,从舆情演化过程中涉及的“人”“事”“势”等要素出发构建较为完善的网络舆情严重性评估指标,以全面反映舆情事件的演变规律.由于舆情事件的具体排序结果不但和评估指标具体数值有关,也和其所处的相对位置相关(例如和不同的网络舆情事件对比),为利用这种空间信息,本文将待排序事件的评估指标按照pairwise的形式进行组织,随后利用局部敏感Hash算法对数据集进行预处理,计算各数据点的邻域信息,构建数据的图结构;同时,利用二阶切比雪夫多项式作为卷积核,计算得到数据点及其一阶邻域的混合特征.更近一步,我们针对有标签数据和无标签数据设计了不同的损失函数,充分利用有标签数据和无标签数据之间的联系.针对有标签数据,其损失函数定义为数据对优先关系概率和其标签分布的交叉熵;针对无标签数据,其损失函数定义为数据对评分值的相似度和其特征空间相似度分布的交叉熵,通过超参构建模型的损失函数.最后,利用Adam算法和反向传播算法对模型进行迭代训练,建立排序模型.

为验证本文所提算法的性能,本文构建了2种类型的数据集.第1种为公开有标注的数据集,包括微软信息检索数据集MQ2007-semi和MQ2008-semi,利用这类数据验证本文算法的有效性;第2种为自主构建的舆情数据集,包括10个在2019-06-07—2019-06-14期间传播于新浪微博的典型舆情事件,利用这类数据验证本文所提算法在真实场景中的实用性和泛化性.2类数据集上的实验结果显示,本文所设计的算法具有良好的性能,能够在标签有限的情况下实现真实环境中舆情事件演化的重要性评估,为舆情事件的管控提供决策支持.

1 相关工作

舆情事件演化趋势重要性评估是指根据所设计的指标体系,量化舆情事件的影响范围或者危害程度.在过去几年,舆情事件管控逐渐引起了学术界的重视,和本文相关的主要工作简单总结如下:

在舆情演化态势评估指标构建方面,高承实等人[2]综合考查了社会类指标与技术类指标、舆情主体与舆情受众之间的关系构建了舆情监测指标体系;Jin 等人[3]设计了一种社交媒体中用户情感计算指标体系,并设计了相应的用户情感计算方法,以衡量社交媒体中用户情绪的影响;张一文等人[4]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.但是目前这些指标体系构建工作主要是以舆情事件中的特定方面要素为中心,导致态势评估的结果存在片面性;此外,所构建的指标体系大多同时包含可量化的数值型指标和不可量化的模糊性指标,这不利于舆情事件重要性的统一度量.

在评估指标的基础上,可以结合专家知识实现舆情演化态势重要性评估.郝楠等人[5]综合应用层次分析法和模糊理论构建基于模糊综合评价的网络舆情预警模型,并选取 3 个典型舆情事件进行了案例分析;但是这类网络舆情演化趋势重要性评估算法多依赖于专家知识,可扩展性和泛化性较差,对实施人员也有较高的专业性和知识性要求.

随着机器学习的发展,近几年也出现了一些将机器学习方法应用于舆情评估领域的研究,游丹丹等人[6]利用粒子群算法对建立在时间序列上的舆情演化趋势值进行预测;张和平等人[7]利用舆情事件的百度指数作为训练数据,建立了基于灰色 Markov 的舆情事件演化趋势预测模型.但是这类方法往往无法实现从评估指标到演化趋势的直接映射,实质上仍然利用了标注质量较高的数据进行训练和学习,所构建的模型面对真实环境下大规模、高并发的舆情演化趋势分析并没有良好的效果.

舆情事件态势评估可以转化为多指标排序任务,即根据指标体系,筛选出急需管控的舆情事件.虽然将排序学习算法应用于舆情研究领域的研究较少,但是有许多相关且可迁移的方法.Burges 等人[8]提出利用神经网络进行排序任务的 RankNet 算法并推导了对应的损失函数;之后 Burges 对 RankNet 进行了改进,使之可以优化 NDCG(normalized discounted cumulative gain)等非连续的信息检索指标;Pan 等人[9]提出了 Semi-RankSVM 算法,该算法是支持向量机排序学习的半监督拓展,主要创新是利用拉普拉斯正则化将数据结构信息的损失纳入学习目标;Amini 等人[10]提出了基于 RankBoost 的半监督排序算法,该算法首先依据特征向量空间距离较近的数据拥有相似标签的原则,为部分无标签数据赋予标签,然后利用真标记数据和伪标记数据训练模型.Xu 等人[11]提出了 AdaRank-NDCG 算法,它首先由训练集训练得到多个性能较弱的分类器,然后基于提升思想将其集成为更强的最终分类器,是效果较好的监督算法;Cao 等人[12]提出的 ListNet 算法是典型的列表数据形式监督算法,它将每个查询对应的整个数据列表当作一个训练数据,然后用模型预测的数据列表排序和真实列表排序之间的交叉熵作为损失函数.秦涛等人[13]利用排序算法对多指标舆情事件的严重程度进行排序,并利用主曲线模型构建了一种无监督排序模型.但这些工作都没解决如何利用少量有标签数据中专家知识的难题,以及如何利用有标签和无标签数据的关联特征训练建立具有泛化能力的舆情演化态势评估模型.

结合相关研究现状和舆情监控需求,本文在构建舆情事件演化趋势评估指标体系的基础上,设计了一种面向少量标注数据的演化趋势评估算法,利用标注数据中的专家知识以及标注和无标注数据之间的关联关系,提升态势排序模型的性能.

2 网络舆情演化趋势评价指标构建与计算

构建高质量的指标体系,可以将不同性质的舆情事件进行横向比较,有助于整体上掌握舆情的发展变化趋势,在此基础上制定引导和控制策略.

2.1 层次化指标体系构建

结合舆情管控的实际需求、前期研究基础及舆情管控经验,设计了涵盖舆情事件 3 个成因:“因人”“因事”“因势”的演化态势评估指标.“因人”是指和舆情事件发起者或者参与者相关的特征,例如舆情事件参与人的年龄,地域等特征,这部分特征主要由事件参与者的平台注册属性获取;“因事”是指舆情事件涉及的事件类型、已经存在的时长、话题主题等,这部分特征主要通过对博文的处理获取;“因势”

是指当前监控时刻舆情的具体演变态势,例如帖子数和参与人数呈现的增长态势等,这部分主要通过对所捕获帖子和参与人在时间维度的变化趋势获取.

据此,本文构建了包含 14 个指标的评估指标体系,如图 1 所示,所设计的指标综合考虑了舆情事件的传播特征和监管需求,涵盖了静态特征,例如参与人的粉丝数,也涵盖了事件演变的动态特征,例如事件传播的飙升度.此外,所构建的指标体系更加注重舆情事件传播的动态变化特点,更适合用于舆情事件演化趋势评估.

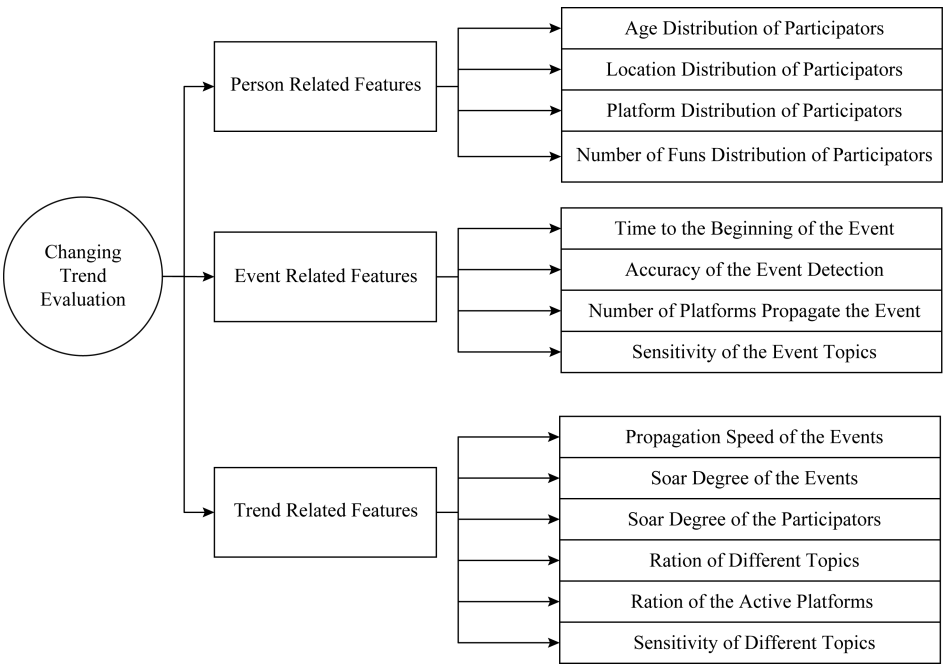


Fig. 1　The indices for public opinion changing trend evaluation

图 1　舆情演化趋势评价指标

2.2　评估指标体系量化

从有含义、易获取、易理解的角度,我们对不同指标的量化方法进行了不同的定义,其中,“因人”相关的特征量化方式为:

1) Feature a_1. 参与人群的年龄分布,其定义为所有该事件参与者年龄的标准差.

2) Feature a_2. 参与人群的地域分布,其定义为每个地域参与人数比率的标准差.

3) Feature a_3. 参与人在各平台的分布,其定义为由每个平台参与人数比率的标准差.

4) Feature a_4. 参与人的粉丝数量,其定义为所有参与人粉丝数量的平均值.

因舆情事件参与者众多,故本文采用特征得分来刻画和“人”相关的特征.在具体的计算过程中,利用标准差反映某一特征的分布,利用均值刻画粉丝的数量,在一定程度上降低了特征的计算复杂度.

“因事”相关的特征量化方式为:

1) Feature b_1. 事件发现时已存在的时间,其定义为从事件发生到态势评估时所经历的时间 t_n-t_0.

2) Feature b_2. 事件识别准确率,其定义为爬取的帖子中符合事件主题的帖子数量占总帖子数量的百分比.

3) Feature b_3. 事件发现时存在的平台数,其定义为爬取的数据来源站点的数量.

4) Feature b_4. 事件的严重或敏感程度,其定义为事件的严重或敏感程度评级,主要利用关键词的频繁度刻画.

在上述特征的计算过程中,特征 b_2 和 b_4 的计算需要用到专家知识,在一定程度上需要有标签数据,例如事件的敏感程度依赖于敏感的定义和敏感词语义的标注.

“因势”相关的特征量化方式为:

1) Feature c_1. 舆情事件的传播速度,其定义为符合事件主题的帖子在单位时间间隔内的新增量与时间间隔之比$(TP_n-TP_{n-1})/T$,即:

$$c_1=(TP_n-TP_{n-1})/T,$$

其中 TP_n 表示第 n 个时刻获得的符合目标主题的帖子数量.

2) Feature c_2. 舆情事件传播的飙升度,其定义为符合目标主题的帖子数量在第 n 个时间间隔内

的新增量和在前一个时间间隔内的新增量之差与时间间隔之比,即:

$$c_2=((TP_n-TP_{n-1})-(TP_{n-1}-TP_{n-2}))/T. \quad (1)$$

3) Feature c_3. 舆情事件参与人群飙升度,其定义为参与人数量在第 n 个时间间隔内的新增量和前一个时间间隔内的新增量之差与时间间隔之比,即:

$$c_3=((H_n-H_{n-1})-(H_{n-1}-H_{n-2}))/T, \quad (2)$$

其中,H_n 代表第 n 个时刻的参与人数量.

4) Feature c_4. 舆情事件话题倾向性比率,其定义为爬取的帖子中负向情感帖子数量占总帖子数量的比率,即:

$$c_4=TP_{ne}/TP.$$

5) Feature c_5. 舆情事件平台活跃度比率,其定义为爬取到符帖子数量超过设定值的平台的数量 NE_0 与总平台数量 NE 之比,即:

$$c_5=NE_0/NE.$$

6) Feature c_6. 舆情事件话题敏感度比率,其定义帖子中的敏感帖子的数量占总帖子数量之比 TP_{se}/TP,即:

$$c_6=TP_{se}/TP.$$

在具体的计算过程中,特征 c_4 和 c_6 的计算需要用到专家知识,在一定程度上需要有标签数据,所用到的标签数据和"因事"相关特征的标签数据相同.

根据上述特征定义,结合舆情监控的实际需求,以时间窗口 T 为时间单位将舆情数据集分段(在本文中 $T=1$ 天),并根据每个事件窗口内的数据量化指标.同时,为了克服指标量纲不同带来的影响,我们对抽取的特征进行了归一化处理:

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

其中,d_i 为某评估指标的值;d_i' 为归一化后的特征值;$d_i^{\min}$ 和 $d_i^{\max}$ 分别为该指标的最小值和最大值.

3 基于图卷积的排序神经网络模型

3.1 问题建模

多个舆情事件演化态势评估问题可以转化为排序问题,即根据指标体系量化舆情事件态势的严重性,并据此实现排序,筛选出急需管控的舆情事件.

首先,本文采用 pairwise 的形式重构指标数据集,即通过 2 个数据点之间的排序优先关系构建数据对.对于数据点 x_i 和 x_j 来说,其标签信息可有 3 种形式:$(\langle x_i,x_j\rangle,+1)$,$(\langle x_i,x_j\rangle,-1)$,$(\langle x_i,x_j\rangle,0)$,分别代表 x_i 在排序上优先于 x_j,x_i 在排序上落后于 x_j,以及 x_i 和 x_j 在排序优先度上无法区分.演化态势重要性排序模型的目标是通过训练建立一个评分函数 $f(\boldsymbol{x};\theta)$,评分函数将对待排序数据集中的每个数据点进行评分,评分值越高则代表该数据点在本次排序中拥有更高的重要度,亦即对应的舆情事件更加严重,最后根据评分值获取待排序数据集的排序.

3.2 基于排序学习的严重性评估模型整体框架

为利用标注数据和无标注数据之间的关联关系,在模型训练过程中充分利用有限标签数据中的专家知识,本文设计了如图 2 所示的舆情演化态势重要性评估模型,具体包括 4 个步骤:

Step1. 评估指标量化.结合第 2 节所设计的指标体系和量化方法,计算舆情事件的指标值,作为排序模型的输入.

Step2. 数据点邻域混合特征提取.利用局部敏感 Hash 算法构建数据点的邻域信息,以建立有标签

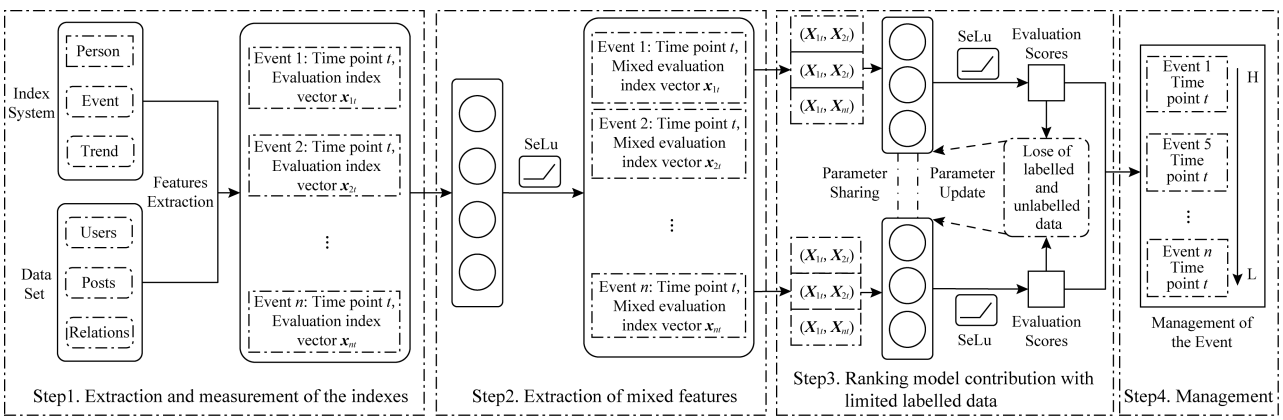


Fig. 2 Framework of the public opinion events changing trend evaluation

图 2 舆情演化态势严重性评估模型框架

数据点和无标签数据点之间的联系.

Step3. 排序神经网络模型构建.将指标数据向量重构为数据对形式,针对有标签数据和无标签数据设计不用的损失函数,提升模型效率.

Step4. 评估结果管控.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.

3.3　基于局部敏感 Hash 的数据邻域结构求取

在排序模型中,排序结果不但和具体的指标数据有关,也和排序点所处的相对位置有关,为利用这种知识,本文构建各数据点以欧氏距离度量邻域,从而挖掘数据分布的结构信息,并以此定义数据特征向量间的相似性.为降低算法时间复杂度,本文采用局部敏感 Hash 算法[14]来求取数据点的邻域.

在原始数据空间中距离较近的点会以一个高于指定阈值的概率被 Hash 至同一个值区间,距离较远的点会以一个低于指定阈值的概率被 Hash 至同一个值区间.我们利用 p 稳定分布[8]构建 Hash 函数族,对于一个数据向量 $\boldsymbol{v}=(v_1,v_2,\cdots,v_n)$,从 p 稳定分布中,随机选取与 $\boldsymbol{v}$ 的维度相同数量的随机变量 $(X_1,X_2,\cdots,X_n)$构成向量 $\boldsymbol{a}$,定义 $\boldsymbol{a}\cdot\boldsymbol{v}=\sum_{i=1}^{n}v_iX_i$,此时 $\boldsymbol{a}\cdot\boldsymbol{v}$ 与 $\|\boldsymbol{v}\|_pX$ 同分布.因此就可以通过生成一定数量的向量 $\boldsymbol{a}$,来计算一定数量的 $\boldsymbol{a}\cdot\boldsymbol{v}$,从而来估计 $\|\boldsymbol{v}\|_pX$ 的值.对于欧氏距离,即当范数 $p=2$ 时,标准正态分布就是一个 p 稳定分布.文献[15]中提出了一种欧氏距离下的 Hash 函数族:

$$h_{a,b}(\boldsymbol{v})=\left\lfloor\frac{\boldsymbol{a}\cdot\boldsymbol{v}+b}{r}\right\rfloor,\tag{4}$$

其中,$\boldsymbol{a}$ 为含义同上文所述;b 为一个属于$(0,r)$的随机数;r 为 Hash 系数.因为 $\boldsymbol{a}\cdot\boldsymbol{v}$ 可以估计 $\|\boldsymbol{v}\|_p$,那么 $\boldsymbol{a}\cdot v_1-\boldsymbol{a}\cdot v_2=\boldsymbol{a}\cdot(v_1-v_2)$可以估计 $\|v_1-v_2\|_p$.即当空间中 2 个数据点的距离 $c=\|v_1-v_2\|_p$ 小于一定值时,经过 Hash 函数 $h_{a,b}(\boldsymbol{v})$可以被以一定概率映射为同一值.Hash 表构建和数据邻域计算算法包括 5 个步骤:

Step1. 构建 L 组 Hash 函数族,每组由 k 个 Hash 函数组成.

Step2. 每个数据经过一个 Hash 函数族映射后,得到一个整型向量.

Step3. 整型向量经过一次散列后得到对应的 *key* 值,*key* 值经二次散列后得到其在 Hash 表中的索引,索引下的数据结构为字典,以存储同 *key* 值的不同数据.

Step4. 对于数据集中的每个数据,进行邻域计算时,依次经过 2 次散列,得到其在 Hash 表中的存储位置,将该位置中的所有数据取出.

Step5. 对取出的数据按照与查询数据的距离进行排序,取距离最小的 K 个作为该查询数据的邻域.

3.4　图卷积提取混合特征

所提取的邻域特征和原始的指标数值特征共同决定了排序结果,本文采用图卷积神经网络将空间特征和数值特征形成混合特征,以进行排序模型训练.

为了实现图上的卷积[16-17],首先要定义图的拉普拉斯矩阵 $\boldsymbol{L}$:

$$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A},$$

其中,$\boldsymbol{D}$ 为以图结构中各点的度作为对角线上值的对角矩阵;$\boldsymbol{A}$ 为图的邻接矩阵,表示不同数据点的连接关系,如果 2 个数据点均不在对方的邻域内,则邻接矩阵中对应元素为 0,否则为 1.对 $\boldsymbol{L}$ 进行谱分解可得:

$$\boldsymbol{L}=\boldsymbol{U}\begin{pmatrix}\lambda_1 & & \\ & \cdots & \\ & & \lambda_n\end{pmatrix}\boldsymbol{U}^{\mathrm{T}},\tag{5}$$

其中,$\boldsymbol{U}=(\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_n)$,$\boldsymbol{u}_n$ 为 $\boldsymbol{L}$ 的单位特征向量;λ_n 为 $\boldsymbol{L}$ 的特征值.

对于输出为 y,输入为 x,激活函数为 σ 的 GCN 网络层产生的变换为:

$$y=\sigma(\boldsymbol{U}g(\boldsymbol{\Lambda})\boldsymbol{U}^{\mathrm{T}}x),\tag{6}$$

式(6)为卷积核的一般形式,但是它有学习的参数多、需进行拉普拉斯矩阵分解等缺点.本文采用切比雪夫多项式展开近似卷积核,则 g 可近似为

$$g(\boldsymbol{\Lambda})=\sum_{m=0}^{k-1}\beta_mT_m(\boldsymbol{\Lambda}'),\tag{7}$$

其中,T_m 为 $m+1$ 阶的切比雪夫多项式;β_m 为对应的系数,同时也是需要学习的参数;$\boldsymbol{\Lambda}'$为经过数值变换至$[-1,1]$之间的特征值对角矩阵.在本文中仅考虑一阶邻域,则最终可得卷积层变换为

$$y=\sigma(\beta_0x-\beta_1\boldsymbol{L}'x).\tag{8}$$

经过卷积层后,可以提取出结合了邻域特征向量和评估指标特征向量的混合特征向量,混合特征可以帮助排序模型更好地进行学习.

4　排序模型损失函数设计

4.1　针对有标签数据构建损失函数

针对有标签数据,根据 2 个数据点的排序评分值和 2 个数据的实际优先关系设计相应的损失函数.

记数据点 x_i 经过排序模型输出的该数据点的评分值 s_i. $P(i\succ j)$ 为数据点 x_i 优先级高于 x_j 的概率,该概率依赖于两者的排序评分:

$$P(i\succ j)=\frac{1}{1+\mathrm{e}^{-(s_i-s_j)}}. \tag{9}$$

此模型满足序的传递性,即若 $P(i\succ j)>0.5$, $P(j\succ k)>0.5$,可以推出 $P(i\succ j)>0.5$.据此,有标签数据的损失函数可由数据对的优先关系标签和其概率预测分布的交叉熵度量,即:

$$C_{\mathrm{L}}=-\sum_{(i,j)\in \boldsymbol{L}} l_{i\succ j}\log P(i\succ j), \tag{10}$$

其中,$l_{i\succ j}$ 表示数据对的标签.当 $i=j$ 时,$l_{i\succ j}=0$;当 $i\succ j$ 时,$l_{i\succ j}=1$;当 $i\prec j$ 时,$l_{i\succ j}=2$.

4.2 针对无标签数据构建损失函数

针对无标签数据,如何将有标签数据和无标签数据联系起来是在标签有限的情况下实现训练效果提升的可靠方法[18].根据流形假设[19],相似的数据点应具有相似的评分和排序优先级,据此在设计无标签数据的损失函数时,通过添加流形正则化项,使模型在根据有标签数据进行训练时可以利用无标签数据的结构信息,以此提高模型效果.

针对每个数据对,都有以概率形式的输入特征相似度和概率形式的排序优先级相似度.因特征空间的相似度与排序优先级的相似度应趋于一致,故要根据输入数据的相似度对损失函数的输出进行惩罚.为此,首先定义数据点 x_i 和 x_j 之间的距离为

$$d_{ij}^2=\|x_i-x_j\|. \tag{11}$$

同时,流形正则项只在数据点的最近的 K 个近邻生效[20],根据流形假设,只有局部邻域(利用 4.1 节所述局部敏感 Hash 算法计算得出)内的样本数据拥有相似性特征.因此,数据点 x_i 和 x_j 间的转移概率为

$$q_{j|i}=\begin{cases}\dfrac{\mathrm{e}^{-\frac{d_{ij}^2}{\sigma_i^2}}}{\sum\limits_{k\in N_k(i)}\mathrm{e}^{-\frac{d_{ij}^2}{\sigma_i^2}}}, & j\in N_k(i),\\ 0, & j\notin N_k(i),\end{cases} \tag{12}$$

其中,$N_k(i)$ 为 x_i 的邻域;σ_i^2 为尺度放缩系数.则数据点 x_i 和 x_j 间的相似性可以用 $q_{ij}=\dfrac{\max(q_{i|j},q_{j|i})}{Z}$ 衡量,其中 Z 是归一化系数.

然后,以概率形式定义数据对在排序优先级上的相似度:

$$r_{ij}=P(i\not\succ j)\cdot P(j\not\succ i) \tag{13}$$

其中,$P(i\not\succ j)=1-P(i\succ j)=P(j\succ i)$,代表了 x_i 不优先于 x_j 的概率,$P(j\not\succ i)$ 同理.则 $P(i\not\succ j)+P(j\not\succ i)=1$. $P(i\not\succ j)\cdot P(j\not\succ i)$ 越大,即两者和一定情况下,乘积越大,则 $P(i\not\succ j)$ 和 $P(j\not\succ i)$ 越接近 0.5,也即 x_i 和 x_j 的排序优先级相同的概率越大.

利用数据点交叉熵衡量特征空间的相似度与排序优先级的相似度这两者分布的差异,并作为惩罚的依据,由此可得无标签数据的损失函数为

$$C_{\mathrm{UL}}=-\sum_{(i,j)\in \boldsymbol{U}} q_{ij}\log P(i\not\succ j)P(j\not\succ i). \tag{14}$$

通过超参将有标签数据的损失函数和无标签数据的损失函数结合起来,即可得到整个排序模型的损失函数.

5 实验与分析

5.1 数据集

因为真实的舆情数据集缺乏权威标注,无法衡量本文所提方法的性能,本文首先利用公开数据集验证本文模型性能,之后在真实的舆情数据集上验证方法的实用性,综合两者来评判本文所提算法在舆情演化趋势评估中的可用性.

公开数据集采用微软 MQ2007-semi[21] 和 MQ2008-semi[21],是文档信息检索领域的半监督数据集,其中有标签数据由查询 ID、数据相关性标注和数据特征向量组成,无标签数据由查询 ID 和数据特征向量组成.MQ2007-semi 数据集包含 1 693 个查询 ID,MQ2008-semi 数据集包含 785 个查询 ID,对于一个查询 ID 来说,对应有标签数据数量约为 40 个,对应无标签数据数量约为 1 000 个;数据相关性标注分为{0,1,2},其中 0 代表该数据与查询完全无关,1 代表两者间有一定相关性,2 代表两者完全相关;数据特征向量共有 46 维,对应 46 个数据评价指标,主要包括:词频(term frequency, TF)、逆向文件频率(inverse document frequency, IDF)、二元独立模型(binary independence model, BIM)、信息检索语言模型(language model for information retrieval, IMIR)等.2 个数据集都被等量地划分成 5 个子集,选取其中 3 个子集作为训练集,在训练集上对模型进行训练;选取其余 2 个子集其中的 1 个作为验证集,在验证集上进行参数选取,选用评价指标最高的模型的参数作为最终参数;最后一个子集作为测试集,在测试集上应用模型来评估模型的泛化性能.

真实的舆情数据集由新浪微博中相关典型舆情事件的帖子构成,数据集详细信息如表 1 所示.在 2019-06-07—2019-06-14 间,共采集了 10 个典型舆情事件的帖子 43 042 条,事件主题和事件描述如表 1 所示.所选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、娱乐等多种舆情事件,具有一定的代表性.由表 1 我们可以看出,关于中美贸易战舆情事件的帖子数最多,关于俄部署导弹的帖子数最少.如按照帖子数量大小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.

Table 1　Data Set of Public Opinion Events

表 1　舆情事件数据集

Event	主题	事件概述	帖子数量
Event1	华为	华为公司在贸易战中被打压	10 758
Event2	赵志勇	强迫卖淫、强奸犯被处死刑	3 857
Event 3	高考	高考泄题、漏题等	4 428
Event 4	林志玲	娱乐明星林志玲结婚	4 038
Event 5	王源	明星王源学会盛饭而上热搜	4 611
Event 6	孙宇晨	孙宇晨炒作比特币	2 510
Event 7	NBA	NBA 总决赛信息	4 084
Event 8	普京	俄疑部署针对我国导弹	1 734
Event 9	扫黑	打击黑恶势力专项行动	3 487
Event 10	百度	百度人事变动股价大跌	3 535

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,利用本文所提算法对 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.

5.2　评估指标

1) NDCG:归一化折损累计增益,考虑排序结果的相关度和位置计算增益,并进行归一化计算[22]:

$$NDCG_k=\frac{DCG_k}{IDCG_k}=\frac{\sum_{i=1}^{k}\frac{2^{rel_{i-1}}}{\mathrm{lb}(i+1)}}{\sum_{i=1}^{|REL|}\frac{2^{rel_{i-1}}}{\mathrm{lb}(i+1)}},\quad(15)$$

其中,i 表示数据在排序结果中的位置,rel_i 表示第 i 个位置上的相关度,$|REL|$ 表示最佳排序结果.

2) $P@n$(precision at position n):它是排序列表的前 n 个数据中与查询相关的数据数量与 n 的比值[23].即:

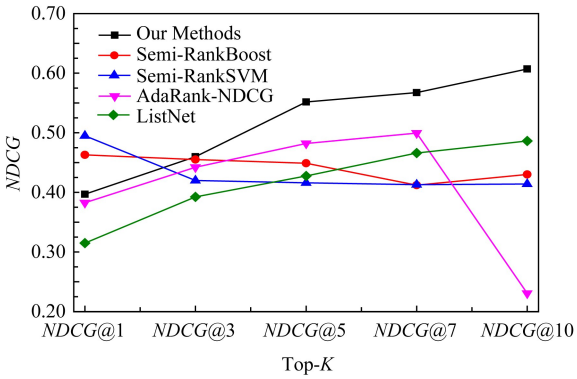
$$P@n=\frac{1}{n}\sum_{i=1}^{n}rel(i),\quad(16)$$

其中,$rel(i)$ 表明排序结果中第 i 个位置的数据是否与查询相关,有关时值为 1,无关时值为 0.

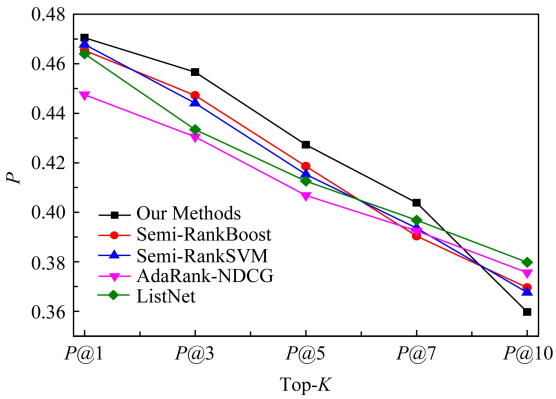
5.3　实验结果分析

5.3.1　公开数据集实验结果分析

利用公开有标注的数据集验证本文所提算法的性能,同时选择 Semi-RankSVM[9],Semi-RankBoost[10],AdaRank-NDCG[11],ListNet[12] 等 4 种方法进行了对比分析.在 MQ2007-semi 和 MQ2008-semi 上的实验结果如图 3、图 4 所示.由图 3 所示结果表明,本文所提出的算法在 $NDCG@3,5,7,10$ 四个指标上都表现出了良好的性能;与此同时在 $P@1,3,5,7$ 四个指标上也表现出了良好的性能.通过对图 4 所示的结果分析,可以得到相似的结论.更近一步求取 NDCG 和 P 的平均值,用来度量所提模型的性能,结果如表 2 所示.和传统的若监督或无监督算法相比,本文所提算法在 Mean $NDCG$ 和 Mean P 指标上均有不同程度的提升,验证了本文所提算法在少标签数据的排序任务有更好的性能.



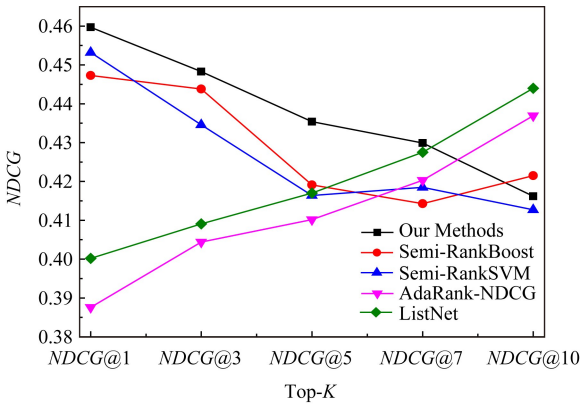
(a) Experimental results on *NDCG*



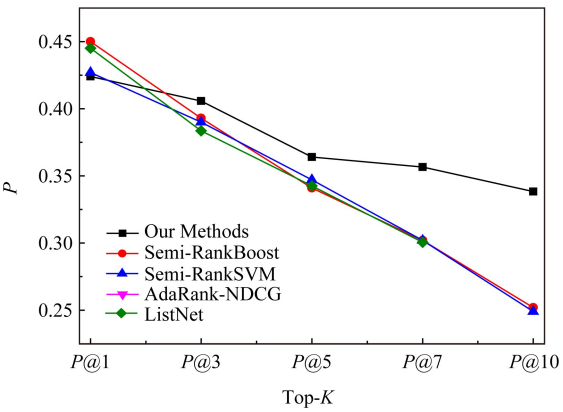
(b) Experimental results on *P*

Fig. 3　Experimental results using MQ2007-semi data set

图 3　MQ2007-semi 数据集分析实验结果



(a) Experimental results on *NDCG*



(b) Experimental results on *P*

Fig. 4 Experimental results using MQ2008-semi data set

图 4 MQ2008-semi 数据集分析实验结果

Table 2 Comparison Experiment Results

表 2 对比实验结果统计表

Methods	MQ2007-semi		MQ2008-semi	
	Mean *NDCG*	Mean *P*	Mean *NDCG*	Mean *P*
Our-Methods	**0.4379**	**0.4236**	**0.5166**	**0.3777**
Semi-RankBoost	0.4292	0.4182	0.4418	0.3476
Semi-RankSVM	0.4271	0.4177	0.4316	0.3430
AdaRank-NDCG	0.4119	0.4106	0.4073	0.3444
ListNet	0.4196	0.4173	0.4174	0.3439

Notes: The best results are in bold.

5.3.2 消融实验分析

为了验证数据邻域和不同损失函数设计的作用,将两者分别消去进行同参数条件下的实验,实验结果如表 3 所示,其中 Semi-RankNet 相对本文方法消去了数据邻域特征,GCN-RankNet 相对本文方法消去了有差别的损失函数.表 3 表明消去邻域特征和采用无差别损失函数后效果均有不同程度的下降.因此引入混合特征和差别化损失函数后,捕捉到了数据的结构信息,建立了有标签和无标签数据之间的联系,利用数据点的绝对数值和其在特征空间的相对位置,增强了有标签数据较少情况下模型的学习能力.

Table 3 Experiment Results on Ablation Analysis (Mean *NDCG*)

表 3 消融实验分析结果(Mean *NDCG*)

Methods	MQ2007-semi	MQ2008-semi
Our Method	**0.4379**	**0.5166**
Semi-RankNet	0.3931	0.3943
GCN-RankNet	0.4119	0.4460

Notes: The best results are in bold.

5.3.3 舆情数据集实验结果分析

在舆情数据集上进行实验,根据实际管控需要将舆情事件管控等级分为 0,1,2 三级,分别代表无需调控,需要关注和亟需调控,管控优先级的设置和每个优先级中的舆情事件个数取决于管控资源.其中管控优先级 2 为最高,表示管理部门需要对相应的舆情事件进行管控,结合管控经验,随机选取少量数据点(10 个)进行标注.以归一化后的模型输出评分作为该舆情事件在该时间段的舆情事件演化趋势值.2019-06-07—2019-06-13,舆情事件演化趋势变化如图 5 所示:

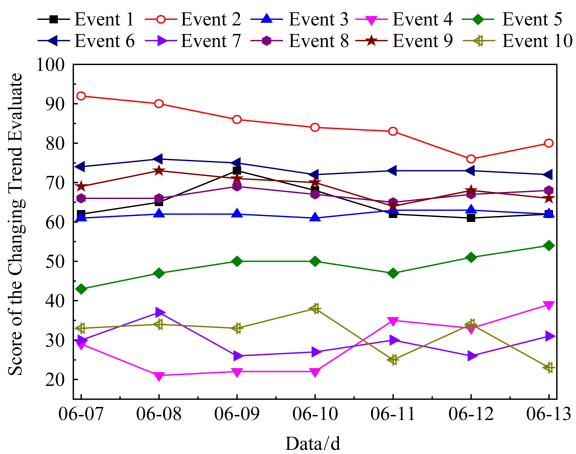


Fig. 5 Dynamic changing trends of the public opinion events

图 5 舆情事件演化趋势变化图

由图 5 中可以看出,赵志勇和孙宇晨事件演化趋势较为敏感,全时段监控需求均为 2 级,从事件本身来看,赵志勇因为其恶劣的行径被判处死刑,孙宇晨因利用比特币进行欺骗和炒卖使多个家庭破产分裂.这 2 个事件的发展速度较快,且热度一直维持在

较高水平,网民言论中存在较多偏激的观点,情绪倾向更为敏感,所以其演化趋势评估分数较高也符合常理认知.华为和扫黑除恶事件在部分日期监控需求为2级,其余时段为1级;高考和普京事件全时段监控需求为1级.这些事件的热度和情感倾向没有过分突出,因而监控等级也更弱.林志玲、王源、NBA、百度等舆情事件,因为其事件性质并不敏感,网民的态度较为中立,情感倾向也趋于正面,整体热度不高,所以模型对其演化趋势评估较低同样也符合常理认识.

6　总结与展望

本文针对舆情事件演化趋势评估任务中可供学习的有标签数据有限所造成的困难,利用排序学习模型解决并发舆情事件演化趋势严重性或管控优先级的筛选问题.本文首先设计了舆情演化趋势评估指标体系,然后提出了一种基于神经网络的半监督舆情事件演化趋势重要性评估算法,实现演化趋势及危害严重舆情事件的筛选.在公开数据集和真实舆情数据集上进行了实验分析,实验结果表明,本文方法具有良好的性能,并且对评估网络舆情事件演化趋势具有一定的有效性.在下一步的工作中,我们将针对降低模型时间复杂度,扩大图卷积邻域范围以增强模型性能进行进一步的研究.

参　考　文　献

[1] China Internet Network Information Center. The 45th statistical report on Chinese Internet Development [EB/OL]. [2020-04-28]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/202004/P020200428596599037028.pdf (in Chinese)
(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[EB/OL]. [2020-04-28]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg /202004/P020200428596599037028.pdf)

[2] 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 (in Chinese)
(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)

[3] Jin Rui, Zhang Hongli, Zhang Yu. The social negative mood index for social networks [C] //Proc of the 3rd 2018 IEEE Int Conf on Data Science in Cyberspace. Piscataway, NJ: IEEE, 2018: 1-5. doi: 10.1109/DSC.2018.8570298

[4] 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 (in Chinese)
(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)

[5] Hao Nan, Feng Jing, Gao Yuan. Study on method of network public opinion early warning based on fuzzy comprehensive evaluation [J]. Journal of Chongqing University of Technology: Natural Science, 2019, 33(8): 227-231 (in Chinese)
(郝楠, 冯晶, 高媛. 基于模糊综合评价的网络舆情预警方法研究[J]. 重庆理工大学学报: 自然科学, 2019, 33(8): 227-231)

[6] You Dandan, Chen Fuji. Research on the prediction of network public opinion based on improved PSO and BP neural network [J]. Journal of Intelligence, 2016, 35(8): 156-161 (in Chinese)
(游丹丹, 陈福集. 基于改进粒子群和BP神经网络的网络舆情预测研究[J]. 情报杂志, 2016, 35(8): 156-161)

[7] Zhang Heping, Chen Qihai. Research on the prediction of network public opinion based on grey Markov model [J]. Information Science, 2018, 36(1): 75-79 (in Chinese)
(张和平, 陈齐海. 基于灰色马尔可夫模型的网络舆情预测研究[J]. 情报科学, 2018, 36(1): 75-79)

[8] Burges C, Shaked T, Renshaw E, et al. Learning to rank using gradient descent [C]//Proc of the 22nd Int Conf on Machine Learning. New York: ACM, 2005: 89-96

[9] Pan Zhibin, You Xing, Chen Hong, et al. Generalization performance of magnitude-preserving semi-supervised ranking with graph-based regularization [J]. Information Sciences, 2013, 221(2): 284-296

[10] Amini M R, Truong T V, Goutte C. A boosting algorithm for learning bipartite ranking functions with partially labelled data [C] //Proc of the 31st Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2008: 99-106

[11] Xu J, Li H. AdaRank: A boosting algorithm for information retrieval [C] //Proc of the 30th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2007: 391-398

[12] Cao Z, Qin T, Liu T Y, et al. Learning to rank: From pairwise approach to listwise approach [C] //Proc of the 24th Int Conf on Machine Learning. New York: ACM, 2007: 129-136

[13] Qin Tao, Wang Xifeng, Shen Zhuang, et.al. Research on unsupervised method for the importance of changing trend evaluation of Internet public opinion events [J]. Journal of Xi'an Jiaotong University, 2020, (11): 113-120 (in Chinese)
(秦涛, 王熙凤, 沈壮, 等. 面向无监督的网络舆情事件演化趋势重要性评估方法研究[J]. 西安交通大学学报, 2020, (11): 113-120)

[14] Datar M, Immorlica N, Indyk P, et al. Locality-sensitive hashing scheme based on p-stable distributions [C] //Proc of the 20th Annual Symp on Computational Geometry. New York: ACM, 2004: 253-262

[15] Hu B G, Mann G K I, Gosine R G. Control curve design for nonlinear(or fuzzy) proportional actions using spline-based functions [J].Automatica, 1998, 34(9): 1125-1133

[16] Bruna J, Zaremba W, Szlam A, et al. Spectral Networks and Locally Connected Networks on Graphs [EB/OL]. [2009-01-01]. https://arxiv.org/abs/1312.6203

[17] Zeng Yifu, Mu Qilin, Zhou Le, et al. Graph embedding based session perception model for next-click recommendation [J]. Journal of Computer Research and Development, 2020, 57(3): 590-603 (in Chinese)
(曾义夫, 牟其林, 周乐, 等. 基于图表示学习的会话感知推荐模型[J]. 计算机研究与发展, 2020, 57(3): 590-603)

[18] Liu Jianwei, Liu Yuan, Luo Xionglin. Semi-supervised learning methods [J] Chinese Journal of Computers, 2015, 38(8): 1592-1617 (in Chinese)
(刘建伟, 刘媛, 罗雄麟. 半监督学习方法[J]. 计算机学报, 2015, 38(8): 1592-1617)

[19] Liu Yufeng, Li Renfa. Graph regularized semi-supervised learning on heterogeneous information networks [J]. Journal of Computer Research and Development, 2015, 52(3): 606-613 (in Chinese)
(刘钰峰, 李仁发. 异构信息网络上基于图正则化的半监督学习[J]. 计算机研究与发展, 2015, 52(3): 606-613)

[20] Szummer M, Yilmaz E. Semi-supervised learning to rank with preference regularization [C] //Proc of the 20th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2011: 269-278

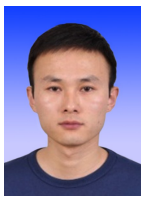
[21] Tao Q, Tie-Yan L. LETOR: Learning to rank for information retrieval [EB/OL]. [2009-01-01]. http://www.microsoft.com/en-us/research/project/letor-learning-rank-information-retrieval

[22] Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques [J]. ACM Transactions on Information Systems, 2002, 20(4): 422-446

[23] Manning D, Raghavan P, Schütze H. Introduction to Information Retrieval [M]. Beijing: The People's Posts and Telecommunications Press, 2010

Qin Tao, born in 1982. PhD, associate professor. Member of CCF. His main research interests include network measurement, online behavior monitoring and management.

Shen Zhuang, born in 1995. Master. His main research interests include online behavior monitoring and management. (s62951413@163.com)



Liu Huan, born in 1990. PhD, assistant professor. His main research interests include machine learning, computer vision and public opinion analysis. (liulaha@qq.com)

Chen Zhouguo, born in 1980. Master, senior engineer. His main research interests include social network analysis, big data and network forensics. (czgexcel@163.com)