计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(12):235−241 [doi: 10.15888/j.cnki.csa.008816]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

面向特定观点的网络舆情信息有用性排序①

齐宝森, 杜义华

(中国科学院 计算机网络信息中心, 北京 100190)
通信作者: 杜义华, E-mail: yhdu@cashq.ac.cn

摘　要: 网络舆情信息挖掘是舆情研究的重要课题. 在大量的信息面前, 为了快速发掘有用性高的舆情信息为舆情的分析、决策提供助力, 提出一种面向特定观点的舆情信息有用性排序方法, 实现快速发掘特定观点下有用舆情信息的目的. 该方法针对舆情信息的具体观点进行分析计算, 同时根据舆情信息可信度和关注度、传播者的影响力, 并且结合信息时效性等因素, 利用排序方法进行打分, 根据舆情信息的得分进行有用性排序. 实验结果表明, 该方法能很好的完成对舆情信息的推荐排序. 本研究理论上对舆情信息挖掘的研究理论进行补充, 现实意义对舆情管理者有很好的辅助作用, 能够为网络舆情引导工作提供助力.

关键词: 网络舆情; 有用性; 特定观点; 排序方法; 数据挖掘

引用格式: 齐宝森,杜义华.面向特定观点的网络舆情信息有用性排序.计算机系统应用,2022,31(12):235–241. http://www.c-s-a.org.cn/1003-3254/8816.html

Helpfulness Ranking of Network Public Opinion Information for Specific Viewpoints

QI Bao-Sen, DU Yi-Hua

(Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: The mining of network public opinion information is an important subject of public opinion research. To quickly discover useful public opinion information among a large amount of information and support public opinion analysis and corresponding decision-making, this study proposes a method of ranking the usefulness of public opinion information for specific viewpoints to achieve the purpose of quickly discovering useful public opinion information under specific viewpoints. This method analyzes and calculates the specific viewpoints of public opinion information and assigns scores to the information by a ranking method according to its credibility and attention and the influence of the disseminator with due consideration of the timeliness of the information and other factors. Then, the public information is ranked in terms of usefulness according to its score. The experimental results show that the proposed method performs well in the recommendation ranking of public opinion information. This research theoretically supplements the research theory of public opinion information mining. Concerning practical significance, it can well assist public opinion managers as it provides a boost to the guidance of network public opinions.

Key words: network public opinion; helpfulness; specific viewpoint; ranking method; data mining

随着互联网技术的快速发展, 网络成为舆论传播的主阵地与主战场, 网民在各种媒体平台自由的发表言论、传播自己的观点, 网络舆情信息的数量呈指数级的增长. 面对各种观点的海量网络舆情信息, 舆情决策者很难及时获取相关有效的信息, 容易出现信息不对称的情况, 不利于舆情决策者掌握舆论导向来进行引导工作. 如何有选择性的发掘可能对舆情工作者有用性高的网络舆情信息, 筛选掉有用性低的数据内容, 是网络舆情信息挖掘阶段的一个难题. 发掘有用性高的网络舆情信息内容可以辅助舆论引导工作者进行更好的决策, 实现

① 基金项目: 中国科学院战略性先导科技专项 (C 类)(XDC02060100)
收稿时间: 2022-03-14; 修改时间: 2022-04-12; 采用时间: 2022-04-22; csa 在线出版时间: 2022-07-14

网络舆论的管控与引导, 具有重要的现实意义.

目前针对发掘有用的网络舆情信息的问题, 众多学者进行了相关研究. 例如米昂[1] 利用舆情信息内容相似度和用户特征, 在海量的舆情信息中快速找到舆论传播的源头. 刘一飞[2] 通过文本摘要技术和文本分类技术分析网络数据中对相关部门有价值的舆情信息. 吴春琼等[3] 则提出要运用大数据的技术, 来帮助在海量网络舆情信息中快速找到大量隐含的、具有潜在价值的数据信息. 黄微等[4] 利用推文内容和定量数据构建网络舆情推文热度测度模型, 快速筛选出为网络舆情研究所用的数据. 同时众多学者也对信息有用性做了很多研究. Luo 等[5] 探究了信息对第三方论坛中信息阅读者有用性感知的影响. Malik[6] 针对评论数量增加造成的信息过载问题, 使用多元自适应回归、分类和神经网络等方法对评论有用性进行预测. 郭顺利等[7] 为了解决信息多样化和答案冗余的问题, 提出融合 Word2Vec 和 WGRA 的方法对社区的答案进行有用性排序; 部分学者[8,9] 也是为了能够准确发现满足用户信息需求的有用评论, 采取不同的方法对评论进行有用性排序, 为用户提供更大的参考价值.

总体来说, 对于发掘有用的网络舆情信息的研究主要在舆情信息溯源与内容分析等方面. 对高效的发掘有用的网络舆情信息的相关研究较少, 同时针对舆情信息的分析主要基于内容特征和传播的拓扑结构关系, 没有具体到某个观点的舆情信息的研究, 没有考虑到舆情信息可信度和传播者影响力等其他影响因素. 在对信息有用性的研究中, 学者大部分研究对象是在用户评论和社区问答的领域, 主要针对有用性排序方法模型的研究, 目前有用性排序的方法并不完全适用于舆情信息领域, 并不具备较好的通用性.

基于以上的研究背景, 提出一种面向特定观点的舆情信息有用性排序方法. 首先从舆情信息的特定观点研究出发, 分析当前舆情信息在特定观点下的态度, 然后结合舆情信息可信度、舆情信息关注度、传播者影响力的计算, 综合考虑时效性对舆情信息的影响, 最后利用客观赋权的熵值法和改进的排序算法对舆情信息进行有用性打分, 根据有用性排序分值的高低, 发掘对舆论引导工作者有用性高的舆情信息.

1 面向特定观点的舆情信息有用性排序方法

舆情信息有用性排序是指在给定某个舆情事件的舆情信息数据集的情况下, 根据一定的指标和方法, 进行有用性计算, 快速发掘特定观点下有用性高的舆情信息, 达到排序推荐的效果. 本文舆情信息有用性是指在媒体平台中, 舆情信息在传播过程中所体现的影响力. 在传播过程中有影响力的文章、图片、视频等受到的关注越高、可信度越高、传播的范围也更广泛. 这些在舆情传播过程中发挥影响力的信息, 对于舆情决策者分析网络舆情动态、舆情走向、舆情决策的制定等方面更具有价值, 对应舆情信息的有用性越高. 同时对于舆情信息有用性分析, 可以过滤掉受关注度低、可信度低、传播范围窄等方面的舆情信息, 防止舆情管理者面对海量舆情信息, 很难及时获取相关有效的信息来进行分析的情况, 避免数据灾难.

基于对舆情信息传播过程中影响力的研究, 发现舆情信息本身的可信度大、受众对舆情信息的关注度高、传播者本身的影响力大和发布时间较晚的舆情信息有用性就越高. 本文设定舆情信息可信度、舆情信息关注度、传播者影响力及舆情信息时效性 4 个方面作为舆情信息有用性的评价指标, 为后面的研究奠定基础. 在某一舆情事件中, 受众面对海量的舆情信息会充分考虑本身可信度的问题, 国内外的学者认为信息来源的可信度对信息扩散产生的影响是十分巨大的[10,11], 所以可信度越高的舆情信息, 越有利于信息的扩散传播, 在传播过程中所体现的影响力越大. 同样舆情信息关注度越高, 其更多的用户会浏览相关舆情信息, 信息更易于被转发、被评论, 针对这些高阅读量、高转发和高评论的信息, 舆情信息的有用性越高. 部分学者研究发现, 对于“精英用户”或者“意见领袖”, 更利于引导信息的流向, 所发布的信息更易被转发和讨论[12,13]. 而对于传播者本身的影响力, 受众更易于关注影响力大的传播者, 在舆情事件发展中扮演举足轻重的作用, 发布的舆情信息的有用性也高于其他普通用户. 舆情信息发布的时间对舆情发展的动态和走向产生不同程度的影响, 发掘有用的舆情信息需要做到时效性. 所以舆情信息的发布时间是舆情信息有用性评价的重要指标.

本文提出的舆情信息有用性排序方法 (SPHR) 首先对舆情信息的特定观点进行分析, 然后综合考虑了舆情信息的可信度与关注度、传播者的影响力和舆情信息时效性多个因素. 通过排序方法计算有用性得分, 通过得分对舆情信息进行有用性排序, SPHR 方法流程如图 1 所示.

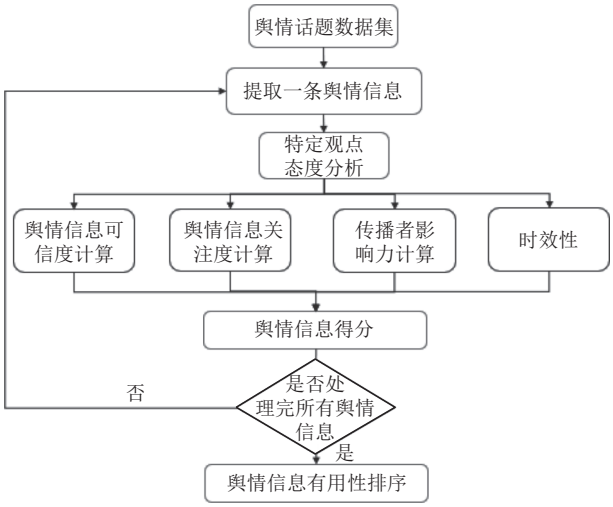

图 1 SPHR 方法流程图

1.1 网络舆情信息特定观点态度分析

针对某一个话题的舆情事件所产生的舆情信息是数以万计的, 每一条舆情信息本身所代表的观点态度是不同的. 当前信息的传播者对待事件的态度通过舆情信息的观点体现. 舆情决策者根据设置议题和引导口径的差别, 反映了决策者侧重于不同的观点, 发掘特定观点下有用的舆情信息, 可以深入的了解某个观点对于舆情事件的影响程度, 便于把握舆情的走向和动态变化, 有利于进行舆论决策的制定. 所以针对性的初步筛选出特定观点下的舆情信息是必要的.

通过对大量的舆情话题信息数据分析, 特定观点下的网络舆情信息态度主要包括支持和反对. 面对某一话题下海量的网络舆情信息, 发掘特定观点下有用的舆情信息, 需要根据舆情决策者设置议题或引导口径表明的观点, 进行网络舆情信息分析, 初步筛选出特定观点下的舆情信息. 针对舆情信息下特定观点的态度分析, 首先对文本信息进行词频特征提取, 然后利用 Word2Vec 训练词向量获取句子的深层语义信息, 最后利用支持向量机模型对其训练和预测完成最终的舆情信息特定观点态度分析任务.

首先对于舆情信息的文本数据, 使用 TextRank 的算法对所有文本特征进行文本关键词提取. TextRank 算法将文档看作词的网络, 利用词语之间的相邻关系来构建图网络, 作为表示词与词之间的语义关联[14]; 相邻之间的节点连接赋予权重, 即代表语义关联程度的高低; 然后采用算法迭代计算得到图中的每一个节点的 rank 值, 直至收敛, 最后通过 rank 值倒序排序, 得到文章的关键词. TextRank 迭代计算公式如式 (1):

$$WS(V_i)=(1-d)+d\times\sum_{v_j\in In(V_i)}\frac{W_{ji}}{\sum_{V_k\in Out(V_j)}W_{jk}}WS(V_j) \tag{1}$$

本文使用 TextRank 方法对舆情信息的文本进行特征提取, 通过该方法提取后, 文本信息具备更强的表示性, 因此可以用于后续的模型输入. 首先基于所有文本数据生成文本关键词, 然后对于每一组文本数据, 计算该文本中的基于文本关键词的权重矩阵, 因而将纯文本数据转换为矩阵形式. 为了限制特征规模, 对于文本特征获取最高为 100 个关键词, 因此后续形成的文本特征规模为 100 维.

上述方法没有获取文本深度语义信息, 因此采用 Word2Vec 模型来获取深层语义信息[15]. Word2Vec 是轻量级的神经网络, 模型共有 3 层: 输入层、隐藏层和输出层, 根据输入输出的不同, 包括 CBOW 模型和 Skip-gram 模型. 该模型能有效的提取文本内容的深层语义含义, 近年来已经有很多学者使用该技术用于自然语言处理中, 并取得了非常出色的实验效果[15].

Skip-gram 模型是根据已知词 w_t 的前提下, 对 w_t 的上下文 $w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$ 进行预测, 而 CBOW 模型正好相反[11]. 本文采用的是 Skip-gram 模型, 它的目标函数取对数似然函数为: $L=\sum \log p(Context(w_c)|w_t)$. 本文利用 Gensim 工具对舆情信息语料用 Word2Vec 训练词向量, 词向量的维数设置为 200 维. 训练完成得到最终的词向量 Skip-gram 模型.

最后利用支持向量机进行特定观点态度的分类任务. 对于文本分类问题, 支持向量机与其他分类算法相比在处理非线性及高维分类中有较好的分类效果[16]. 首先对训练文本进行关键词特征提取, 构建 100 维的浅层文本特征向量, 然后利用 Word2Vec 模型提取文本深层特征, 构建 200 维的特征向量, 综合考虑文本浅层特征和深层语义特征得到最终的特征向量. 支持向量机模型根据特征向量和数据标注训练出对应模型, 并利用训练好的模型对数据进行特定观点态度的分析, 得到舆情话题下特定观点的具体态度.

1.2 网络舆情信息有用性计算

1.2.1 舆情信息可信度计算

在特定观点下的舆情信息, 可信度越高, 用户越容易予以信任, 相对其有用性就越高. 舆情信息可信度通过信息源的权威性和信息的独立性指标来计算. 信息源的权威性代表信息的来源是否来自于官方平台, 是

否有平台认证, 来自于官方平台或者有平台认证的账号发布的信息可信度会越高. 划分信息源的权威性为 3 类: 官方主流信息源 (比如人民日报、人民网、新华网等) S1, 非官方主流认证信息源 S2, 未认证的普通信息源 S3. 信息的独立性表示信息是否为独立发表, 是否为转载内容. 相似度高的两个信息, 独立发表的比信息转载的可信度高. 划分信息独立性类别为 4 类: 原创 D1, 没有标注来源的转载 D2, 有来源信息的转载 D3, 未表明原创和转载的信息 D4. 对于以上的信息可信度通过启发式规则衡量, 具体如表 1.

表 1 舆情信息可信度分类

特点	第1类	第2类	第3类
权威性	S1	S2	S3
独立性	D1	D2、D3	D4
可信度	高	中	低

基于以上的分析, 针对舆情信息可信度计算如式 (2) 所示:

$$Cre\text{-}Info = u_1 \times Info\text{-}So_i + u_2 \times Info\text{-}In_i \tag{2}$$

其中, 可信度从低到高的可信度值设置为 0–2. $Info\text{-}So_i$ 为第 i 条信息源权威性的可信度, $Info\text{-}In_i$ 为第 i 条信息独立性的可信度. 另外 u_1 和 u_2 为其权重值. u_1 和 u_2 的值利用 SPSS 工具, 通过主成分分析方法对其进行量化.

1.2.2 舆情信息关注度计算

舆情信息的关注度利用信息的定量数据来计算, 通过信息的阅读数、点赞数和评论数来反映关注度的大小. 用户浏览相关的舆情信息, 会产生阅读数量的增加, 阅读量越大, 代表着更多的用户关注本条舆情信息, 对应的关注度越高. 同理用户浏览不仅会增加阅读量, 也会增加评论量和点赞量, 用户在阅读的同时, 可能会发表评论表达自己的观点, 对于同一观点下的用户也会产生点赞行为. 所以利用舆情信息的阅读量、评论量和点赞量来反映舆情信息的关注度的高低是合理的. 因此将阅读量、点赞量和评论量作为舆情信息关注度计算指标. 基于舆情信息关注度计算如式 (3) 所示:

$$Atte\text{-}Info = \log_a(Read_i + Comment_i + Like_i) \tag{3}$$

其中, $Read_i$ 为第 i 条舆情信息的阅读量, $Comment_i$ 为第 i 条舆情信息的评论量, $Like_i$ 为第 i 条舆情信息的点赞量.

1.2.3 传播者影响力计算

传播者的影响力由传播者的定量数据和传播者的相关信息计算得到. 主要由 3 个指标构成: 传播者的粉丝数量、传播者的领域与舆情话题的相关度和传播者最近发表文章的频率. 传播者的粉丝数越高, 受关注的程度越大, 更多的用户会浏览到相关的信息, 传播的范围也越广. 传播者的领域与舆情话题的相关度反映了传播者的权威可信性, 发布的信息与自己领域相关度越高, 传播者的影响力越大. 同时传播者的发博率反映了传播者的活跃性, 活跃性高的用户发布文章的频率越高, 参与传播的积极性越高. 则针对传播者影响力的计算如式 (4) 所示:

$$Com\text{-}Info = (\log_a(Fans+1)) \times FieldSim_i \times \frac{PostNum}{T} \tag{4}$$

其中, $Fans$ 为传播者粉丝数量, $PostNum$ 为传播者最近一段时间发布的文章数量, T 为发布 $PostNum$ 的文章数量所用的时间. $FieldSim_i$ 为传播者的领域与第 i 条舆情信息的相关度, 利用 Gensim 模块计算传播者的简介和认证信息与发布舆情信息的相关度.

1.2.4 信息时效性

舆情信息发布的时间是判定舆情信息有用性的重要指标之一, 发布时间越早的信息, 越容易在传播过程中被其他信息覆盖, 信息的传播力会降低. 相反, 发布时间越晚的信息, 信息比较新, 受关注的程度较高. 所以在发掘有用性高的舆情信息时需要考虑到发布时间的指标, 舆情信息的发布时间可以转为相应的持续时间来计算 (以小时为单位), 针对所有舆情信息的持续时间需要进行统一处理, 使用 Python 的 datetime 模块中的 timedelta 函数做持续时间的计算, 持续时间的计算如式 (5) 所示:

$$T = \frac{Delta(time_i - time_s) + 1}{3600} \tag{5}$$

其中, $time_i$ 为当前舆情信息的发布时间, $time_s$ 为所有舆情信息中最早发布的时间. $Delta$ 则代表用于计算当前发布时间和最早发布时间的时间差.

1.3 网络舆情信息有用性排序

SPHR 算法综合考虑了舆情信息的特定观点、可信度和关注度, 传播者的影响力及信息时效性多个因素, 利用熵值法和改进的 Reddit 排序算法对每条舆情信息进行打分, 根据得分进行有用性排序, 将得分高的 N 条舆情信息 (N 是根据专家经验预设的信息数量) 输出为有用性排序列表, N 可以动态调整.

Reddit 排序算法是一种基于投票的网络社区文章

排序方法, 根据每篇文章得票和发布时间进行评分, 实现“最佳文章”的排名[17]. 通过比较其他文章排序算法, Reddit 方法根据投票结果, 将赞成票和反对票巧妙的结合起来, 同时联合发布文章的时间对所有文章进行评分排序, 与提出 SPHR方法的契合度较高. 本文主要对 Reddit 算法进行改进, 通过对 Reddit 算法公式的改进, 进行参数重新定义, 使得更加适合本文的研究思路, 则改进后的排序算法公式如式 (6):

$$Score = \left(\log_a Info + \frac{T}{12.5}\right) \times S \tag{6}$$

其中, T 为舆情信息在舆情事件中的一个时效性指标, 由式 (5) 计算得到, T 越大, 说明信息发布的时间较新, 即信息有用性得分就越高. S 表示特定观点下的态度, 由第 1.1 节分析得到. *Info* 表示综合影响力, *Info* 越大, 舆情信息有用性得分越高. *Info* 是由舆情信息可信度、舆情信息关注度和传播者影响力计算得到. 本文使用信息有用性 *Info* 代替原 Reddit 算法公式中的投票数, 用来更加准确的表示舆情信息和传播者对有用性排序的影响. 舆情信息综合影响力 *Info* 的计算如式 (7) 所示:

$$Info = k_1 Cre\text{-}Info + k_2 Atte\text{-}Info + k_3 Com\text{-}Info \tag{7}$$

其中, *Cre-Info* 为舆情信息可信度, 由式 (2) 计算可得, *Atte-Info* 为舆情信息的关注度, 由式 (3) 计算可得, *Com-Info* 为传播者影响力, 由式 (4) 计算可得. 并且 k_1−k_3 作为 3 项的影响权重.

式 (7) 中的各项权重 k_1、k_2 和 k_3 采用熵值法得到其权重值. 熵值法是一种多指标的综合加权评价方法, 根据各项指标观测值所提供信息的大小确定指标权重值的一种客观赋值方法.

熵值法是依据指标所提供的信息, 根据指标离散性的大小来客观确定权重. 某个指标计算得到的信息熵值越小, 离散程度越大, 提供的信息量也越多, 代表该指标在综合体系中的权重值也就越大. 反之, 权重值也就越小[18]. 熵值法首先对数据各个指标进行去量钢化处理, 然后根据信息熵的计算公式得到各个指标的信息熵, 最后通过信息熵计算各指标的权重. 是一种客观的赋权方法, 避免了人为因素带来的权重误差, 精度较高客观性也更强, 能够很好的解释所得到结果. 具体权重计算如式 (8) 所示:

$$W_i = \frac{1 - E_i}{k - \sum E_i},\ i = 1, 2, \cdots, k \tag{8}$$

其中, E_i 表示第 i 个指标的信息熵, k 指的是指标的个数.

利用熵值法对各项指标进行权重赋值, 具体结果如表 2 所示.

表 2 熵值法计算权重结果

项目	熵值	效用值	权重占比 (%)
舆情信息可信度	0.9709	0.0291	36.43
舆情信息关注度	0.9769	0.0231	28.94
传播者影响力	0.9724	0.0276	34.63

通过表 2 计算结果, 得到权重占比, 即权重值, 所以可以确定各个项的权重值: k_1=0.3643, k_2=0.2894, k_3=0.3463, 将权重值代入式 (7) 计算每条舆情信息的综合影响力, 然后利用改进的排序算法式 (6) 进行舆情信息有用性计算, 按照得分高低进行排序, 并把排序靠前的 N 条舆情信息返回作为舆情信息有用性推荐列表.

2 实验与分析

2.1 数据集与实验设置

实验选用今日头条作为数据平台, 今日头条是一个通用信息平台, 作为国内综合类新闻资讯平台之一, 在数据支撑上具有较好的代表性. 本文采用编写的爬虫程序, 并结合爬虫软件, 从今日头条平台上采集“新冠病毒溯源”“新冠疫苗”两个舆情话题下的数据用于测试所提出的 SPHR 方法. 实验采集的数据共计 7 600 条, 包含文章信息数据, 文章阅读、点赞、评论数据及相关的发布者信息. 然后对数据进行了标注工作. 主要包括信息观点态度的标注和信息有用性的标注. 为后续实验的开展奠定基础.

基于现有的数据集, 设计了两组实验进行分析, 具体如下:

实验 1. 利用提出的 SPHR 方法对舆情信息进行有用性排序, 对排序结果进行评价.

实验 2. 针对提出的 SPHR 方法与传播指数 TGI 的计算方法进行对比实验.

2.2 实验结果与分析

实验 1. 为了验证提出的 SPHR 方法的合理有效, 利用提出的 SPHR 排序方法对舆情信息进行有用性排序, 排序结果使用归一化折损累计增益 (*NDCG*) 进行评估. *NDCG* 常用于推荐任务中, 推荐任务中常返回一个 item 列表, *NDCG* 用来衡量这个返回列表的好坏程度, 是一种用来评估排序结果的评价指标. 针对 SPHR 排序方法返回的舆情信息有用性排序列表, 采用 *NDCG*

方法来衡量排序的好坏是合理化的.

NDCG 是由折损累计增益与理想的折损累计增益比值得到的, 即$NDCG = \frac{DCG}{IDCG}$, *DCG* 考虑到排序顺序的因素, 使得排名靠前的项增益更高, 对排名靠后的项进行折损[19]. *DCG* 的计算如式 (9) 所示:

$$DCG = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)} \tag{9}$$

其中, k 为返回列表的数目, i 为当前列表的项; rel_i 表示第 i 个 *item* 的相关性得分.

SPHR 算法中实验设置 N (返回排序列表信息的数量) 分别设置 10, 15, 20, 30. 不同的 N 值对应不同的 NDCG 评估结果, 实验结果如图 2 所示.

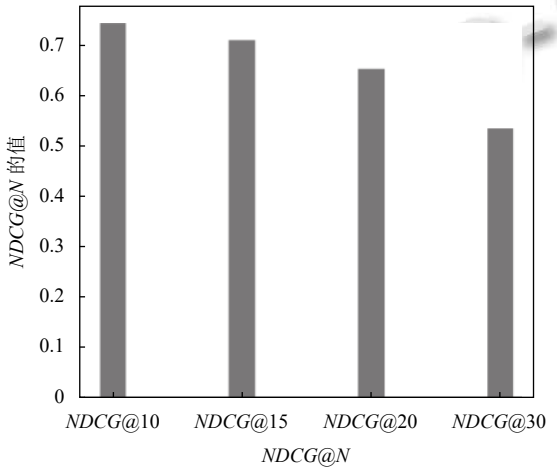

图 2 *NDCG@N* 实验结果

通过图 2 实验结果可以看出, 所提出的 SPHR 算法在舆情信息有用性排序上有一定的效果, 并且当返回列表在 20 以下时, *NDCG* 的评价值达到了 0.7 以上, 这是非常可观的效果, 所以提出的 SPHR 算法具有合理和可用性. 但是也能够发现, 随着返回列表值的增加, *NDCG* 的值有所降低, 主要原因是因为随着返回列表数量的增加, 排序出来的列表信息的折损累计增益在下降, 所以这种现象是正常的, 符合实际情况.

实验 2. 为了进一步验证 SPHR 方法对舆情信息有用性具有很好的排序效果, 引入清博舆情头条号传播指数 TGI 作为 SPHR 方法的对比方法, 并且依据人工标注样本辅助进行实验结果分析. 人工排序的方法较直观的体现了用户的需求, 因此被认为是最佳排序结果[20]. 在实际场景中, 舆情管理者会倾向于关注平台中传播能力和效果较好的文章, 这些文章一般通过相关指数计算得到. TGI 是通过计算头条号的活跃指数, 传播指数, 互动指数来反映传播能力和效果. 选取 TGI 指数作为对比方法分析具有一定的合理性.

首先通过提出的 SPHR 算法对舆情信息进行有用性排序, 然后根据头条号传播指数 TGI 计算结果对舆情信息进行排序, 最后比较两次排序结果与人工标注排序的重合率. 设置的重合率计算如式 (10) 所示:

$$OverlapRatio = \frac{RankInfo(label=2)}{RankNum} \times 100\% \tag{10}$$

其中, *RankInfo*(*label*=2) 为排序后的列表中有用性高的舆情信息数量, *RankNum* 为实验中预设返回列表的舆情信息总数量.

实验 2 中设置的返回列表个数 *RankNum* (即 N 值) 依次为 10, 15, 20, 30. 利用 SPHR 算法和 TGI 指数分别得到舆情信息排序列表, 通过重合率计算得到的实验结果如图 3 所示.

从上述实验对比分析中可以看出, SPHR 算法对舆情信息有用性排序与人工标注结果的重合率均在 70% 左右, 整体上来看 SPHR 算法对舆情信息的有用性排序有良好的效果, 能够有效的对舆情信息进行有用性排序推荐. 同时 SPHR 算法对舆情信息有用性排序与人工标注的重合率高于 TGI 排序的结果. TGI 算法利用了当前可获取发文数、阅读数和评论数进行数据量化, 对舆情信息进行排序, 没有对舆情信息内容进行特定观点态度分析, 忽略了信息源的权威性、信息独立性和信息的时效性等因素. SPHR 算法则综合考虑了多个指标特征, 对舆情信息进行有用性排序.

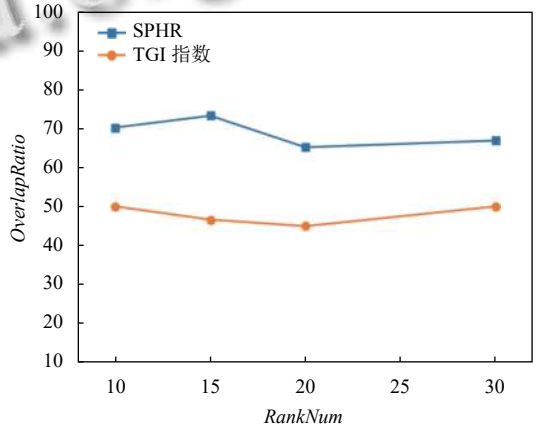

图 3 重合率实验结果

通过上述两个实验结果分析, SPHR 算法的排序结果利用 *NDCG* 评价指标进行评估, 实验结果表明能很好的完成对舆情信息的推荐排序; 同时也将 SPHR 方

法与清博舆情的头条号传播指数 TGI 做对比实验, SPHR 方法排序效果良好, 优于 TGI 指数计算得到的排序结果. 所提出的 SPHR 算法能够合理、有效的对舆情信息进行有用性排序, 对舆情决策者有很好的辅助作用, 具有一定的理论和实践意义.

3 结论与展望

本文主要研究了舆情信息的有用性排序问题, 针对如何有选择性的发掘可能对舆情工作者有用性高的网络舆情信息, 提出了一种面向特定观点的网络舆情信息有用性排序方法 (SPHR). 该方法针对舆情信息的特定观点分析, 可以针对性深入的了解特定观点对于整个舆情事件的影响程度; 同时综合考虑了舆情信息可信度、关注度以及传播者的影响力, 结合时效性因素, 利用改进的 Riddit 排序方法对舆情信息进行有用性排序, 实现快速发掘特定观点下有用性高的舆情信息. 实验结果表明, 本文提出的 SPHR 算法能够合理有效的对舆情信息进行有用性排序, 验证了算法的可行性和有效性. 具有一定的理论和实践意义.

本实验的数据规模较为有限, 主要针对今日头条平台的两类舆情话题的数据进行抓取与分析, 鉴于舆情话题众多, 下一步的研究将扩大实验数据的规模, 同时舆情信息有用性影响因素的相关研究也有待完善, 在此基础上, 后需的研究要考虑到舆情信息传播内容的针对性、内容多样性等因素对舆情信息有用性的影响, 用来更好的发掘有用性高的舆情信息.

参考文献

1 米昂. 结合影响力分析的微博舆情溯源研究 [硕士学位论文]. 北京: 北京交通大学, 2015.

2 刘一飞. 网络舆情信息识别与分析的关键技术研究 [硕士学位论文]. 成都: 电子科技大学, 2020. [doi: 10.27005/d.cnki.gdzku.2020.000640]

3 吴春琼, 鄢冰文, 郁榕睿, 等. 基于大数据的网络群体信息认知研究——海量网络舆情信息主题提取研究. 信息系统工程, 2020, (12): 139–140. [doi: 10.3969/j.issn.1001-2362.2020.12.062]

4 黄微, 刘熠, 许烨婧, 等. 网络舆情推文的热度测度模型构建. 图书情报工作, 2019, 63(20): 17–25. [doi: 10.13266/j.issn.0252-3116.2019.20.002]

5 Luo C, Luo XR, Bose R. Information usefulness in online third party forums. Computers in Human Behavior, 2018, 85: 61–73. [doi: 10.1016/j.chb.2018.02.041]

6 Malik MSI. Predicting users' review helpfulness: The role of significant review and reviewer characteristics. Soft Computing, 2020, 24(18): 13913–13928. [doi: 10.1007/s00500-020-04767-1]

7 郭顺利, 步辉. 融合 Word2Vec 和 WGRA 的社会化问答社区答案有用性排序方法研究——以携程问答为例. 图书情报工作, 2021, 65(23): 126–135. [doi: 10.13266/j.issn.0252-3116.2021.23.014]

8 王建文. 基于信息采纳视角的在线评论有用性排序研究. 现代计算机, 2019, (11): 67–71. [doi: 10.3969/j.issn.1007-1423.2019.11.013]

9 李学明, 张朝阳, 余维军. 基于用户回复内容观点支持度的评论有用性计算. 计算机应用, 2016, 36(10): 2767–2771, 2776. [doi: 10.11772/j.issn.1001-9081.2016.10.2767]

10 Gilly MC, Graham JL, Wolfinbarger MF, *et al*. A dyadic study of interpersonal information search. Journal of the Academy of Marketing Science, 1998, 26(2): 83–100. [doi: 10.1177/0092070398262001]

11 张博, 李竹君. 微博信息传播效果研究综述. 现代情报, 2017, 37(1): 165–171. [doi: 10.3969/j.issn.1008-0821.2017.01.031]

12 Wu SM, Hofman JM, Mason WA, *et al*. Who says what to whom on Twitter. Proceedings of the 20th International Conference on World Wide Web. Hyderabad: ACM, 2011. 705–714.

13 周庆山, 梁兴堃, 曹雨佳. 微博中意见领袖甄别与内容特征的实证研究. 山东图书馆学刊, 2012, (1): 22–27, 35. [doi: 10.3969/j.issn.1002-5197.2012.01.006]

14 Mihalcea R, Tarau P. TextRank: Bringing order into text. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona: Association for Computational Linguistics, 2004. 404–411.

15 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. arXiv:1301.3781, 2013.

16 郭丽娟, 孙世宇, 段修生. 支持向量机及核函数研究. 科学技术与工程, 2008, 8(2): 487–490. [doi: 10.3969/j.issn.1671-1815.2008.02.041]

17 Salihefendic A. How Reddit ranking algorithms work. Hacking and Gonzo. 2010: 23.

18 刘肇民, 郑红玲. 基于熵权 TOPSIS 的区域经济新动能测度研究——以河北省为例. 河北北方学院学报 (社会科学版), 2021, 37(4): 65–71.

19 黄红涛. 校园信息精准推送系统设计与实践. 现代信息科技, 2021, 5(8): 1–4. [doi: 10.19850/j.cnki.2096-4706.2021.08.001]

20 程亚男, 王宇. 基于语义情感相似度的问答社区答案排序研究. 情报科学, 2018, 36(8): 72–76, 83. [doi: 10.13833/j.issn.1007-7634.2018.08.012]

(校对责编: 孙君艳)