



研究与开发

一种基于信息 DNA 的互联网信息 内容传播及演化追溯方法

李攀攀¹, 谢正霞¹, 王赠凯¹, 靳锐²

(1. 嘉兴学院, 浙江 嘉兴 314001; 2. 哈尔滨工业大学, 黑龙江 哈尔滨 150001)

摘要: 针对如何解决互联网上信息内容传播及演化过程追溯的问题, 提出了一种基于信息 DNA 的互联网信息传播及演化追溯方法。首先, 根据领域知识对互联网信息内容进行语义抽取, 形成信息内容关键特征集; 然后, 使用信息内容关键特征集, 提出基于局部敏感哈希的信息 DNA 构建方法; 最后, 通过公开数据集验证了所提方法的可用性及有效性。以信息 DNA 为核心标识符解决了互联网同源信息传播及演化过程可追溯的问题, 对研究互联网信息内容传播、演化追溯及网络舆情事件的治理与引导等有重要的现实意义。

关键词: 信息 DNA; 信息演化; 信息传播; 计算传播学

中图分类号: G206

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2022280

An information-DNA based method of information dissemination and evolution on Internet

LI Panpan¹, XIE Zhengxia¹, WANG Zengkai¹, JIN Rui²

1. Jiaxing University, Jiaxing 314001, China

2. Harbin Institute of Technology, Harbin 150001, China

Abstract: To solve the problem of how to trace the information content dissemination and evolution process on the Internet, an information DNA-based method of information dissemination and evolution on Internet was proposed. Firstly, semantic extraction of Internet information content was performed based on domain knowledge to form a key feature set of information content. Then, using the key feature set of information content, an information DNA construction method based on locally sensitive hashing was proposed. Finally, the usability and effectiveness of the proposed method were verified by public dataset. The problem of traceability of Internet homologous information dissemination and evolution process was solved by using information DNA as the core identifier, which was of great practical significance for the study of Internet information content dissemination, evolution tracing and the governance and guidance of Internet public opinion events.

Key words: information DNA, information evolution, information dissemination, computational communication

收稿日期: 2022-05-11; 修回日期: 2022-10-20

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), Zhejiang Provincial Natural Science Foundation of China (No.LY18F020021)

0 引言

以互联网为代表的信息技术改变了信息的传播模式,相较于传统媒体,信息传播渠道呈现开放性和虚拟化的典型特点。信息在互联网中的传播速度更快、影响范围更广,传播渠道也更加多样^[1-2]。但是,在开放的互联网中,网络谣言传播、数字知识产权剽窃、信息的不当引用等问题导致虚假和不实信息的蔓延和泛滥^[3],给互联网上信息内容的监管带来严峻挑战,因此,研究面向信息内容的传播及演化溯源具有重要的现实意义^[4]。

在互联网信息传播溯源的研究领域中,传统方法主要使用信息发布时间、发布地址、引用(转载或转述)时间等关键特征构建信息传播链,但是,这种方式通常忽略对信息内容的理解,导致对信息内容的不当引用或剽窃难以被溯源^[5]。开放的互联网中,信息的传播往往不是独立的个体行为^[6],传播环境的开放性、复杂性,特别是信息传播过程中受其他相关信息(如融合、引用、评述等因素)的影响,都给信息内容的传播及演化过程追溯带来困难^[7]。总的来说,在开放的互联网中研究信息的传播和演化过程的追溯面临如下 3 个方面的挑战。

- 互联网上信息的传播渠道多样,信息在复杂多样的传播渠道流转过程中,元数据变化或信息内容发生演化使得同源信息的追溯变得困难。
- 信息在融合传播过程中不仅受多个前置信息内容的影响,还受与这些前置信息之间传播渠道多样性的影响,这都给信息传播及演化过程追溯带来挑战。
- 信息在传播过程中发生“变异”后,传统使用串匹配、距离向量法等同源信息相似度检测方法精确率低,难以适用。

为解决上述问题,本文提出了信息 DNA 的概念,作为信息内容的标识符,信息 DNA 是识别信

息传播及其演化过程的可辨识标识,主要采用关键信息特征抽取以及局部敏感哈希算法的思路构建描述信息内容特征的信息 DNA,使用信息 DNA 有效地解决在开放互联网中的信息内容传播和演化追溯问题,实现了信息在开放互联网上传播过程中内容的可追溯。本文的主要贡献有如下 3 个方面。

- 提出了面向互联网信息内容 DNA 的概念,将其作为信息传播及其演化追溯的标识符。
- 提出了信息在互联网传播过程中面向内容的可追溯方法,实现了对信息直接传播和间接传播的可追溯性。
- 提出了互联网上信息变异传播可追溯的方法,能追溯同源信息的传播路径及其内容的演化,并能在一定程度上容忍信息传播过程中内容的变异。

网络空间作为陆、海、空、天之外的第五空间,泛在网络空间打破了信息传播的时空束缚,信息的传播内容、传播范围、传播方式、传播载体、传播时效、传播效果及传播渠道等均呈现新的特点,研究互联网上信息内容的传播是对其进行风险管控及延伸控制的主要手段之一,具有重要的现实意义,期望本文方法能为在开放互联网中研究信息的传播及演化追溯提供一定的参考和借鉴。

1 面向信息 DNA 的多维特征抽取方法

1.1 互联网信息传播模型

理论上,互联网中信息之间的引用、转述、评论、转发等均会产生传播影响力,信息的传播模式主要有直接传播和间接传播两种方式^[8],互联网上信息的传播模型如图 1 所示。

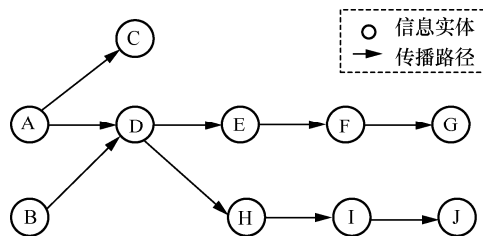


图 1 互联网上信息的传播模型



在图 1 中, 信息实体 A 到信息实体 D 的传播是直接传播, 信息实体 A 到信息实体 G 的传播是间接传播。

信息传播模型用 $G=(V, E)$ 表示, 其中, V 表示信息的集合, $E \subset V \times V$ 表示边的集合, $v_i \in V$ 表示信息实体, $e_{ij} \in E$ 表示信息实体 v_i 与信息实体 v_j 之间的关系, 信息在互联网上的传播以 e_{ij} 为表现形式。

信息变异是指信息实体 v_i 通过 e_{ij} 影响信息实体 v_j 时, 信息实体 v_i 受传播渠道或传播形式的影响出现损失或附加的情况, 这种影响因素包括信息因素、人为因素或网络环境因素。

信息变异率指 v_i 通过 e_{ij} 影响到 v_j 的信息传播过程中, v_i 信息的内容、意义或形式发生变化的程度。

随着信息技术向生产生活领域的全面渗透, 特别是移动互联网及移动社交媒体的普及, 受众获取及传播渠道的多元化和碎片化, 信息在互联网传播过程中更容易发生变异, 如信息形式或内容上的变化, 或未保留全部原信息实体的内容, 如新闻的部分转载。

1.2 信息 DNA 元素多维特征抽取与平行扩展

信息 DNA 的定义: 信息 DNA 是指互联网上所传播信息内容的标识, 是衡量信息传播及演化过程中是否为同源信息的标识符。

设信息实体的特征集记作 D , 其内容记作 D_c 、元数据集记作 D_m , 显然, 信息特征集由其内容和元数据两部分组成, 其中内容是指信息要传递的内容, 元数据是标识信息本身的属性, 如信息长度、信息产生者、信息发布时间等。因此, 则有 $D=D_m \cup D_c$, 那么构建信息 DNA 是从其特征集 D 中抽取关键特征变量, 处理后使其成为信息传播及演化过程的标志。那么, 提取信息特征集也即转化成如何使用语料库描述信息内容及其元数据, 但是, 传统以“词”为最小单位构建语料库的方法中, 很多高频词(如虚词)对信息内容没有实际意义, 且需要较高维度的张量才能完整表

达出信息实体特征集 D 的内容, 这将导致严重的维度灾难^[9], 继而给信息传播路径识别与追溯带来技术上的挑战。

为了解决维度灾难的问题, 本文采用词频—逆向文档频率 (term frequency-inverse document frequency, TF-IDF) 方法, 该方法注重衡量每个“特征词”对信息内容的贡献程度^[10]。通过对信息实体中词语出现的次数进行“全局”归一化处理后, 再使用 TF-IDF 方法能有效地避免维度灾难^[11]。信息实体特征集多维特征的抽取过程如式 (1) 所示。

$$\text{tf} \times \text{idf}(i, j) = \text{tf}_{ij} \times \text{idf}_i = \frac{n_{ij}}{\sum_k n_{kj}} \times \log \left(\frac{|D|}{1 + |D_i|} \right) \quad (1)$$

其中, n_{ij} 表示词 i 在信息 j 中出现的频次, $\sum_k n_{kj}$ 表示信息 j 中每个词出现频次的总和, 即信息 j 中的词数, $|D|$ 为信息集中的信息实体总数, $|D_i|$ 为信息集中出现词 i 的信息实体, 需要说明的是式 (1) 中除以信息实体总词数是为了进行归一化处理。

根据前文所述, 信息 DNA 具备对信息实体内容理解及形式变异后的追踪和识别能力, 因此, 信息 DNA 从信息中抽取的多维特征关键词集包含两个方面。

(1) 信息元数据的标识符, 是信息实体全体属性的集合。

(2) 对信息内容的理解, 即采用自然语言理解技术对信息的内容进行分析处理。

因此, 面向信息 DNA 的多维特征抽取过程示意图如图 2 所示。

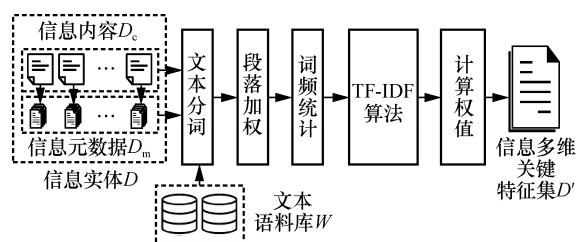


图 2 面向信息 DNA 的多维特征抽取过程示意图

从图 2 可以看出, 经过特征抽取之后, 由于剔除了对信息特征无贡献的特征, $|D'| \ll D$ 。信息 DNA 具备了信息变异后的理解能力, 这为追溯信息演化及变异传播提供了基础依据。

但是, 由于信息在互联网上传播过程的不确定性, 如元数据的缺失、表示方法或格式不一等, 单纯抽取的信息多维特征关键词集仅仅是对当前信息实体的描述。信息在传播过程中可能发生变化, 如时间元数据可能由“2021 年 1 月 3 日”变成“2021/1/3”, 地理元数据可能由“上海”变成“中国上海”或“沪”等。为解决上述问题, 需要对 D' 进行规则化, 即数据抽取转换和加载 (extract-transform-load, ETL), 针对特定的信息实体通过一系列规则模型将 D' 进行规格化处理并将其更新, 信息 ETL 的更新过程如式 (2) 所示。

$$D' = f(D', \text{RULE}) \quad (2)$$

其中, RULE 为规则化集, 由领域专家确定。

开放互联网实现了信息跨地域、跨语种的传播、交流与共享, 如不同语种新闻之间的转载评论、学术论文的跨语种引用等。为了提升本文方法的适用性, 基于信息 DNA 的溯源方法应支持信息实体的跨语种传播。因此, 需使用多语言模型库对规则化的 D' 进行多语言系统扩展, 本文使用平行语料库对信息多维特征关键词集进行平行扩充, 这就实现了对信息内容的平行语料信息增强, 面向信息内容的平行语料信息增强示意图如图 3 所示。

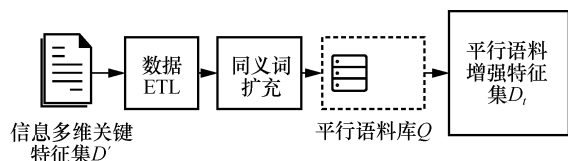


图 3 面向信息内容的平行语料信息增强示意图

在图 3 中, 平行语料库 Q 采用支持多语种语言的平行语料库, 将多维关键特征集 D' 扩充到平行语料增强特征集 D_i , 实现从词、语句乃至篇章级语料中自动抽取平行语句对的功能^[12-13]。在平

行语料的选择上, 采用余弦相似度计算 D' 与文本语料库 Q 的相似性, 余弦相似度通过两个向量之间夹角的余弦值评估语料之间的相似度, 两个向量的夹角越小, 余弦值越接近于 1, 语料相似度也就越高^[14]。另一方面, 为了解决 D' 与 Q 向量长度不一致的问题, 采用传统的向量填充法^[15], 使得经过填充后的 D' 或 Q 长度相同, 填充的长度为 $|D' - Q|$, 那么相似度计算方法如式 (3) 所示。

$$\cos \theta = \frac{\sum d'_i q_i}{\sqrt{\sum (d'_i)^2} \times \sqrt{\sum (q_i)^2}} \quad (3)$$

根据式 (3), 若 $\cos \theta > k$, 则将平行语料库 Q 添加到平行语料集 Q' , 因此, 平行语料增强关键词集 $D_i = D' \cup Q'$, 且有 $|D_i| > |D'|$, 且从语义层面对信息 D 进行了扩充, 具有更强的泛化表现能力, 能有效应对并解决信息在互联网传播过程中因元数据的变化、缺失以及信息内容“变异”等所导致的难以追溯的问题。

2 基于信息 DNA 的信息内容传播追溯方法

信息在互联网传播的过程中, 信息实体元数据及其内容的变化具有随机性和不确定性, 因此, 信息 DNA 要能容忍并追溯信息传播的变化, 本文引入了局部敏感哈希的方法, 构建信息内容传播及演化过程中的“遗传物质”, 即信息 DNA。

2.1 信息传播“遗传物质”与局部敏感哈希算法

构建信息 DNA 的思路是, 使用信息的平行语料增强关键词集 D_i 构建出描述信息内容的唯一标志, 信息 DNA 要能容忍信息实体的元数据及其内容在一定范围内的变异, 是追溯互联网上信息传播的线索。传统的哈希算法不具备这种“容忍”信息内容或形式变化的能力, 信息元数据或内容在形式上的任何微小变化都会导致其哈希值的巨大变化^[16-17]。因此, 无法将信息元数据或内容的哈希值作为其传播过程中的“遗传物质”用于追溯传播及演化路径。

为了应对上述问题, 本文引入局部敏感哈希



(locality sensitive hashing, LSH) 方法, LSH 方法常用于近似最近邻查询, 不仅在高维空间中有优异的性能表现, 而且在克服维度灾难的同时, 还能保持可接受的时间和空间复杂度, 在图形图像、音视频、海量文本等领域的相似性查询算法中有广泛的应用^[18]。根据 LSH 方法的思想, 在原空间中很近 (相似) 的两个点, 经过 LSH 方法中哈希函数的映射后, 哈希值有很大的概率是相同的, 而两个距离很远的点 (相似性弱) 映射后, 哈希值相等的概率很小, 即 LSH 方法的哈希函数 $f_{\text{ls}}h$ 满足如下性质。

对于在 $f_{\text{ls}}h$ 高维空间的任意两点 x 和 y , 则满足:

(1) 如果 $d(x, y) \leq R$, 则 $p(h(x)=h(y)) \geq P_1$;

(2) 如果 $d(x, y) \geq cR$, 则 $p(h(x)=h(y)) \leq P_2$ 。

其中, $c > 1$, $P_1 > P_2$ 。

基于 LSH 方法的数据距离与冲突概率示意图如图 4 所示。

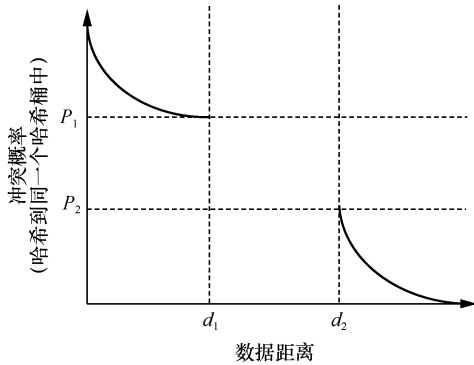


图 4 基于 LSH 方法的数据距离与冲突概率示意图

通过上述对 LSH 方法的分析, 根据一定的策略或方法从信息实体中选择适当的参数, 再采用 LSH 方法构建信息内容传播的“遗传物质”, 能在一定程度范围内容忍信息元数据或内容的变化。信息实体 A 和信息实体 B 的信息 DNA 分别表示为 DNA_A 和 DNA_B , 在信息 A 向信息 B 的传播过程中, 当相似度发生变化时, $d(\text{DNA}_A, \text{DNA}_B) < R$, 则 $h(\text{DNA}_A)$ 等于 $h(\text{DNA}_B)$ 的概率大于 P_1 , 这样可以将 P_1 作为判断同源信息的阈值。

2.2 信息实体语义抽取

根据前文讨论, 信息 DNA 要同时具备追溯信息传播过程中元数据和及其内容双重变化的能力, 因此, 信息 DNA 还要包含信息实体的语义, 需要提取信息实体的语义信息。

信息实体语义的理解通常有分布式语义表示、模型论语义表示和框架语义表示 3 种基本方法^[19-20], 前两种方法通常将信息实体的词或句子用高维向量表示, 这导致在技术上信息特征集提取变得困难, 特别是互联网上信息传播的碎片化, 使得上述前两种方法难以适用于碎片化的互联网信息实体语义理解。由于采用分层的思路, 框架语义方法对信息实体的语义碎片化有较强的表示能力, 因此, 本文使用框架语义表示方法分析信息实体的语义, 把整个信息实体分成 3 个层次: 信息领域 (domain)、信息意图 (intent) 和语义槽 (slot), 其中, 信息领域指信息所处的上下文环境, 信息意图指信息实体隐式表达的潜在内容, 语义槽指使用预定义关键词集合增强信息语法的扩展能力。针对信息实体的具体领域, 可由领域专家将信息实体划分成其他表示形式, 通过框架语义表示方法抽取信息实体的语义信息用 G 表示。

2.3 基于 LSH 的信息 DNA 构建方法

设信息实体 $D = \{d_i\} (i \in |D|)$, 信息实体 D 经过 LSH 方法计算后得到其对应的信息 DNA, 信息实体 D 的 DNA 信息 D_{DNA} 定义为式 (4)。

$$D_{\text{DNA}} := \{D_{\text{ID}}, G_{\text{ID}}, D_{i\text{DNA}}\} \quad (4)$$

其中, D_{ID} 表示信息实体的标识符, G_{ID} 表示信息 ID 的语义信息, $D_{i\text{DNA}}$ 表示信息集中每个信息的信息 DNA。根据前文信息 DNA 的定义, 信息 DNA 是信息内容演化传播过程中的“遗传物质”, 是研究信息传播路径及其内容演化规律的重要指标, 因此信息 DNA 构建的流程及其所选取的参数至关重要。信息 DNA 必须具有信息元数据及其内容的强关联性, 这是评估信息内容传播的基础, 显然这与信息元数据及其内容的表达

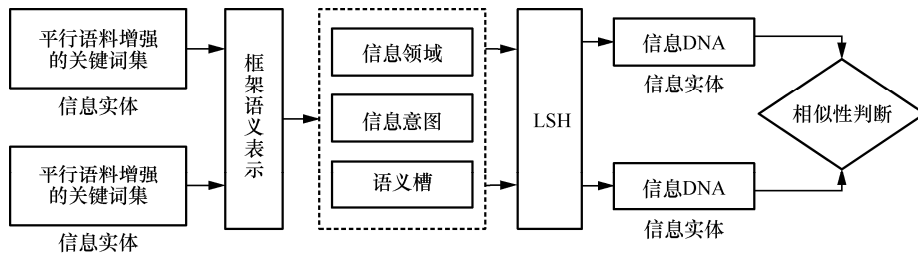


图5 基于 LSH 方法的信息 DNA 计算流程

形式弱关联, 基于 LSH 方法的信息 DNA 计算流程如图 5 所示。

此外, 为了加深对信息内容的理解, 在采用框架语义表示各信息实体 D_i 之后, 再分析出所属领域, 用于领域专家给出信息意图和语义槽, 这就在语义层面实现了对 D_i 的扩充, 之后再使用 LSH 方法计算出每个信息实体的信息 DNA, 具体来讲, 信息 DNA 的构建算法如下。

算法 1 信息 DNA 构建算法

输入: 信息实体 D 、ID 标志 D_{ID} 及其内容 D_c 和元数据集 D_m , 文本语料库 W , $Q' = \text{NULL}$, 平行语料库 Q , 相相似度判断阈值 k ;

输出: 信息 DNA $\{D_{ID}, D_{DNA}\}$;

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

 使用 W 对 $D_c \cup D_m$ 进行文本分词;

 段落加权;

 词频统计;

 使用 DF-IDF 方法计算权值;

 输出信息实体的关键多维特征集 D' ;

 输出 D' 的语义信息集 G ;

$\}$

for $i=0, i < |D'|, i++\{$

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

 计算 $D'[i]$ 和 $Q[j]$ 的相似度 k' ;

 if ($k' \geq k$)

$Q' = Q' \cup Q_j$;

$\}$

$\}$

$D_i = D' \cup Q'$

使用 LSH 方法计算 D_i , 得到信息实体 D 的 DNA 信息 D_{DNA} ;

return $\{D_{ID}, D_{DNA}\}$

综合前文所述本文采用 LSH 方法构建信息 DNA, 能有效地解决信息在互联网传播过程中元数据缺失、差异化表述、内容剽窃、不恰当引用等情况下的追溯问题。

2.4 基于信息 DNA 的信息传播链构建及溯源方法

此外, 除构建信息 DNA 外, 构建信息 DNA “遗传物质”的传播链, 用于信息传播内容及演化追溯, 需要两个主要阶段, 即建立信息 DNA 传播链阶段、信息 DNA 验证阶段。

(1) 建立信息 DNA 传播链阶段

步骤 1 提取信息实体的平行语料增强关键词集 D_i 。

步骤 2 使用 LSH 方法计算信息实体的平行语料增强关键词集, 对 D_i 中每个关键词进行投影映射, 并将映射的桶号作为该信息实体对象的编号来建立哈希索引表, 并将哈希索引表向量存储到对应的哈希桶中。

(2) 信息 DNA 验证阶段

在信息内容传播及演化追溯研究领域, 传统方法采用信息内容相似性分析与度量为主要手段, 本文使用的信息 DNA 突破了这种相对“刚硬”的思路。通过引入 LSH 方法, 本文有效地解决了信息在互联网传播过程中各种不确定性的问题, 并能支持信息内容的传播演化追溯。对于信息实体 A 和 B, 若存在 e_{AB} , 那么同源信息的判断方法如下。



- 若 $d(A_{\text{DNA}}, B_{\text{DNA}}) \leq k_{\text{DNA}}$, 则可判断信息实体 A、B 为同源信息, 即信息实体 B 受信息实体 A 的影响。
- 若 $d(A_{\text{DNA}}, B_{\text{DNA}}) > k_{\text{DNA}}$, 则可判断信息实体 A、B 不是同源信息, 即信息实体 B 不受信息实体 A 的影响。

其中, k_{DNA} 为同源信息的判断阈值, 其具体值由领域专家根据不同的应用环境、应用场合等综合分析后指定。

3 信息传播及演化过程中可追溯性理论证明

信息在互联网传播过程中, 其内容被转述、转载等, 信息传播主要面临如下 3 个风险。

- (1) 信息元数据丢失或形式发生变化。
- (2) 内容被不当引用、评述或转载乃至歪曲。
- (3) 信息杂交融合传播后引起的追溯困难。

下面从理论上证明信息 DNA 用于信息传播及演化过程中可追溯的有效性。

(1) 对于同源信息在传播过程中形式的变化, 如元数据形式的变化 (丢失、格式变化等) 以及信息内容的变化, 如何验证信息实体 A 和信息实体 B 为同源信息?

证明 1: 信息实体 A 到信息实体 B 的变异传播过程中, 在信息元数据或内容变化后, 因为有平行语料库的作用, 信息实体 B 在信息实体 A 基础上进行扩充, 会使得 $|A| < |B|$, 根据 LSH 方法的性质, 可判定 $d(A_{\text{DNA}}, B_{\text{DNA}}) < k_{\text{DNA}}$, 这样可认定传播链上的信息内容具有一定的相似性, 证毕。

(2) 信息传播过程中, 若信息实体 A 被信息实体 B 全文转载或引用, 通过本文方法可分析出信息实体 B 受信息实体 A 的影响。

证明 2: 根据前文描述, 信息实体 A 的扩展集真包含于信息实体 B 的扩展集, 即 $A_{\text{DNA}} \subseteq B_{\text{DNA}}$, 根据 LSH 方法的性质, 则有 $d(A_{\text{DNA}}, B_{\text{DNA}}) = 0$, 根据同源信息遗传物质验证方法, 可判断 A 和 B 为同源信息, 证毕。

(3) 对于信息杂交变异后传播的可追溯问题, 设信息实体 A 和信息实体 B 相融合后共同影响信息实体 C, 通过本文方法可通过信息实体 C 的信息 DNA 判断其受信息实体 A 或信息实体 B 的影响。

证明 3: 显然由证明 2, 可以确定信息实体 C 受影响于信息实体 A, 则有 $d(A_{\text{DNA}}, C_{\text{DNA}}) < k_{\text{DNA}}$, 同理, 也可以确定信息实体 C 受影响于信息实体 B, $d(B_{\text{DNA}}, C_{\text{DNA}}) < k_{\text{DNA}}$, 显然, 对于融合传播后影响的信息实体 C, 受影响于信息实体 A 和信息实体 B, 证毕。

因此, 通过上述证明, 从理论上验证了本文方法具有追溯互联网中信息内容传播及演化的能力。

4 实验分析

下面进一步采用公开数据集及仿真实验验证本文方法的有效性和可用性。

4.1 实验数据

使用 MATLAB 仿真本文方法, 实验数据采用来自互联网的公开数据集 SogouT 互联网信息语料库^[21], 其包含互联网原始网页、引用和评论等, 信息实体 A、B 和 C 随机选自 SogouT 的网页信息, 平行语料库使用联合国平行语料库^[22], 同源信息的判断阈值 $k_{\text{DNA}}=0.6$ 。

4.2 信息直接演化传播的可追溯性

网络环境的开放性 & 传播路径的复杂性导致信息在网络上的变异传播, 首先验证信息传播过程中不同变异程度下的可追溯性, 选取 100 组信息 A_i 影响信息 B_i 的传播过程, 其中, $|A_i| \geq 300$, $|B_i| \leq 400$, $0 < i \leq 100$, 为了更加体现实验结果的有效性, 将本文方法与使用关键词方法对信息传播内容追溯准确性进行对比分析, 关键词方法即通过传统的对信息传播过程中信息主体和受影响信息客体的关键词做余弦相似度分析, 本实验中关键词余弦相似度 $\alpha=0.6$ 。

在信息 A 到信息 B 的传播过程中, 信息的元数据和内容上的变异范围为 10%~70% 的情况

下, 量化评估本文方法及关键词法的可追溯性, 即通过信息实体 B 的 DNA 信息分析出其受信息实体 A 影响的概率, 信息直接演化传播过程中追溯的准确性如图 6 所示。

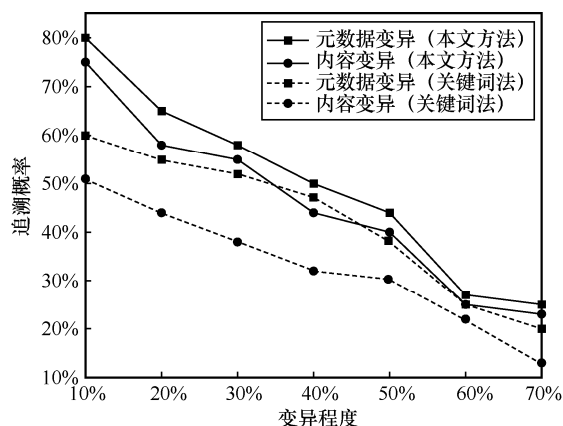


图 6 信息直接演化传播过程中追溯的准确性

从图 6 可以看出, 随着信息变异程度的增大, 本文方法追溯的准确性逐渐降低, 但是当变异程度增大到一定程度时, 追溯的准确性保持在相对稳定的区间, 这验证了本文方法的可用性。较比关键词方法, 本文方法对信息内容可追溯性识别率更高, 主要原因是本文方法使用了同义词语料库和平行语料库对信息实体的特征集进行了扩充, 使得本文方法能够在一定程度上容忍信息内容的变异化传播。另外, 还可以看出, 相较于信息内容变异传播, 本文方法对元数据变异传播的容忍程度更大, 主要是元数据变异后对其进行同义词扩充相对容易, 扩充的程度更加全面, 这也验证了使用平行语料增强关键词集能提升对原始信息实体内容的表达能力。

下面进一步分析本文方法追溯的误报率, 即假阳性和假阴性, 其中假阳性指当信息实体 A 未影响信息 B 时, 而本文方法推断出信息实体 B 受信息实体 A 影响的概率, 假阴性则指当信息实体 A 影响信息实体 B 时, 而本文方法未能推断出信息实体 B 受信息实体 A 影响的概率, 在元数据和内容传播过程中不同变异程度的情况下, 信息直接演化传播过程中追溯的误报率如图 7 所示。

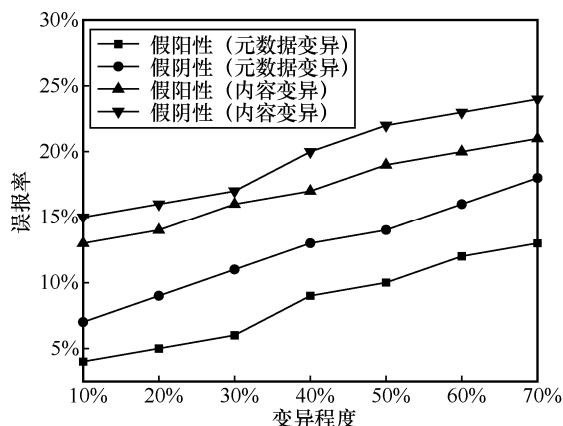


图 7 信息直接演化传播过程中追溯的误报率

从图 7 中可以看出, 本文方法的误报率与信息的变异程度密切相关, 随着变异程度的增大, 误报率呈现上升的趋势, 同时也可以看出, 相较于元数据的变异, 内容变异所带来的误报率更高, 这与图 6 的结论相吻合, 验证了本文方法对信息传播中内容变异的追溯能力稍逊于对元数据变异的追溯能力。

4.3 信息间接演化传播的可追溯性

进一步地, 通过实验分析间接传播内容的可追溯性, 即对传播链上信息实体 A, 经过信息实体 B 影响信息实体 C, 其中 $|C|=530$, 即信息实体 C 的个数为 530, 验证本文方法检测到信息实体 C 受影响于信息实体 A 的概率, 为了充分验证本文方法的有效性, 在信息实体 A 到信息实体 B 的传播过程中分别有 10%、30% 和 50% 的信息变异情况下, 信息 B 到信息 C 传播变异为 10%~70% 的情况下, 通过信息实体 C 的信息 DNA 分析出其受信息实体 A 影响的概率, 信息间接演化传播过程中追溯的准确性如图 8 所示。

信息在传播链上的变异程度, 直接影响着信息追溯的准确性, 对比图 7 中的实验结果, 验证了本文方法的有效性, 但是随着信息变异程度的增加, 追溯准确性受到一定程度的挑战, 主要由于实验中采用了相对单薄的同义词语料扩展库, 影响了对信息内容的扩展, 继而给信息内容变异传播的分析带来消极影响。在实际环境中, 随着同义词语料库的丰富, 本文方法在信息间接传播的

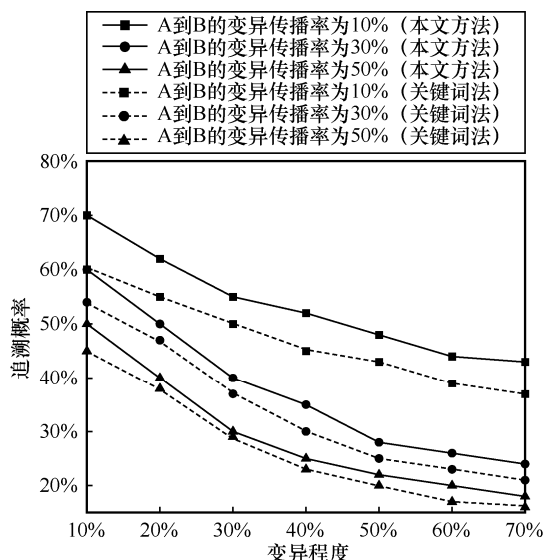


图8 信息间接演化传播过程中追溯的准确性

追溯准确性上会随之提升。通过对图8进一步分析还可以发现,在不同变异传播的情况下,本文方法的追溯准确性高于关键词法,其原因也与图7的分析结论类似。

下面进一步分析在不同信息变异程度下间接传播过程中追溯的误报率,信息间接演化传播追溯的误报率如图9所示。

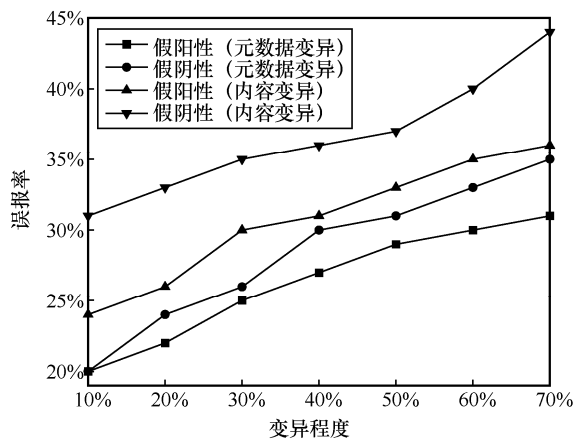


图9 信息间接演化传播追溯的误报率

从图9中可以看出,在元数据变异和内容变异范围内,误报率均在可接受的范围内。进一步地,结合图7和图9,可以看出本文方法对直接演化传播和间接演化传播在一定变异范围内追溯的有效性。

4.4 信息融合传播的可追溯性

相较于信息在某一传播链上的影响,信息融合后对受影响信息的追溯性判断更具有挑战性。假设信息实体A和信息实体B相融合后共同影响信息实体C,验证信息实体C与信息实体A和信息实体B均为同源信息,其中,融合率指信息实体A和信息实体B分别到信息实体C传播变异率(或变异程度)的最小值。在实验过程中,信息实体A和信息实体B融合率分别为10%、30%、50%以及融合后信息在对信息实体C的传播过程中变异程度为10%~70%的情况下,对信息融合传播过程中进行可追溯性分析,即通过信息DNA检测出信息实体C受信息实体A和信息实体B共同影响的概率,信息融合演化传播过程中追溯的准确性如图10所示。

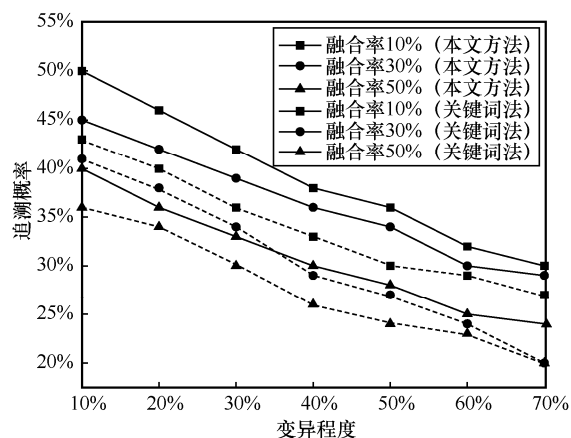


图10 信息融合演化传播过程中追溯的准确性

信息融合传播受限于前序信息实体对后续信息的影响程度,进一步地,从图10可以看出,本文方法在低变异融合传播过程中,能以较大概率识别得到同源信息。与前文类似,同源信息识别的精度与信息传播过程的变异程度相关,识别的精度随着变异程度的增大而降低,但当变异程度大于30%后,识别精度的下降趋势变得相对平缓,这样验证了当信息变异足够大时,本文方法仍能以一定的概率验证同源信息。与图6和图8的实验结果类似,在不同的融合率传播情况下,本文

方法比关键词法的追溯精度要高,其原因也与图 6 的分析类似。

下面进一步分析信息融合传播过程下追溯的误报率,信息融合演化传播过程中追溯的误报率如图 11 所示。

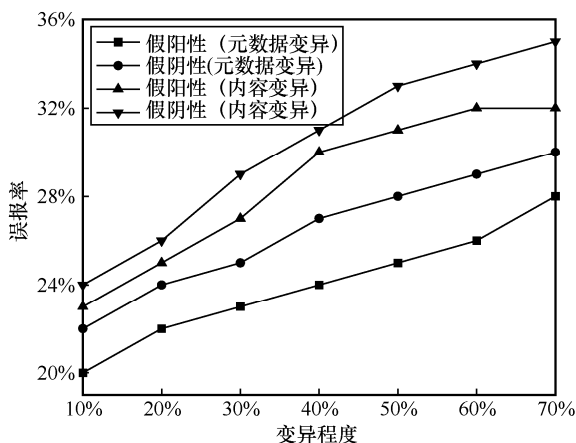


图 11 信息融合演化传播过程中追溯的误报率

从图 11 中可以看出,信息融合传播情况下追溯的误报率保持在 10%~35%。需要指出的是,随着变异程度的增加,误报率也始终保持了相对缓和增长的趋势,这也验证了本文方法在信息融合传播环境下的可用性。进一步地,结合图 7、图 9 和图 11 可以看出,各种信息传播模式下,本文方法均能在一定范围内实现信息内容传播的可追溯性,即验证了本文方法的可用性和有效性。

5 结束语

当前,互联网以其开放性、时效性、共享性等特点改变了信息知晓、交流与共享的模式,网络空间承载了海量的信息,研究互联网上信息内容传播及演化过程的可追溯性在网络舆情治理、数字内容版权管理等领域具有重要的现实意义。针对互联网上信息内容的传播及演化过程,本文提出了信息 DNA 的概念,在对信息内容多维特征抽取和语料平行扩展的基础上,引入了基于 LSH 的信息 DNA 构建方法,解决了互联网上信息内容传播及其演化过程可追溯的重要

问题。期待本文的研究思路为本领域的研究人员提供一定的参考,共同促进本领域的发展。

参考文献:

- [1] 曹玖新,高庆清,夏蓉清,等. 社交网络信息传播预测与特定信息抑制[J]. 计算机研究与发展, 2021, 58(7): 1490-1503.
CAO J X, GAO Q Q, XIA R Q, et al. Information propagation prediction and specific information suppression in social networks[J]. Journal of Computer Research and Development, 2021, 58(7): 1490-1503.
- [2] 李晋,杨子龙. 微博转发网络中的节点特征和传播模型[J]. 电信科学, 2016, 32(1): 40-45.
LI J, YANG Z L. Node characteristic and propagation model in microblog forwarding network[J]. Telecommunications Science, 2016, 32(1): 40-45.
- [3] 徐铭达,张子柯,许小可. 基于模体度的社交网络虚假信息传播机制研究[J]. 计算机研究与发展, 2021, 58(7): 1425-1435.
XU M D, ZHANG Z K, XU X K. Research on spreading mechanism of false information in social networks by motif degree[J]. Journal of Computer Research and Development, 2021, 58(7): 1425-1435.
- [4] 刘小洋,何道兵. 基于突发公共事件的信息传播动力学模型与舆情演化研究[J]. 计算机科学, 2019, 46(5): 320-326.
LIU X Y, HE D B. Study on information propagation dynamics model and opinion evolution based on public emergencies[J]. Computer Science, 2019, 46(5): 320-326.
- [5] 曹峰,张真继,关晓兰. 基于系统动力学的网络舆情驱动力模型研究[J]. 电信科学, 2020, 36(12): 49-58.
CAO F, ZHANG Z J, GUAN X L. Research on the driving force model of Internet public opinion based on system dynamics[J]. Telecommunications Science, 2020, 36(12): 49-58.
- [6] ZHANG H J, DONG Y C, XIAO J, et al. Consensus and opinion evolution-based failure mode and effect analysis approach for reliability management in social network and uncertainty contexts[J]. Reliability Engineering & System Safety, 2021(208): 107425.
- [7] FANI H, JIANG E, BAGHERI E, et al. User community detection via embedding of social network structure and temporal content[J]. Information Processing & Management, 2020, 57(2): 102056.
- [8] 李攀攀,谢正霞,王赠凯,等. 开放互联网环境基于信息熵的信息传播影响力计算方法[J]. 电信科学, 2022, 38(4): 90-100.
LI P P, XIE Z X, WANG Z K, et al. Calculation method of information dissemination based on information entropy in public Internet[J]. Telecommunications Science, 2022, 38(4): 90-100.
- [9] 王乃钰,叶育鑫,刘露,等. 基于深度学习的语言模型研究进展[J]. 软件学报, 2021, 32(4): 1082-1115.
WANG N Y, YE Y X, LIU L, et al. Language models based on



- deep learning: a review[J]. Journal of Software, 2021, 32(4): 1082-1115.
- [10] 丁晓阳, 王兰成. 网络论坛文本特征词权重计算优化方法研究[J]. 情报理论与实践, 2021, 44(5): 187-192.
DING X Y, WANG L C. Research on optimized calculation method for weight of terms in BBS text[J]. Information Studies: Theory & Application, 2021, 44(5): 187-192.
- [11] 孟青, 刘波, 张恒远, 等. 在线社交网络中群体影响力的建模与分析[J]. 计算机学报, 2021, 44(6): 1064-1079.
MENG Q, LIU B, ZHANG H Y, et al. Multi-relational group influence modeling and analysis in online social networks[J]. Chinese Journal of Computers, 2021, 44(6): 1064-1079.
- [12] 贾承勋, 赖华, 余正涛, 等. 融合单语语言模型的汉越伪平行语料生成[J]. 计算机应用, 2021, 41(6): 1652-1658.
JIA C X, LAI H, YU Z T, et al. Chinese-Vietnamese pseudo-parallel corpus generation based on monolingual language model[J]. Journal of Computer Applications, 2021, 41(6): 1652-1658.
- [13] 黄水清, 王东波. 国内语料库研究综述[J]. 信息资源管理学报, 2021, 11(3): 4-17, 87.
HUANG S Q, WANG D B. Review of corpus research in China[J]. Journal of Information Resources Management, 2021, 11(3): 4-17, 87.
- [14] BAÑÓN M, CHEN P Z, HADDOW B, et al. ParaCrawl: web-scale acquisition of parallel corpora[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 4555-4567.
- [15] 孙留倩, 魏玉良, 王佰玲. 基于图卷积网络的多源本体相似度计算方法[J]. 网络与信息安全学报, 2021, 7(5): 149-155.
SUN L Q, WEI Y L, WANG B L. Novel similarity calculation method of multisource ontology based on graph convolution network[J]. Chinese Journal of Network and Information Security, 2021, 7(5): 149-155.
- [16] 郭一村, 陈华辉. 在线哈希算法研究综述[J]. 计算机应用, 2021, 41(4): 1106-1112.
GUO Y C, CHEN H H. Survey on online hashing algorithm[J]. Journal of Computer Applications, 2021, 41(4): 1106-1112.
- [17] WANG J, ZHANG T, SONG J, et al. A survey on learning to hash[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 769-790.
- [18] JAFARI O, MAURYA P, NAGARKAR P, et al. A survey on locality sensitive hashing algorithms and their applications[J]. arXiv preprint, 2021, arXiv: 2102.08942.
- [19] 徐戈, 杨晓燕, 汪涛. 单词语义相似性计算综述[J]. 计算机工程与应用, 2020, 56(4): 9-15.
XU G, YANG X Y, WANG T. Survey on semantic similarity calculation of words[J]. Computer Engineering and Applications, 2020, 56(4): 9-15.
- [20] 由丽萍, 刘荟, 刘焘. 基于框架的情感语义表示模型设计与标注实验[J]. 情报科学, 2014, 32(6): 143-147.
YOU L P, LIU H, LIU T. Frame-based sentiment semantic representation model design and annotating experiments[J]. Information Science, 2014, 32(6): 143-147.
- [21] LUO C, ZHANG Y K, LIU Y Q, et al. SogouT-16: a new web corpus to embrace IR research[C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2017: 1233-1236.
- [22] 联合国大会和会议管理部. 联合国平行语料库[EB]. 2021. Department for General Assembly and Conference Management of United Nations. A six-language Parallel Corpus [EB]. 2021.

[作者简介]



李攀攀 (1983-), 男, 博士, 嘉兴学院讲师, 主要研究方向为社会计算、开源情报、网络空间安全等。



谢正霞 (1982-), 女, 嘉兴学院工程师, 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。