



开放互联网环境基于信息熵的信息传播影响力计算方法

李攀攀¹, 谢正霞², 王赠凯¹, 靳锐³

(1. 嘉兴学院信息科学与工程学院, 浙江 嘉兴 314001;

2. 嘉兴学院建筑工程学院, 浙江 嘉兴 314001;

3. 哈尔滨工业大学计算机科学与技术学院, 黑龙江 哈尔滨 150001)

摘要: 研究了开放互联网环境下衡量信息传播的影响力, 提出了一种基于最大熵理论的信息传播影响力计算方法。首先, 对信息传播影响力所涉及的要素进行建模; 然后, 以香农信息熵理论为基础, 根据“奥卡姆剃刀原理”选取信息影响力关联指标; 最后, 基于信息熵理论给出了信息传播影响力的计算方法。对学术论文引文和网络舆情事件的传播影响力进行了分析, 验证了所提方法的有效性, 所提方法对探究信息在开放互联网环境中的传播影响力以及对受众的影响程度具有重要的现实意义。

关键词: 开放互联网; 传播影响力; 信息熵; 最大熵理论; 计算传播学

中图分类号: G206

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2022026

Calculation method of information dissemination based on information entropy in public internet

LI Panpan¹, XIE Zhengxia², WANG Zengkai¹, JIN Rui³

1. College of Information Science and Engineering, Jiaxing University, Jiaxing 314001, China

2. College of CML Engineering and Architecture, Jiaxing University, Jiaxing 314001, China

3. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

Abstract: A method of how to calculate the influence of information dissemination based on the maximum entropy theory was presented, which aimed at evaluating the impacts of the information dissemination in the open internet environment. Firstly, the influential factors of the information dissemination were analyzed. Then, the factors that have high correlation with the calculation of information dissemination were selected by the principle of “Occam’s Razor”, based on Shannon’s information entropy theory. Finally, the method of computing the influence of information dissemination was proposed based on the information entropy theory. The effectiveness of the proposed method was validated by the experiments of analyzing the influence of the information dissemination of academic paper citations and hot public opinion events. The experiments show that the proposed method is of great significance for investigating the influence of information dissemination and the extent of impacts on the audience as well as the trends of information evolution in the open internet environment.

Key words: public internet, communication influence, information entropy, maximum entropy theory, computational communication

收稿日期: 2021-11-04; 修回日期: 2022-01-29

基金项目: 国家自然科学基金资助项目 (No.61902226); 浙江省自然科学基金资助项目 (No.LY18F020021)

Foundation Items: The National Natural Science Foundation of China (No.61902226), The Natural Science Foundation of Zhejiang Province (No.LY18F020021)

0 引言

信息技术的更新迭代一直是社会发展的核心推动力之一,科学技术的关键创新推动了信息传播方式的重大变化,信息传播一直朝着速度更快、受众面更广、传播成本更低的方向发展^[1]。当今,在泛互联网空间环境下,信息传播的门槛被极大地降低,信息传播速度及传播模式的变革深深影响了人类社会的发展^[2-3]。作为衡量信息传播能量的重要指标,信息影响力的度量在研究信息传播过程中扮演着重要的角色,因此,研究信息在开放互联网环境中的传播影响力、对受众的影响程度以及探究信息演化的趋势具有重要的现实意义。

互联网的出现改变了信息的传播模式,一方面,信息传播的广度显著增加,所传播范围的不确定性日益突出,相比传统环境,信息共享与传播的边界不断扩大,人们能更便捷地享受信息高效传播带来的便利;另一方面,共享开放的网络空间给信息影响力的传播带来了深刻的变革,信息传播的跨时空性给量化信息的传播影响力带来更为严峻的挑战^[4-5]。

传统的信息传播影响力计算方法有如下研究。在信息影响力关键要素抽取方面,文献[6]使用信息熵建立了中文公众事件的抽取方法,未对传播影响力进一步研究;在传播影响力分析方面,文献[7]和 InfoIBP (influence propagation Indian Buffet process) 模型^[8]侧重于信息主体与信息内容之间相互影响力的分析,忽视了信息之间的传播影响力,而基于粒子群的 LAPSO-IM (a learning automata based discrete particle swarm optimization influence maximization) 方法^[9]或基于蚁群优化的 ACO-IM (ant colony optimization influence maximization) 方法^[10]通常依赖大规模特定领域的数据集作为初始训练集,应用范围受限;在关键传播节点或关键事件影响力分析方面,文献[11]和文献[12]分别使用贝叶斯方法和实测方法进行了研

究,该类方法通常忽略了信息的跨领域传播。

针对上述问题,本文研究了开放互联网环境下多维信息因素的建模,充分考虑了互联网环境的开放性、多维性等特点,研究了基于最大熵理论的信息传播影响力计算方法。互联网中海量的信息并非互相独立,随着互联网信息共享能力的提升,大多数传统的信息“孤岛”和数据“壁垒”已被打破,用户通过互联网接收传播信息的同时,其认知也同样经历着影响与被影响的过程。互联网信息传播的影响力涉及信息对信息、信息对用户、用户对用户等多个方面^[13-15]。理论上讲,传播影响力是衡量信息重要程度的内在特征属性,如何量化这种影响力的传播过程,探究信息传播的影响力是研究信息传播及其演化模型的基础。

在传播速度更快、环境复杂性更高的公共互联网环境中,评价与量化信息传播影响力主要面临的挑战有以下方面。

- 互联网信息影响力涉及因素众多,如信息主体、传播路径、客体影响等,在开放的公共互联网中难以获得完备的信息传播影响力要素集。
- 互联网信息类型多样,如文本、图片、音/视频等,信息影响力各因素极其复杂,尤其是各因素相互叠加影响,难以统一量化。
- 信息传播渠道多样化,如社交网络转发引用、新闻评论、学术论文引用与评述等,不同渠道难以用统一的量纲描述信息的影响力。

鉴于此,本文提出了一种在公共互联网环境基于信息熵的信息传播影响力计算方法,通过“奥卡姆剃刀原理”选取关键特征。针对多维变量权重的相互影响,难以用传统的统计学函数分布描述问题展开研究,本文采用最大熵理论探究并量化信息在公共互联网中传播的影响力,为信息传播影响力的可度量、可计算方面的研究提供了理论基础和实践指导。



1 互联网信息传播动力学模型

1.1 信息传播影响力模型

经典的影响力传播模型主要有两种：线性阈值模型（linear threshold model, LT）和独立级联模型（independent cascade model, IC）。这两种基于概率的传播影响力模型^[16]在计算过程中，通过引入随机性参数在一定程度上体现了信息传播的特点，LT 在信息传播前为节点随机地选择概率的阈值，IC 中一个节点通常以概率 p 激活邻居节点。信息实体 A, B, ..., H 基于 LT 的信息影响力传播过程如图 1 所示。

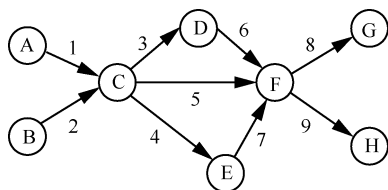


图 1 信息实体 A, B, ..., H 基于 LT 的信息影响力传播过程

从图 1 可以看出，LT 通过随机地选择阈值作为计算信息传播影响力的基础，显然这种随机性难以准确反映信息传播的具体环境特点，尤其在开放的互联网环境。从理论上讲，信息传播影响力最大化求解过程可抽象为一个子集上面向特定目标的查找组合优化问题，因此，基于 LT 和 IC 的影响力最大化问题属于非确定性多项式（non-deterministic polynomial, NP）难问题^[7]，难以适用于开放互联网中信息传播影响力的计算，主要原因有以下两点。

- 这两种模型的相对随机性难以适用于互联网信息影响力的计算，主要因为开放互联网中信息影响力涉及因素多且这些因素无统一可计算的量纲。
- LT 概率阈值及 IC 激活概率的随机性使得采用这些模型计算的传播影响力通用性不高。

信息传播影响力涉及内、外部因素诸多，难以量化这些因素与传播影响力之间的关系，

如浏览、引用、点击量等。因此，需要使用多维可计算性衡量信息传播影响力，以使影响力具有通用性。信息传播影响力具备以下 3 个主要特征。

- 可度量性：对信息传播影响力的指标进行综合分析并量化而成的特性，是信息传播影响力可计算的基础。
- 可计算性：不同类型信息传播影响力具备可计算能力，实现信息传播过程中影响力的计算。
- 多维性：信息传播影响力不仅包含较多的因素，而且不同领域的信息影响力具有各自的特性，信息传播力衡量的特征要具备容纳多维特征的能力。

根据文献[18]，信息传播影响力关系可抽象成有向图 $G(V, E, W)$ ，其中， V 是节点（信息实体）的集合， E 是有向边的集合（信息个体之间的影响关系）， W 是信息传播影响力的集合，节点 $v \in V$ 表示一个信息实体，边 $(u, v) \in E$ 表示节点 u 到节点 v 的影响关系^[19]。互联网信息传播过程可从以下两个角度描述。

- 传播广度：信息在互联网被引用、转载、评述等的次数，是对信息直接影响力的度量。
- 传播深度：信息被引用、转载、评述后，还能被后续信息引用、转载、评述的次数，是对信息间接影响力的度量。

信息的传播过程由信息及信息之间的引用、转载、引述、评论等关系组成一个复杂网络，这种复杂的信息传播拓扑结构对信息影响力的传播和扩散起着至关重要的作用。

1.2 影响力传播形式化表示方法

本书详细讨论特征参数的选取规则。设信息为全集 U ，由 n 个子集构成： $U_1, U_2, \dots, U_i, \dots, U_n$ 。假设 $n=10$ ， U_1 表示信息主题集合， U_2 表示信息类型集合， U_3 表示所有者集合， U_4 表示出版源集合， U_5 表示时间集合， U_6 表示被引用或被转载次

数, U_7 表示所属领域集合, U_8 表示关键词集合, U_9 表示摘要信息集合, U_{10} 表示其他信息 (未抽取信息) 集合。显然, $U=U_1 \cup U_2 \cup \dots \cup U_i \cup \dots \cup U_n$ 。上述形式化描述是一种对互联网信息的简化表述方式, 最大限度地包含了涉及信息传播影响力的要素。

信息传播是信息主体与被影响主体之间强度的表示方式, 由信息影响力元数据 (如信息主体、信息传播路径等) 指标综合而成, 公共互联网信息传播影响力关联要素示意图如图 2 所示, 信息的影响力传播建模基础工作是从影响信息传播的众多因素中选取对传播影响力有较大贡献的特征变量, 即将信息传播影响力量化为求解影响力边际增益或收益最大的问题。

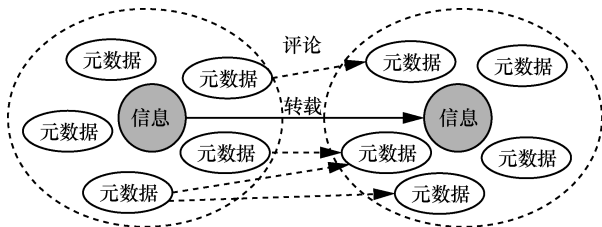


图2 公共互联网信息传播影响力关联要素示意图

在评估信息传播影响力时涉及信息的引用、转发、评论等多个参数, 这些参数难以分别具体量化, 参数之间的相互影响程度难以评估, 因此, 本文采用最大熵理论作为信息影响力传播的重要理论依据, 即从信息熵的原理出发, 在评估信息的传播影响力时, 选择在满足全部已知条件的情况下, 对未知的情况不做任何主观假设^[7-8,19]。

1.3 面向传播影响力的多维特征选择

合理的特征选择是量化信息传播影响力的基础, 根据“奥卡姆剃刀原理”进行特征工程选取, 即在特定的环境下, 选用信息传播影响力最大的特征作为其数学模型的参数, 基于信息熵和“奥卡姆剃刀原理”信息传播影响力特征选择流程如图 3 所示。

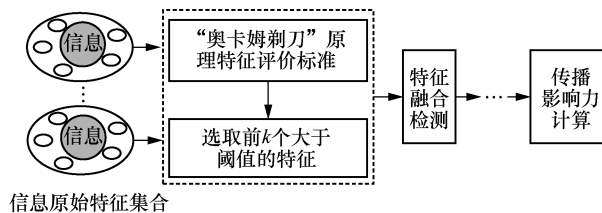


图3 基于信息熵和“奥卡姆剃刀原理”信息传播影响力特征选择流程

在图 3 中, 信息影响力可表示为 $ID_{INIA \rightarrow INIB}$, 显然, 影响力是具有方向性的矢量, 因此, 信息客体受信息主体的影响力如式 (1) 所示。

$$ID_{INIA \rightarrow INIB} := \{ID_A, ID_B, f: ID_A \rightarrow ID_B\} \quad (1)$$

其中, $f: ID_A \rightarrow ID_B$ 表示信息 A 对信息 B 的影响力, 需要指出的是 $f: ID_A \rightarrow ID_B \neq f: ID_B \rightarrow ID_A$ 。

需要进一步指出的是, 在参数规则建模过程中, 没有选取信息类型, 是因为通过信息摘要能确定信息的所属领域。根据上述分析, 对于有 n 个随机变量组成的信息实体, 其传播影响力可以表示为 $(X_1, X_2, \dots, X_n)$, 信息的影响力组成一个随机信息系统 X 。因此, X 等价于 $(X_1, X_2, \dots, X_n)$, 其联合概率分布为 $(p_1, p_2, \dots, p_i, \dots, p_n)$ 。

2 基于信息熵的信息传播影响力计算方法

作为一种系统中信息含量的分析方法, 信息熵理论已被广泛运用于物理学、信息技术、统计学等领域^[3,20]。在互联网信息传播领域, 信息传播影响力实质是信息传播力的熵值随着信息的传播和扩散阶段不断减小的过程, 消除了信息传播影响力的不确定性和无序性, 从而形成信息传播影响力不断提升的过程, 因此, 信息熵理论非常适合在信息传播影响力场景中展开研究^[21-22]。

2.1 互联网信息影响力的信息熵计算方法

在开放互联网环境下, 信息的影响力涉及较多的因素, 如信息的发布者、信息的转发者、信息的评论次数、引用者身份、发布渠道等, 这就造成了在衡量信息传播影响力时, 难以量化各指标的权重及其相互之间影响程度的权重, 因此,



本文使用最大熵模型解决此问题,用最大熵模型描述并量化信息传播的影响力。

针对不同的信息构建完备合理的语料库是提取信息特征的基础,本文通过不同方面的领域专家构建语料库,并针对不同的应用环境使用不同的语料库。信息的特征选择主要包括以下3个步骤。

步骤 1 使用领域知识对信息进行文本分词以及词频权重统计,分词 t_i 词频权重统计 $\text{wf}(t_i)$ 方法如式(2)所示。

$$\text{wf}(t_i) = \lg\left(\frac{N}{n_i} + 1\right) \times \lg\left(\frac{m_i}{M_i} + 1\right) \quad (2)$$

其中, N 为总的信息数, n_i 为包含 t_i 信息的总数, m_i 为特征词在该类中出现的次数, M_i 为该类信息的总数。

步骤 2 计算分词之间的逐点互信息(pointwise mutual information, PMI), PMI 用于衡量信息变量间相互依赖的程度,因此 n_i 和 n_j 使用联合概率分布的计算过程如式(3)所示。

$$\text{PMI}(n_i, n_j) = \lg \frac{p(n_i, n_j)}{p(n_i)p(n_j)} \quad (3)$$

其中,若 n_i 和 n_j 无关,则有 $p(n_i, n_j) = p(n_i)p(n_j)$, 此时 $\text{PMI}(n_i, n_j) = 0$; 若 n_i 和 n_j 的相关性越大,那么 $p(n_i, n_j) / (p(n_i)p(n_j))$ 越大,随之 $\text{PMI}(n_i, n_j)$ 越大。

步骤 3 根据“奥卡姆剃刀原理”筛选关键特征,构成信息的特征集。

经过上述步骤,信息的特征是离散的随机变量,用 X 表示影响信息传播影响力的离散特征值集合,设离散变量 X 的概率分布是 $P(X)$, 根据香农信息熵理论,离散随机变量 X 的熵 $H(X)$ 如式(4)所示。

$$H(X) = -\sum_x P(x) \lg P(x) = -\lg p(x_1, x_2, \dots, x_n) \quad (4)$$

随机变量 X 的熵 $H(X)$ 是 $p(x)$ 概率分布函数,用于衡量信息的传播强度。当 $p(x)=0$ 时,定义

$0 \lg(0)=0$ 。从式(4)可以看出, $H(X)$ 仅依赖于 X 的分布,而与 X 的具体取值无关。还可以看出,熵满足 $0 \leq H(P) \leq \lg|X|$, 其中, $|X|$ 是 X 取值的个数,当且仅当 X 均匀分布时右边的等号成立。

采用拉格朗日乘子法求解在一定约束条件下 X 的分布情况,将最大信息熵问题转化为无约束极值问题进行求解,记拉格朗日函数 $L(p(x_i)) = H(X)$, 文献[6]采用拉格朗日乘子法求解该问题,得到熵函数 $H(x)$ 如式(5)所示。

$$H(x) = -\sum_x P(x) \lg(P(x) = \ln(x)) \quad (5)$$

从式(5)可以看出,对于取值有限的离散型随机变量 X 来说,当每个变量的取值相等时,其信息熵最大,即当 X 服从均匀分布时熵最大。信息的影响深度可以表示该信息对后续信息的影响强度,如学术领域的高被引论文通常被认为具有较高的影响力,新闻信息被转载或评论的次数愈多其影响力也就愈大。因此,若使信息的传播影响力最大,则式(5)满足 $P(x) = \frac{1}{|X|}$ 。

通过上述推理证明,可得出以下两个结论。

结论 1 X 均匀分布时, $H(X) = \lg n$, 根据定义,此时信息传播影响力达到最大。

因此, $H(x)$ 极值点在 $p_1 = p_2 = \dots = p_n = \frac{1}{|X|}$ 时,此时有 $H(X) = \lg n$ 。

结论 2 $H(x)' = \frac{1}{x}$, $x > 0$, 推导出 $H(X)' > 0$,

因此,信息熵 $H(X)$ 是关于 x 的严格单调函数。

其中,结论1和结论2保证了信息传播影响力计算结果的严格单调性,是本文方法可用性的基础。

举例说明,在 $n=2$ 时, $H(P)$ 的表达式可写成: $H(P) = -\sum_x P(x) \lg(P(x))$ 。

最大熵理论认为所采用的概率模型首先必须满足既存事实,即约束条件,同时不对未知的情况做任何主观假设,即在所有可能的概率模型(分布)中。因此,使用最大熵模型评估信息影

响力在理论上是可行的。此外,从信息论相关理论上,信息传播的最大熵模型中往往包含更多的关联信息影响力特征,用最大熵计算的影响力也相应最大。包含更大信息量的信息容易造成较大的影响力,根据前文,在(0~1)分布的函数、概率 $r=0.5$ 时,熵值 $H(X)$ 最大,这也很好地解释了互联网热点事件或有重大影响属性的事件有较高传播影响力的原因,与文献[6]的研究结论吻合。

2.2 基于最大熵理论的信息影响力定量计算流程

基于最大熵理论的信息影响力定量分析方法需要经过一定的预处理流程,以便于将信息分解为离散、可被计算的参数^[23],开放互联网环境基于信息熵的影响力计算示意图如图4所示。

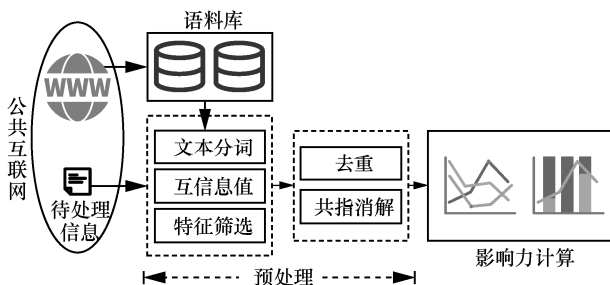


图4 开放互联网环境下基于信息熵的影响力计算示意图

从图4可以看出,首先,结合社会学理论以及领域知识,利用搜索引擎、网络爬虫等工具在互联网中提取海量领域信息建立合理且完备的语料库;其次,使用语料库对待处理的信息进行分词处理,并由领域专家计算逐点互信息;再次,由领域专家使用基于“奥卡姆剃刀原理”的关键特征筛选方法进行预处理,经过上述预处理后,待处理信息被量化为可计算的离散变量;最后,将这些离散变量进行去重和共指消解以提升后续影响力计算的准确性^[6]。

具体来说,互联网信息影响力的信息熵计算方法如算法1所示。

算法1 互联网信息影响力的信息熵计算方法

输入 信息 $t_i (i \in |G|)$ 及其传播拓扑图 G 、

语料库 D

输出 信息 t_i 的影响力 p_i

for $i=0, i < |t_i|, i++ \{$

for $j=0, j < |t_j|, j++ \{$

使用语料库 D 对信息 t_i 进行分词

词频及权重统计

计算逐点互信息PMI

提取 t_i 特征集

$\}$

对信息 G 采用广度优先遍历分析信息之间的影响与被影响的关系

执行去重流程

执行共指消解流程

根据拉格朗日乘子法计算 t_i 的影响力信息熵 p_i

return p_i

$\}$

综上分析,本文所提方法支持了信息传播过程中影响力的可度量、可计算并支持多维性因素,并通过领域专家支持了信息的跨领域传播过程中影响力的计算。

3 信息传播力影响力熵定量分析

3.1 信息传播影响力建模

使用随机变量 X 表示信息传播影响力,假设 $n=10$,那么 X 等价于 $(X_1, X_2, \dots, X_{10})$,其中, $X_1, X_2, \dots, X_{10}$ 是信息传播影响力指标的参数化表示。 $X_1, X_2, \dots, X_{10}$ 作为影响自变量与熵值 H 的关系无法确定,难以通过具体的函数模型或量化指标描述 $X_1, X_2, \dots, X_{10}$ 与熵值 H 的关系,因此,本问题适用文献[6]所提的最大熵模型加以解决,这使最大熵表示信息影响力的熵值更接近信息在公开互联网传播的实际情况。

3.2 基于信息熵的传播力计算方法

信息传播影响力要素由 $(X_1, X_2, \dots, X_{10})$ 构成,为了计算传播影响力信息熵,需要分别确定各分变量的取值范围,则传播影响力由各分要素组合



构成一个多维向量空间。使用 c_i 表示随机变量 X_i 的取值次数, 当 X_i 的取值次数为 0 时, c_i 为空, 对计算结果无影响。因此, 根据文献[6], 信息传播影响力涉及的多维变量取值见表 1。

表 1 信息传播影响力涉及的多维随机变量取值

X_i	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
c_i	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}

在计算信息传播影响力的最大熵时, 受 $\sum p(c)=1$ 的约束。

根据最大熵理论, 为了获得信息的最大传播影响力, 各影响力要素的概率均匀分布, 有 $p(c_1)=p(c_2)=\dots=p(c_{10})$, 根据约束条件, 则有 $p(c_1, c_2, \dots, c_{10})=1/n$, 因此, 熵函数如式 (6) 所示。

$$H(X) = -\lg p(c_1, c_2, \dots, c_{10}) = -\lg \frac{1}{c_1 c_2 c_3 \dots c_{10}} = \lg(c_1 c_2 c_3 \dots c_{10}) \quad (6)$$

式 (6) 是由互联网信息影响力各特征要素构成的传播影响力计算方法。

4 实验分析

为了衡量本文方法的有效性, 本节使用不同领域的数据集对本文方法进行量化评估。

4.1 实验数据及实验环境

随着互联网的发展和广泛应用, 学术论文的传播以互联网为主要载体, 学术论文的传播路径单一, 通常以引用、评述为主要传播路径, 因此, 研究学术论文影响力可作为评估本文方法的手段之一。不失一般性, 随着 Web3.0、移动互联网、短视频等媒体技术的普及, 互联网舆情事件传播途径多样化。因此, 选用学术论文、热点事件为互联网的内容实体评估本文方法, 数据集有以下两个。

- 针对学术论文信息, 数据集来自 DBLP (DataBase systems and Logic Programming) 学术社交网络, 收集了 1997 年至今公开发表的计算机类学术论文元数据及引

文记录^[24]。物理学论文数据集选取美国物理协会 (American Physical Society, APS) 公布的物理学领域自 1893 年以来论文元数据及其引文记录^[25]。本文选取长期活跃的研究者为研究对象, 为了体现横向对比性, 本文以自然年为时间单位评价和衡量影响力。

- 针对热点舆情事件, 数据集选用近 3 年来有影响力的舆情事件作为研究对象, 评估本文方法的有效性。

在实验环境方面, 本文使用 MATLAB 仿真本文的算法, 并与 LAPSO-IM 方法^[9]和 ACO-IM 方法^[10]对比分析, 其中, LAPSO-IM 方法和 ACO-IM 方法分别是基于粒子群算法和蚁群算法的影响力最大化分析方法。

4.2 基于信息熵的学术论文的传播影响力分析

在开放互联网中, 信息传播影响力与时间紧密相关, 首先分析信息影响力随时间变化的趋势, 针对此在 DBLP 引文数据库中, 随机选取 2000—2005 年每年各 100 篇会议论文和期刊论文, DBLP 中不同论文影响力随时间变化趋势如图 4 所示。

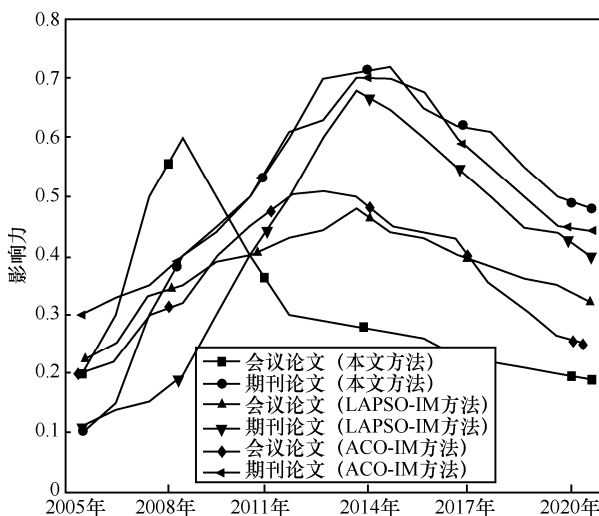


图 4 DBLP 中不同论文影响力随时间变化趋势

从图 4 可以看出, 会议论文的影响力在刊发 1~2 年后影响力持续爆发, 期刊论文的影响力约在 2 年后持续爆发, 这主要是由于会议论

文的评审和刊出时间相比于期刊论文较短,最新的研究成果能更快地在学术会议论文集上刊发,使得其影响力较大,而期刊论文的评审和刊发周期则相对较长,其影响力的爆发时间滞后。与 LAPSO-IM 方法和 ACO-IM 方法对比来看,本文方法计算的影响力对其随时间的变化更为敏感,而另外两种方法基于粒子群和蚁群随时间变化的速度慢,验证了本文方法的高效性。

为了更合理地横向对比各论文的影响力,本文引入了规格化的影响力,将各学术论文的影响力放在统一的量纲对比分析,选取相同主题的论文,按“IPv6”(105篇)、“人工智能”(178篇)两个主题范围进行影响力量化分析,DBLP不同研究方向论文影响力随时间变化趋势如图5所示。

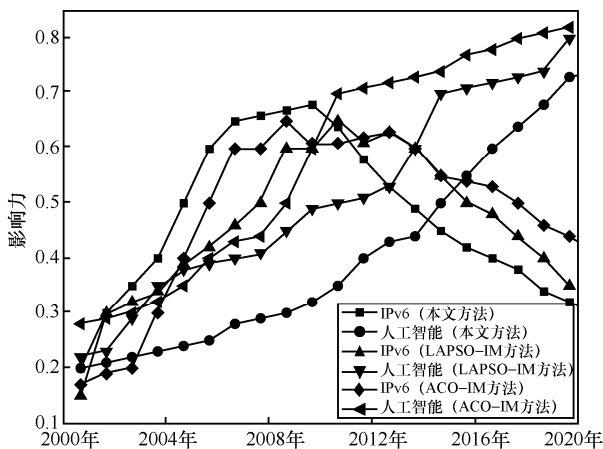


图5 DBLP不同研究方向论文影响力随时间变化趋势

从图5可以看出,“IPv6”向学术论文的影响力约在2010年度达到顶峰,“人工智能”领域的学术论文在2016年左右萌芽,影响力至今仍呈现上升的趋势,这与技术的整体发展趋势相吻合^[6,18],实验结果与LAPSO-IM方法和ACO-IM方法保持一致,验证了本文方法的有效性。

下面进一步地衡量物理学领域相关论文,从物理学引文数据库中随机选取相关论文,物理学领域论文影响力随时间变化趋势如图6所示。

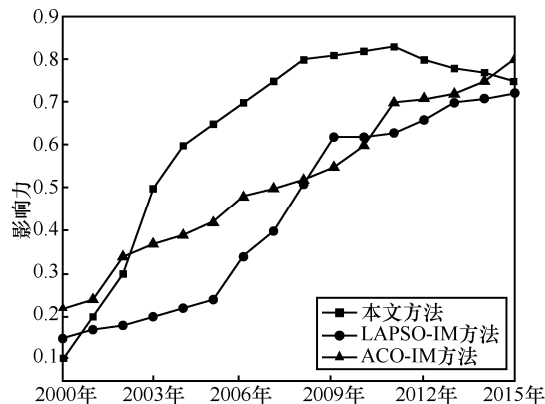


图6 物理学领域论文影响力随时间变化趋势

选取相似领域的多篇文献,以年度为单位进行影响力计算,通过综合对比图4与图6,进一步看出,物理学论文影响力的上升速度较为明显,这主要体现了物理学学术论文的累积效应,而计算机领域的论文的累积效应较弱,这与两个学科的特点有关,物理学学术论文的既往成果对后继成果有较大的影响力,而计算机领域的论文发展迅速,新兴成果的影响力相对较大,本文方法的结果与LAPSO-IM方法和ACO-IM方法计算的结果保持一致。需要指出的是,随着成果应用领域不断拓展,如传统的计算机网络技术向物联网、网络空间安全、社会计算等领域横向扩展的范围不断加大,导致了既往成果对后续关联成果的影响力逐渐降低,而新兴重要成果的影响力较高。

4.3 基于信息熵的网络事件传播影响力分析

学术论文的影响力传播在相对封闭学术团体的范围内,无法综合评估本文方法的有效性。不失一般性,进一步分析开放互联网中网络舆情事件的传播影响力。分别选取2019—2021年热点网络舆情的事件以确保实验样本的丰富性,分别包括文化、科技、民生等领域,范围来自社交网络、大众媒体等,近3年网络舆情事件实验数据集见表2,对舆情事件影响力的量化以发布主题、浏览次数、评论次数和转载次数等指标为特征变量。



表2 近3年网络舆情事件实验数据集

时间	编号	舆情事件	来源
2019 年	事件 1	《哈是佩奇》爆红网络事件	新浪微博
	事件 2	翟天临论文事件	社交网络
2020 年	事件 3	北京一女子开奔驰进故宫事件	新浪微博
	事件 4	浙江温岭槽罐车爆炸事故	人民日报
2021 年	事件 5	比尔·盖茨离婚事件	新浪新闻
	事件 6	第七次人口普查数据公布事件	新华社

通常网络舆情的发展主要有 4 个阶段：潜伏期、扩散期、爆发期和消散期^[6]。针对上述网络舆论事件，选取评论或转载作为受影响的客体衡量其影响力，为了便于横向对比分析，本实验使用加权的影响力计算方法，网络舆情事件随时间影响力变化趋势如图 7 所示。

由图 7 可知，网络舆情事件的影响力随时间呈上升的趋势，但是，影响力增速呈整体下降的趋势，这与直观感受和 PageRank 算法吻合。另外，通过对比事件 1、事件 3、事件 5 与事件 2、事件 4、事件 6，可以进一步发现，事件 4、事件 6 的舆情影响力升上趋势较为强劲，究其原因，这 3 件舆情事件来自有影响力的官方媒体，而另外 3 件的舆情事件影响力发展趋势较为缓和，这 3 件舆情事件主题均来自民间团体，其影响力稍逊于官方媒体，这导致了事件影响力的发展出现差距。与图 4 的研究结果保持一致，对比图 7 (a)、图 7 (b) 和图 7 (c) 可以看出，本文方法较 LAPSO-IM 方法和 ACO-IM 方法所计算的影响力对时间更为敏感，即对事件影响力的响应表现更为积极，这也与粒子群和蚁群算法的收敛速度不高有关。

进一步地，事件 4 和事件 6 官方媒体的舆情事件的扩散期、爆发期影响力上升较为强势，这与官方媒体的权威性有密切联系；而民间力量的潜伏期较长，这与舆情的前期在广大民众中的逐步酝酿的过程有关，互联网中的意见领袖等对事件的评论转述在舆情初期及发展期均能引发更持续的传播影响力。此外，舆情事件的潜伏期与

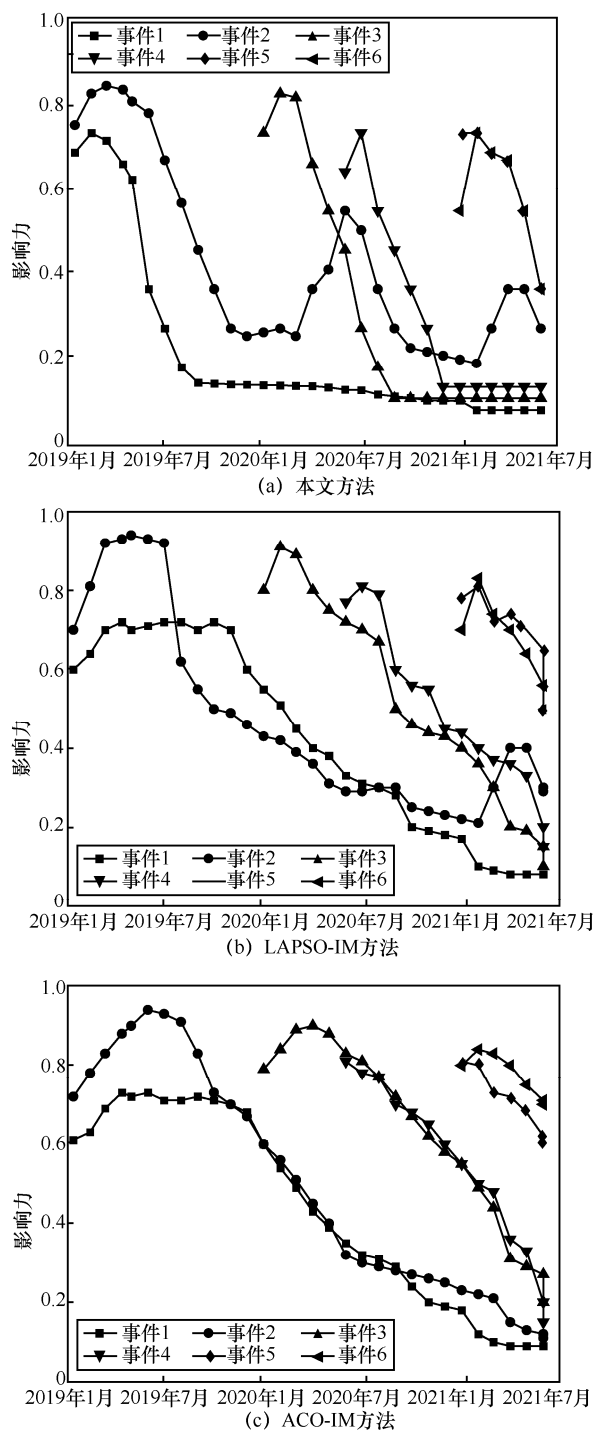


图7 网络舆情事件随时间影响力变化趋势

影响力的大小有关，这也验证了本文方法的有效性。还可以得知，随着时间的变化，不同网络舆情事件的影响力发生一定的变化，影响力不是积累的过程，而是由萌芽期、发生期、爆发期以及平缓期等阶段组成的。

5 结束语

人类社会互联网颠覆了信息传播的方式并深刻影响着社会的发展,研究面向公共互联网的信息传播影响力具有较强的现实意义。本文提出了一种基于香农信息熵的公共互联网信息传播影响力计算模型和方法,首先,根据“奥卡姆剃刀原理”选取信息事件的多维变量,形成信息可计算的形式化表述;接着,使用香农信息熵理论和最大熵理论研究了公共互联网最大影响力建模;最后,提出了基于最大熵的信息传播影响力计算方法,并通过实验验证了方法的有效性和可用性。

参考文献:

- [1] 中国互联网络信息中心. 第47次《中国互联网络发展状况统计报告》[R]. 2021.
CNNIC. The 47th "Statistical Report on the Development of the Internet in China"[R]. 2021.
- [2] 郭宇, 王晰巍, 李师萌, 等. 基于情感分析的社会网络用户影响力模型研究[J]. 情报学报, 2017, 36(11): 1139-1147.
GUO Y, WANG X W, LI S M, et al. Research on the influence model of social network users based on emotional analysis[J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(11): 1139-1147.
- [3] 李根强, 刘莎, 张亚楠, 等. 信息熵理论视角下网络集群行为为主体的观点演化研究[J]. 情报科学, 2020, 38(1): 42-47, 86.
LI G Q, LIU S, ZHANG Y N, et al. Research on the opinion evolution in online collective behavior from the perspective of information entropy[J]. Information Science, 2020, 38(1): 42-47, 86.
- [4] GUILLE A, HACID H, FAVRE C, et al. Information diffusion in online social networks[J]. ACM SIGMOD Record, 2013, 42(2): 17-28.
- [5] 方滨兴, 贾焰, 韩毅. 社交网络分析核心科学问题、研究现状及未来展望[J]. 中国科学院院刊, 2015, 30(2): 187-199.
FANG B X, JIA Y, HAN Y. Social network analysis—key research problems, related work, and future prospects[J]. Bulletin of Chinese Academy of Sciences, 2015, 30(2): 187-199.
- [6] 靳锐, 张宏莉, 张玥, 等. 中文公众事件信息熵计算方法[J]. 软件学报, 2016, 27(11): 2855-2869.
JIN R, ZHANG H L, ZHANG Y, et al. Calculation method of Chinese public event information entropy[J]. Journal of Software, 2016, 27(11): 2855-2869.
- [7] 黄飞虎, 彭舰, 宁黎苗. 基于信息熵的社交网络观点演化模型[J]. 物理学报, 2014, 63(16): 160501.
HUANG F H, PENG J, NING L M. Opinion evolution model of social network based on information entropy[J]. Acta Physica Sinica, 2014, 63(16): 160501.
- [8] 王祯骏, 王树徽, 张维刚, 等. 基于社交内容的潜在影响力传播模型[J]. 计算机学报, 2016, 39(8): 1528-1540.
WANG Z J, WANG S H, ZHANG W G, et al. Social content based latent influence propagation model[J]. Chinese Journal of Computers, 2016, 39(8): 1528-1540.
- [9] SINGH S S, KUMAR A, SINGH K, et al. LAPSO-IM: a learning-based influence maximization approach for social networks[J]. Applied Soft Computing, 2019(82): 105554.
- [10] SINGH S S, SINGH K, KUMAR A, et al. ACO-IM: maximizing influence in social networks using ant colony optimization[J]. Soft Computing, 2020, 24(13): 10181-10203.
- [11] 吴吟, 孙琪. 互联网信息传播中的事件网络关键节点研究: 一项基于网络拓扑学的实证研究[J]. 中国广播电视学刊, 2020(6): 37-42.
WU Y, SUN T. A study of key nodes of event networks in internet information dissemination—an empirical study based on network topology[J]. China Radio & TV Academic Journal, 2020(6): 37-42.
- [12] 王晰巍, 贾若男, 刘婷艳, 等. 新冠肺炎疫情事件信息传播主体特征分析方法及模型: 以“钻石公主号邮轮”事件为例[J]. 情报学报, 2021, 40(3): 245-255.
WANG X W, JIA R N, LIU T Y, et al. Research on the characteristics of the participants of information dissemination during the COVID-19 pandemic—a case study of “diamond princess”[J]. Journal of the China Society for Scientific and Technical Information, 2021, 40(3): 245-255.
- [13] 王晰巍, 张柳, 韦雅楠, 等. 社交网络舆情中意见领袖主题图谱构建及关系路径研究: 基于网络谣言话题的分析[J]. 情报资料工作, 2020(2): 47-55.
WANG X W, ZHANG L, WEI Y N, et al. Research on thematic map construction and relationship path of opinion leaders in social network public opinion: an analysis based on Internet rumor topics[J]. Information and Documentation Services, 2020(2): 47-55.
- [14] 琚春华, 赵凯迪, 鲍福光. 融入紧密度中心性与信用的社交网络用户影响力强度计算模型[J]. 情报学报, 2019, 38(2): 170-177.
JU C H, ZHAO K D, BAO F G. A user influence strength model in E-commerce social networks based on closeness and users' credit[J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(2): 170-177.



- [15] YUAN J J, SHI J W, CHE J Y, et al. Modeling and simulation analysis of public opinion polarization in a dynamic network environment[J]. Concurrency and Computation: Practice and Experience, 2020, 32(19): e5771.
- [16] KEMPE D, KLEINBERG J, TARDOS É. Maximizing the spread of influence through a social network[C]//KDD'03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. [S.l.:s.n.], 2003: 137-146.
- [17] 吴小兰, 章成志. 基于社交媒体的高影响力跨学科用户发现研究[J]. 情报学报, 2017, 36(6): 618-627.
- WU X L, ZHANG C Z. Finding high-impact interdisciplinary users based on social media[J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(6): 618-627.
- [18] 朱毅华, 张超群. 基于影响模型的网络舆情演化与传播仿真研究[J]. 情报杂志, 2015, 34(2): 28-36.
- ZHU Y H, ZHANG C Q. An application of impact model to simulating the evolution and dissemination of network public opinion[J]. Journal of Intelligence, 2015, 34(2): 28-36.
- [19] SHANNON C E. A mathematical theory of communication[J]. The Bell System Technical Journal, 1948, 27(3): 379-423.
- [20] ZHANG H L, JIN R, ZHANG Y, et al. A public psychological pressure index for social networks[J]. IEEE Access, 2020(8): 23457-23469.
- [21] 吴瑞明, 胡代平. 评价系统结构的信息熵方法[J]. 情报杂志, 2008, 27(11): 3-6.
- WU R M, HU D P. The method of evaluating and choosing for systems structures based on entropy[J]. Journal of Information, 2008, 27(11): 3-6.
- [22] RIM H, LEE Y, YOO S. Polarized public opinion responding to corporate social advocacy: social network analysis of boycotters and advocates[J]. Public Relations Review, 2020, 46(2): 101869.
- [23] 宋卓远, 阚乾超, 赵凯, 等. 基于共词分析的国内文本挖掘研究[J]. 图书馆学刊, 2021, 43(4): 104-111.
- SONG Z Y, KAN Q C, ZHAO K, et al. Research on domestic text mining based on CO-word analysis[J]. Journal of Library Science, 2021, 43(4): 104-111.

- [24] DBLP Computer Science Bibliography. DataBase systems and Logic Programming[EB]. 2021.
- [25] American Physical Society. APS data sets for research[EB]. 2021.

[作者简介]



李攀攀(1983-), 男, 博士, 嘉兴学院信息科学与工程学院讲师, 主要研究方向为社会计算、网络空间安全、人工智能等。



谢正霞(1982-), 女, 嘉兴学院建筑工程学院工程师, 主要研究方向为社会计算、网络舆情等。



王赠凯(1980-), 男, 博士, 嘉兴学院信息科学与工程学院讲师, 主要研究方向为社会计算、人工智能等。



靳锐(1976-), 男, 哈尔滨工业大学计算机科学与技术学院博士生, 主要研究方向为信息安全、社交网络分析、机器学习等。