软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(11):2855−2869 [doi: 10.13328/j.cnki.jos.004932]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

中文公众事件信息熵计算方法*

靳 锐, 张宏莉, 张 玥, 王 星

(哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

通讯作者: 靳锐, E-mail: jinrui@pact518.hit.edu.cn, http://www.hit.edu.cn

摘 要: 随着中文社交网络的发展(特别是微博的兴起),互联网中文公众事件越来越深刻地影响现实社会的生产和生活.由于缺乏有效的技术手段,信息处理的效率受到了限制.提出了一种公众事件信息熵的计算方法,其基本思想是:首先,对公众事件信息内容进行建模;然后,以香农信息论为理论基础,对公众事件的多维随机变量信息熵进行计算.这为互联网公众事件的定量化分析提供了一个重要的技术指标,为进一步的研究工作打下基础.

关键词: 社会计算;公众事件;香农信息论;信息熵;最大熵理论

中图法分类号: TP393

中文引用格式: 靳锐,张宏莉,张玥,王星.中文公众事件信息熵计算方法.软件学报,2016,27(11):2855−2869. http://www.jos.org.cn/1000-9825/4932.htm

英文引用格式: Jin R, Zhang HL, Zhang Y, Wang X. Calculation method of Chinese public event information entropy. Ruan Jian Xue Bao/Journal of Software, 2016,27(11):2855−2869 (in Chinese). http://www.jos.org.cn/1000-9825/4932.htm

Calculation Method of Chinese Public Event Information Entropy

JIN Rui, ZHANG Hong-Li, ZHANG Yue, WANG Xing

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: With the proliferation of the Chinese social network (especially the rise of weibo), the productivity and lifestyle of the country's society is more and more profoundly influenced by the Chinese internet public events. Due to the lack of the effective technical means, the efficiency of information processing is limited. This paper proposes a public event information entropy calculation method. First, a mathematical modeling of event information content is built. Then, multidimensional random variable information entropy of the public events is calculated based on Shannon information theory. Furthermore, a new technical index of quantitative analysis to the internet public events is put forward, laying out a foundation for further research work.

Key words: social computing; public event; Shannon information theory; information entropy; principle of maximum entropy

随着互联网技术的发展,Web 2.0 的网络用户信息发布技术引发了社交网络的蓬勃发展,社交网络时代已经到来.Web 2.0 则更注重用户的交互作用,用户既是网站内容的浏览者,也是网站内容的制造者.基于此,在国际互联网产业领域,以 facebook 和 twitter 为代表的新型社交网站成为了社交网络时代成功的典范,以人人网、新浪微博为代表的中文社交网络取得巨大成功,社交网络深入到社会各个角落,深刻地影响着国家的政治、经济、文化、社会活动组织等领域.

技术革命造成了社会生产生活方式的变革,在社交网络的快速信息交互中,非洲大陆与阿拉伯世界经历了一系列的剧烈社会变革[1];在中国,SNS、微博等社交网站的发展如火如荼,各种社会信息在社交网络中快速流

* 基金项目: 国家重点基础研究发展计划(973)(2013CB329602); 国家自然科学基金(61202457, 61472108, 61402149)

Foundation item: National Program on Key Basic Research Project of China (973) (2013CB329602); National Natural Science Foundation of China (61202457, 61472108, 61402149)

收稿时间: 2015-02-17; 修改时间: 2015-05-08, 2015-09-10; 采用时间: 2015-10-17; jos 在线出版时间: 2015-11-18

CNKI 网络优先出版: 2015-11-18 14:58:46, http://www.cnki.net/kcms/detail/11.2560.TP.20151118.1458.001.html

转,互联网公众意见[2−4]得到了快速的表达与形成,担当起前所未有的社会角色,发挥着举足轻重的社会作用.近两年,中国国内的各类公众事件频频爆发,对互联网舆情监控提出了新的要求.如何准确、快速地获取和分析相应的事件信息,成为中文社交网络信息处理领域的一个新的挑战.

互联网公众意见研究,又称为舆情分析,是当前互联网智能信息处理的研究热点之一[5−7].这项技术研究可应用于国家政策的实施预测、政治选举结果的预测与分析[5]、产品的市场销售分析以及个人名誉与发展等.近年来,该领域开始受到国内外研究人员的重视,也逐渐受到各个国家政府、经济实体乃至个人用户的重视.

如何衡量一件公众事件的重要性、计算其影响力或涉及事件部门的事态严重程度,目前还没有一个有效的衡量方法,无法对公众意见事件的内容信息进行定量化计算.仅仅依靠网民的参与程度来衡量事件的重要程度是不够的,这不但不能反映事件的实质内容,而且具有明显的滞后性,存在易被误导的弊端.

随着社交网络的发展,社会计算[7,8]逐渐引起相关的研究人员的重视.2007 年底,在哈佛大学举办了计算社会学研讨会;2008 年 4 月,美国军方在亚利桑那州立大学举办了社会计算、行为建模和预测研讨会.在此基础上,2009 年,Lazer 等人[9]在《Science》上提出了计算社会学的概念,并指出,网络上的大量信息,如博客、论坛、聊天、消费记录、电子邮件等,都是对现实社会的人及组织行为的映射,网络数据可用来分析个人和群体的行为模式,标志着计算科学和社会科学的交叉融合正成为国际瞩目的前沿研究和应用热点[10].

2003 年,美国提出情报与安全信息学的概念,其核心是研究如何开发、研究智能算法通过数据信息处理技术、安全策略的集成等,使情报采集和安全分析更加系统化、科学化,保障国际安全、国家安全、社会安全、商业安全和个人安全.美国亚利桑那大学基于国家社会安全问题考虑,进行“情报与安全信息学(ISI)”[11]研究;卡内基梅隆大学也开展了公共卫生事件等领域的学术研讨.自 2005 年起,中国科学院自动化研究所开始了情报与安全信息学(ISI)的研究,以社会计算理论与计算实验平台为基础,并以开源情报的获取和处理为基础,对社会媒体和舆情信息进行实时监测、分析和预警[10,12−14].

当前,社会计算方法多用于社区发现与社会媒体挖掘,如社交网络用户的信息交互关系计算、社区与意见领袖发现、社交网络用户行为分析等[5−7].

公众意见分析领域的研究仍然处于发展初期阶段,理论体系还没有完全建立起来,尤其是定量化的技术衡量指标还不完备,引入社会计算方法是解决此问题的有效途径之一.

互联网公众事件的文本形式是互联网信息的重要载体[5],其包含的信息量是事件信息的重要技术指标,也是分析其影响力、舆论压力等技术指标的定量化前提.本文通过香农信息论与最大熵理论的方法,对互联网公众事件内容信息量的计算方法进行了研究,该方法属于社会计算范畴.

1 公众事件数学模型

1.1 公众事件的分析模型

网络文本事件的结构如图 1 所示.

为了进行信息量的计算,先分析一下公众事件的构成,如图 1 所示.这里对事件所包含的信息内容进行分析,事件信息有 5 个构成要素:事件主体、时间、地点、数量、未抽取信息.而事件主体又有 4 个属性:社会(自然)角色、社会(自然)关系、所属机构或体系、主体行为.事件本身具有一个重要属性,即事件社会(自然)类别.

- 数学描述

设事件信息为全集 U,由 n 个子集构成为 $U_1,U_2,\ldots,U_i,\ldots,U_n$,其中,$U_i=\overline{U_1\cup U_2\cup\ldots\cup U_{i-1}\cup U_{i+1}\cup\ldots\cup U_{n-1}}$,图 1 中,$n$ 取 10,U 表示事件主体集合,U_2 表示社会(自然)角色集合,U_3 表示社会(自然)关系,U_4 表示所属机构或体系集合,U_5 表示时间信息集合,U_6 表示主体行为集合,U_7 表示地点集合,U_8 表示数量集合,U_9 表示事件社会(自然)类别集合,U_{10} 表示未抽取信息.由集合的性质可知,$U=\overline{U_1\cup U_2\cup\ldots\cup U_n}$.

我们看到如图 1 所示的 5 个构成要素、5 个相关的属性,这是对公众事件最简化的一种表达方式,公众事件文本内容的信息全部包含在其中.

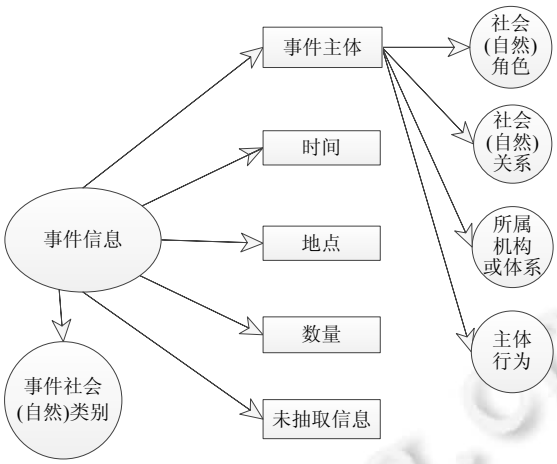

Fig.1 Public event structure

图 1 公众事件结构

经过分析,信息系统内的各个属性和要素之间的相互影响可以导致要素或属性的条件信息量的变化.

图 1 中的信息模块结构来源于文本信息抽取项的研究[15],抽取项有主体、时间、关系、机构等多项研究.

设一个互联网公众事件由 n 个随机变量构成,则事件可以表示为$(X_1,\ldots,X_n)$.事件本身为一个随机信息系统,用X表示,则X等价于$(X_1,\ldots,X_n)$,其联合概率分布为$(p_1,p_2,\ldots,p_i,\ldots,p_n)$.各个分变量之间的函数关系未知,或社会性函数关系极其复杂,无法使用特定的函数关系进行表述,满足 $p_i(X_i|X_k)\neq p_i$,且$(i\neq k)$,即每个分变量之间不存在条件独立关系.

根据哲学的一般原理:在一个系统之内,每一个部分都不是孤立存在的.图 1 中表示了结构图,包含 5 个要素和 5 个属性.5 个要素和 5 个属性相互之间的相互影响关系比图 1 中所表示的要复杂得多,图 1 表示的仅仅是基本的隶属关系.

1.2 应用多维随机变量对公众事件进行建模

设一个互联网公众事件由 n 个随机变量构成,则事件可以表示为$(X_1,\ldots,X_n)$,事件本身为一个随机信息系统,用 X 表示,则 X 等价于$(X_1,\ldots,X_n)$.

这里,我们取 n=10,其中,X_1 表示公众事件的主体名称,X_2 表示主体的社会或自然角色,X_3 表示社会或自然关系,X_4 表示主体所属的机构或体系的名称,X_5 表示时间信息,X_6 表示主体的社会(自然)行为,X_7 表示事件的地址信息,X_8 为事件的数量信息,X_9 为舆情事件的类别,X_{10} 为未抽取信息.

2 公众事件熵的计算方法

2.1 香农信息熵

信息的可度量、可计算,是人类对信息技术掌握的里程碑.香农在信息论的研究中贡献最为显著,下面我们阐述一下相关理论.

香农理论的重要特征是熵(entropy)的概念.香农证明了熵与信息内容的不确定程度有等价关系[16].

定义 1. 一个随机变量 X 的熵 $H(x)$定义为

$$H(X)=-\sum_x p(x)\log p(x) \tag{1}$$

一个随机变量 X 的熵 $H(x)$是概率分布 $p(x)$的函数,它衡量了包含在 X 中的平均信息量.

下面,我们依据此公式计算公众事件的信息熵.

2.2 基于最大熵理论的计算方法

1) 理论描述

最大熵原理最初是由 Jayness 在 1950 年提出来的[17].

结论:对一个随机过程,如果没有任何观测量,即没有任何约束,则解为均匀分布.

2) 最大熵建模

最大熵统计建模是以最大熵理论为基础的一种选择模型的方法,即从符合条件的分布中选择熵最大的分布作为最优的分布:

$$p'=\arg\max H(p).$$

3) 熵函数取最大值时的概率分布

以(0-1)分布的熵函数为例,在概率γ=0.5 时,γ取得均值的位置出现最大熵值.其他类型概率分布函数的熵函数情况相似,也存在最大值,在概率分布取得均值的点获得最大熵(如图 2 所示).

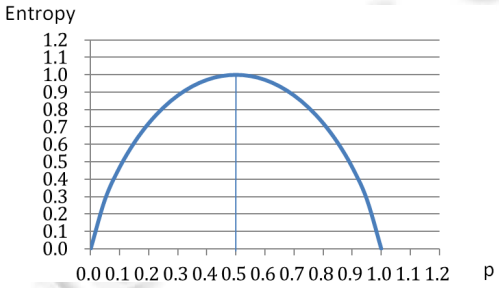

Fig.2 Entropy function of (0-1) probability distribution

图 2 (0-1)概率分布的熵函数

2.3 最大熵原理的数学表示

2.3.1 最大熵的数学表示

1) 在给定的约束条件下,由最大熵原理求解最佳概率分布,即应用拉格朗日乘子法求解条件极值问题[18].

2) 求解过程.

求 n 元函数的 $f(x_1,x_2,\ldots,x_n)$在 $m(m<n)$个约束条件下的条件极值,常数 $1,\lambda_1,\ldots,\lambda_m$ 依次乘 $f,\varphi_1,\ldots,\varphi_m$,然后累加起来得函数 $F(x_1,x_2,\ldots,x_n)$:

$$\begin{cases}\varphi_1(x_1,x_2,...,x_n)=0\\ \varphi_2(x_1,x_2,...,x_n)=0\\ ...\\ \varphi_m(x_1,x_2,...,x_n)=0\end{cases},$$

$$F(x_1,x_2,...,x_n)=f+\lambda_1\varphi_1+\lambda_2\varphi_2+...+\lambda_m\varphi_m.$$

然后列出 $F(x_1,x_2,\ldots,x_n)$无约束条件时具有极值的必要条件:

$$\begin{cases}\dfrac{\partial F}{\partial x_1}=\dfrac{\partial f}{\partial x_1}+\lambda_1\dfrac{\partial\varphi_1}{\partial x_1}+\lambda_2\dfrac{\partial\varphi_2}{\partial x_1}+...+\lambda_m\dfrac{\partial\varphi_m}{\partial x_1}=0\\ \dfrac{\partial F}{\partial x_2}=\dfrac{\partial f}{\partial x_2}+\lambda_1\dfrac{\partial\varphi_1}{\partial x_2}+\lambda_2\dfrac{\partial\varphi_2}{\partial x_2}+...+\lambda_m\dfrac{\partial\varphi_m}{\partial x_2}=0\\ ...\\ \dfrac{\partial F}{\partial x_n}=\dfrac{\partial f}{\partial x_n}+\lambda_1\dfrac{\partial\varphi_1}{\partial x_n}+\lambda_2\dfrac{\partial\varphi_2}{\partial x_n}+...+\lambda_m\dfrac{\partial\varphi_m}{\partial x_n}=0\end{cases}.$$

把这 n 个方程和 m 个约束条件方程进行联立,即可求出 $n+m$ 个 $x_1,x_2,\ldots,x_n,\lambda_1,\lambda_2,\ldots,\lambda_m$ 的值,其中,$x_1,x_2,\ldots,x_n$ 就是可能的极值点,称为驻点.

因为熵函数 $H(x)$是分布函数 $f(x)$的泛函,于是用拉格朗日乘子法求出的解就不再是 x,而是 $f(x)$.

2.3.2 离散型随机变量的最大熵分布形式

设离散型随机变量 X 取得有限个值 $x_1,x_2,\ldots,x_n$,相应的概率记为 $p_1,p_2,\ldots,p_n$,则 $H(x)$最大的充要条件:

$$p_1=p_2=\ldots=p_n=\frac{1}{n}.$$

证明:由于 $\sum_{i=1}^{n}p_i=1$,根据拉格朗日乘子法求解此约束条件下熵最大概率分布.设

$$F(p_1,p_2,\ldots,p_n)=-\sum_{i=1}^{n}p_i\ln p_i+\lambda\left(\sum_{i=1}^{n}p_i-1\right).$$

对 p_i 求偏导数,根据求取最值的必要条件,得到方程组:

$$\partial F/\partial p_i=-\ln p_i-1+\lambda=0,i=1,2,\ldots,n.$$

求解:$p_i=\exp(\lambda-1)$,为常数.

根据约束条件 $\sum_{i=1}^{n}p_i=1$,则 $np_i=1$,即 $p_i=1/n$.

此时,熵函数:

$$H(x)=-\sum_{i=1}^{n}(1/n)\ln(1/n)=\ln(n) \tag{2}$$

对于取值为有限值的离散型随机变量来说,当每一个取值的概率相等时,其信息熵最大,此时的分布为最大熵分布.

重要结论:得到了一个关于 n 的严格单调函数 $H(x)=\ln(n)$.本文利用这个结论进行公众事件信息熵的社会计算,可以保证计算结果具有严格单调性(如图 3 所示). □

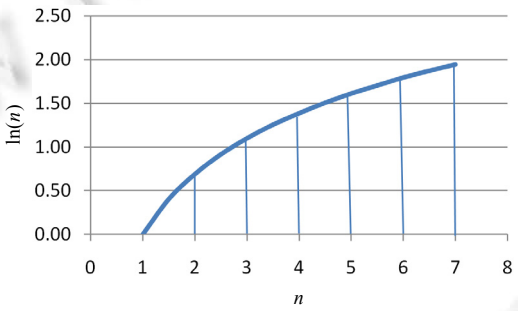

Fig.3 Monotony of the entropy function $H(x)=\ln(n)$

图 3 熵函数 $H(x)=\ln(n)$的单调性

3 应用最大熵理论计算公众事件的信息熵

3.1 公众事件建模

用随机变量 X 表示公众事件表示,X 等价于$(X_1,\ldots,X_{10})$,其中,X_1 为事件主体,X_2 为社会(自然)角色,X_3 为关系,X_4 为所属机构或体系,X_5 为事件发生时间,X_6 为行为,X_7 为发生地点,X_8 为数量,X_9 为事件类别,X_{10} 为未抽取信息.

3.2 多维随机变量的向量空间

设一个公众事件可以由多维随机变量$(X_1,\ldots,X_m)$表示,我们分别确定各个分变量的取值范围,并组合构成一个多维向量空间[19].

定义 2. 当多维随机变量的取值都是基本取值集合内元素时,此事件为元事件,以$(x_1,x_2,\ldots,x_n)$表示.所有的分向量的取值元素集合组合在一起构成了公众事件的多维向量空间.这里的集合元素指的是文本事件抽取项的关键词或短语.

举例说明,以“80 后清华硕士任副局长后受贿 1 600 万,被判无期”事件为例,对文本形式的事件进行信息抽取,得到以下形式:“80 后清华硕士任副局长后受贿 1 600 万,被判无期”事件(见表 1).

Table 1 Distribution of multidimensional random variables

表 1 多维随机变量的分布

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
80后	80后-〈专有名词,敏感词汇〉	〈清华硕士,清华大学〉:校友关系	〈肖明辉,海南省洋浦经济开发区规划建设土地局〉	10月15日(国庆期间,敏感时段)	受贿	海南省-〈国家省级行政域,...区域〉	涉案人员:2人	经济犯罪
清华	清华(大学)-〈专有名词,敏感词汇,国家著名大学一类,国家教育机构,国家事业单位〉	〈海南省第九届十大杰出青年,清华大学〉:校友关系	〈张成梁,海南省洋浦经济开发区规划建设土地局〉	2007年	收1 611万元好处费	海南省洋浦经济开发区-〈国家经济开发区,...〉	受贿:1 611万元	官员违纪
清华硕士	清华硕士-〈专有名词,敏感词汇,国家名牌大学一类毕业生〉	〈副局长,清华大学〉:校友关系	…	2007年底	为他人谋取不正当经济利益	屯昌-〈国家县级行政区域〉	受贿:6万元	官员受贿
…	…	…	…	2008年3月	…	海口-〈国家市级行政区域,...〉	年龄:32岁	…
赵某	赵某-〈个体营业者,专有名词>	〈国家公务员,副局长>:上下级关系	…	2009年	提供虚假发票	…	…	…
该工程	该工程-〈专有名词,敏感词汇,工程类名词〉	…	…	2009年	支付好处费	…	…	…
人民币	人民币-〈专有名词,敏感词汇,财经类名词〉	…	…	2011年	帮助赵某中标工程	…	…	…
6万元	6万元-〈专有名词,敏感词汇,钱款类名词〉	〈司机,副局长〉:紧密上下级关系	〈张成梁,行政机构〉	…	收取6万好处费	…	…	…

X_1-事件主体: 80 后;清华;清华硕士;副局长;1 600 万;无期;科员;副局长;“杰出青年”;“光环”;肖明辉;5 亿元;工程招标;大权;1 611 万元;“好处费”;经济利益;肖明辉;海南省;海南省二中院;无期徒刑;政治权利;…;赵某;该工程;人民币;6 万元.

X_2-社会角色: 80 后-〈专有名词,敏感词汇〉;清华(大学)-〈专有名词,敏感词汇,国家著名大学一类,国家教育机构,国家事业单位〉;清华硕士-〈专有名词,敏感词汇,国家名牌大学一类毕业生〉;副局长-〈专有名词,敏感词汇,国家公务员,国家处级公务员〉;…;肖明辉-〈国家公务员,国家官员,国家处级公务员,工程管理类国家公务员,名牌大学一类毕业生,硕士学历人员〉;…;赵某-〈个体营业者,专有名词〉;该工程-〈专有名词,敏感词汇,工程类名词〉;人民币-〈专有名词,敏感词汇,财经类名词〉;6 万元-〈专有名词,敏感词汇,钱款类名词〉.

X_3-社会关系: 〈清华硕士,清华大学〉:校友关系;〈海南省第九届十大杰出青年,清华大学〉:校友关系;

〈副局长,清华大学〉:校友关系;〈政府官员,清华大学〉:校友关系;〈国家公务员,清华大学〉:校友关系;…;〈国家公务员,副局长〉:上下级关系,…;〈司机,副局长〉:紧密上下级关系.

X_4-所属机构: 〈肖明辉,海南省洋浦经济开发区规划建设土地局〉;〈张成梁,海南省洋浦经济开发区规划建设土地局〉;…;〈张成梁,行政机构〉.

X_5-事件发生时间: 2012 年 10 月 15 日;2007 年;2007 年底;2008 年 3 月;2009 年;2009 年;2011 年.

X_6-行为: 受贿;收 1 611 万元好处费;为他人谋取不正当经济利益;注册空头公司;牵线搭桥;签订 10 份虚假合同;签订虚假劳务合同;被判无期徒刑;剥夺政治权利终生;提供虚假发票;支付好处费;帮助赵某中标工程;收取 6 万好处费.

X_7-事件发生地点: 海南省;海南省洋浦经济开发区;屯昌;海口.

X_8-数量: 涉案人员:2 人;受贿:1 611 万元|6 万元;年龄:32 岁.

X_9-事件类别: 经济犯罪|官员违纪|官员受贿.

由最大熵理论可知,随机变量的取值项数量越多,即内容越“杂乱”,最大熵值就越大.这可以解释为什么一些包含复杂内容(如社会角色和关系等)的公众事件容易引起关注,因为事件本身信息量较大,或者直观解释为事件内容更加丰富,对表 1 中的信息系统进行向量抽取,显然其信息冗余较大,也就是信息量较大.

由于 $X_1,X_2,\ldots,X_{10}$ 之间的函数关系无法确定,所以此问题适合使用最大熵模型来解决,以最大熵表征公众事件的熵值,与实际的情况最为接近.

3.3 公众事件信息熵的计算公式

在公众事件信息量的计算中属于约束条件 $\sum p(x_1,x_2,...,x_9)=1$ 的最大熵问题,其熵函数的形式与一维随机变量的形式类似,信息熵值可以为任意正数.

取最大熵时,其联合概率分布为均匀分布,则计算公式可以表示为

$$H(X_1,X_2,...,X_n)=-\sum_x p(x_1,x_2,...,x_n)\log p(x_1,x_2,...,x_n)=-\log p(x_1,x_2,...,x_n).$$

用 q_i 表示随机变量 X_i 的取值次数总数,当有一次基本集合的取值时,q_i=1(见表 2).

Table 2 Value of subvariables

表 2 分变量的取值

X	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
取值项数q_i	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9

由约束条件可知, $p(x_1,x_2,...,x_9)=\dfrac{1}{q_1q_2...q_9}$,则熵函数表示为

$$H(X_1,X_2,...,X_9)=-\log p(x_1,x_2,...,x_9)=-\log\frac{1}{q_1q_2...q_9}=\log(q_1q_2...q_9) \tag{3}$$

此公式为公众事件多维随机变量的信息熵计算公式,其形式具有严格的单调关系.下面证明其单调性,并分析如何计算出事件信息熵值.

3.4 多维随机变量信息熵的单调性证明

由公式(3),熵函数:$H(X_1,X_2,\ldots,X_9)=-\log p(x_1,x_2,\ldots,x_9)=\log(q_1q_2\ldots q_9)$,证明其具有单调性.

证明:设 $X_1,X_2,\ldots,X_9$ 取得一组 $(q_1...q_i'...q_9)$ 熵函数,取值为 H'',而另一组取得的 $(q_1...q_i''...q_9)$ 熵函数取值为 H'. 当 $q_i''>q_i'$ 时,q 取得正整数值,可知, $q_1...q_i''...q_9>q_1...q_i'...q_9$, 进而得知, $\log(q_1...q_i''...q_9)>\log(q_1...q_i'...q_9)$. 所以,熵函数具有严格的单调性,$H''>H'$.可知,这个多维随机变量的熵函数具有严格单调性.证毕. □

3.5 中文语言特性对公众事件信息熵的影响

中文公众事件的信息熵值必然受到中文语言特性的影响,中文语言是一种意合语言,中文的特点是概括性

强,语言表述往往包含很多汉语成语、典故、常用语等,这样的语言往往简短,但却包含了比词本身丰富得多的含义.这体现在信息熵值计算方面,必然造成信息熵值的增大,这种中文语言特性对事件信息熵影响较大.

例如,有这样一则公众信息表述:

谎称收工程保证金,七旬老汉“指鹿为马”诈骗百万(2015-04-17 09:18:58 来源:胶东在线)

胶东在线网 4 月 17 日讯(记者侯嘉伟通讯员徐忠孙世建)2014 年以来,蓬莱市公安局经侦大队共破获以收取工程保证金为名实施的合同诈骗案件 4 起,抓获犯罪嫌疑人 20 余名,涉案金额达 5 000 余万元.

…

经查,1945 年出生的柯某利利用相同手段共诈骗 150 余万元.2014 年 11 月 6 日因涉嫌合同诈骗被刑事拘留,同年 12 月 12 日被批准逮捕,现该案已移送检察部门审查起诉.

事件中使用了“指鹿为马”这样的成语,在中文社区,这样的表述会激发读者头脑中的语义框架,读者获得了成语中丰富的信息,“指鹿为马”成语中包含的信息就“嵌入”到了事件当中,这在中文公众事件表达当中属于常见现象.

“指鹿为马”这样的成语,包含的信息内容是比较固定的,构成一个封闭的独立语境事件.

同样,我们可以计算其熵值,在计算过程中,可以把这个熵值当作常数,累加到事件信息熵值中.

“指鹿为马”的文本信息摘要描述如下:

“指鹿为马”:出自《史记·秦始皇本纪》,秦始皇死后,赵高试图要谋朝篡位,为了实验朝廷中有哪些大臣顺从他的意愿,特地呈上一只鹿给秦二世胡亥,并说这是马.秦二世不信,赵高便借故问各位大臣.不敢逆赵高意的大臣都说是马,而敢于反对赵高的人则说是鹿.后来说是鹿的大臣都被赵高用各种手段害死了.指鹿为马的故事流传至今,人们便用指鹿为马形容一个人是非不分,颠倒黑白.

经过信息抽取计算后,得到“指鹿为马”典故成语的熵值为 M,事件信息熵值为 H',则最终的事件信息熵为 $H=H'+M$.

这可以解释,为什么使用成语、典故较多的事件描述更容易引起读者的兴趣,其中一个原因是其造成了事件信息熵的增加.

4 公众事件熵的计算过程

在计算公众事件信息熵时,基于社会学理论及一些领域知识,我们可以把它以“关键词或同义词短语”的形式集成到我们的知识库中,这里需要专家的人工知识分析.一旦知识库建立后,会为我们提供很大的便利.

由第 3.4 中的单调性证明,我们这里构建的关键词知识库只需要按社会学知识划分不同的子集合,并进行关键词的匹配或短语的同义词替换,然后进行关键词匹配计算即可.由此而产生的计算属于社会计算.

以 2012 年度全年的互联网中文公众事件为实验数据集进行计算,语料库中统计了中国全年 1 200 个中文事件案例(每个季度 300 件公众事件),这是全年爆发的互联网中文公众事件中引起社会重视较高的事件,文中选取部分事件的计算结果进行分析.

4.1 构建知识库

设 X_n 为公众事件 X 的某个分随机变量(如 X_1),离散型随机变量,假设 X_n 的取值集合为 M,M 包含若干个子集 $M_1,M_2,M_3,\ldots,M_n\subset M$,同时满足 $M_1\cup M_2\cup M_3\cup\ldots\cup M_n=M$.

由于各个国家的历史、文化、习俗、宗教、固有观念等社会状况有很大的区别,所以特定的国家或地区要有特定的分析,相应的随机变量的概率分布情况也会有很大的区别.比如“驻阿富汗美军烧古兰经事件”,如果发生在其他非信仰伊斯兰教的地区,事件就不会这么敏感,不会引起这么大规模和广泛的争端.

本文在使用通用计算方法的基础上,以中国国内社会状况、文化特点为背景进行互联网社会计算研究,如果要计算其他国家的互联网公众事件信息熵,则要根据实际情况进行相应的知识库调整.

下面我们以中文公众事件的计算为例,分别分析 9 个随机变量的取值集合情况,给出一个互联网公众事件信息熵的具体计算方法.集合中的元素都是有代表性的关键词,这些关键词或同义短语构成了知识库,考察 9 个

随机变量的取值集合,可以构建相应的知识库.这里给出简略描述.

4.1.1 分析随机变量 X_1(公众事件中的主体名)的取值范围

事件的主体名往往是人物的名称,也有地名、机构名和其他类型主体的名称.把集合按子集合划分,当有一次关键词匹配时,q_1=1;若没有,则取 q_1=0.

设 X_1 为表示公众事件主体名的随机变量,是离散型随机变量,建立 X_1 的取值集合 M,其中包含若干个子集 $M_1,M_2,M_3,M_4 \subseteq M$,满足 $M_1 \cup M_2 \cup M_3 \cup M_4 = M$.

根据常识,人名或地名等具有公众信息敏感度,我们把知名度分为 4 个等级,分别对应 X_1 的 4 个取值子集合,其中,M_4 为取值的基本集合.

公众信息敏感度级别的划分如下:

M_1 为公众信息敏感度第 1 等级,可继续划分子集 $L_1,L_2,\ldots$;

L_1 历史名人,L_1={曹操,李鸿章};L_2 当代政治人物,L_2={奥巴马,普京,…};

…;

M_4 公众信息敏感度第四等级,可继续划分子集 $L_1,L_2,\ldots$

本文中,划分等级是为了方便说明问题,当有匹配时 q_1 取值相同,取值为 1.社会计算中使用带有加权值的运算方法,留待后续研究中系统介绍.

即,当 $X_1=x_1$ 时,若 $x_1 \in M_1 \sim M_4$,则取得 q_1=1;否则,q_1=0.以下各项情况类似.

形式化命题逻辑判断,如下描述.可以看到,进行匹配计算的过程就是进行一阶谓词逻辑判断的过程.

命题 A. X_1 有一个取值,即,当 $X_1=x_1$ 时,逻辑为真.

命题 B. 当 $x_1 \in M_1$ 或 $x_1 \in M_2 \ldots$,或 $x_1 \in M_4$ 其中一个成立时,逻辑为真.

这样,当 A∧B 的合取式为真时,表示 q_1 有一次取值,为 1.

当 A∧B 的合取式为假时,表示 q_1 有一次取值,为 0.这种情况下,对计算值无贡献.

4.1.2 分析随机变量 X_2(社会(自然)角色)的取值范围

设 X_2 为表示公众事件主体的社会角色的随机变量,是离散型随机变量,我们建立 X_2 的取值集合 M,并包含若干个子集 $M_1,M_2,M_3,\ldots,M_{43} \subseteq M$,满足 $M_1 \cup M_2 \cup M_3 \cup \ldots \cup M_{43} = M$.

由于互联网空间的出现,相应地出现了许多新的社会角色,如互联网的公知人群、意见领袖人群,还有部分网络文化名人等,并担当起了相应的社会责任,发挥着某种社会功能.从社会学的角度来分析,互联网不但改变了人们获取知识的方式,同时新的社会角色也在一定程度上改变了人们之间的关系,产生了新的信息传播与信任方式,比如“公知”、“意见领袖”、“微博大 V”等.

作为 X_1“主体”的属性,X_2 社会(自然)角色是构成事件信息内容的重要因素.因为在互联网公众事件中,主体的角色对事件信息引起关注的程度影响极大,一个事件的主体可以有多个社会角色.

这里取“主体”的职位名称、地名的行政区域身份或属性名称、商业实体名称、商业人士的职位名称或是特殊人群的社会名称等关键词,作为社会(自然)角色的描述.

子集 M_1 为自然灾害类型名称,M_2 为星际名称集合,M_3 为地理名的社会(自然)角色,M_4 为国家自然类别集合,M_5 为特殊国家类别,…,M_{39} 为学生类别集合,M_{40} 为未成年人,M_{41} 为敏感角色(如奶粉业、明胶业、三鹿乳业等),M_{42} 为普通民众集合,M_{43} 为其他角色.

允许 $M_p \cap M_k \neq \varnothing$,$1 \leqslant p,k \leqslant 43$.

我们逐项分析 $M_1,M_2,\ldots,M_{43}$.

M_1 为自然灾害严重程度集合,可继续划分子集 $L_1,L_2,\ldots,L_6$.满足 $L_1,L_2,\ldots,L_6 \subset M_1$,$L_1 \cup L_2 \cup \ldots \cup L_6 = M_1$.

我们按自然灾害的级别进行划分子集:

L_1 为较轻型灾害集合,L_1={霜冻,虫害,降温,干旱,…};

…;

L_6 为其他类型灾害集合,L_6={冰冻,虫害,降温,干旱,…};

…;

M_{42}为普通民众集合,其权值为 1;M_{43}为其他角色集合,体现完备性,权值也为 1.

与第 4.1.1 节中类似,进行如下一阶谓词逻辑判断:

命题 C. X_2有一个取值,即,当$X_2=x_2$时,逻辑为真.

命题 D. 当$x_2\in M_1$,或$x_2\in M_2$,…,或$x_2\in M_{43}$其中一个成立时,逻辑为真.

这样,当 C∧D 的合取式为真时,表示q_2有一次取值,为 1.

当 C∧D 的合取式为假时,表示q_2有一次取值,为 0.这种情况下,对计算值无贡献.

4.1.3 分析随机变量X_3(“自然关系”或“社会关系”)的取值范围

设X_3表示公众事件的社会关系,是离散型随机变量,我们建立X_3的取值集合M,并包含若干个子集$M_1,M_2,M_3,M_4\subseteq M$,满足$M_1\cup M_2\cup M_3\cup M_4=M$.

由于这里考察的随机变量X_3为事件主体的某种关系,“自然关系”或“社会关系”会对公众事件本身的信息量有很大的“贡献”.

我们把关系一项分为:M_1为强关系、M_2为中等关系、M_3为弱关系、M_4为其他关系.

M_1为强关系,M_1={母子关系,父子关系,敌对关系,历史宿怨关系,…};

…;

M_4为其他关系,如下所示.

使用二元组来作形式化的表示如下形式:

设以y表示“实体 1”,以z表示“实体 2”,则它们之间的关系可以表示为$x_3=\langle y,z\rangle$,且$q_3=f(\langle y,z\rangle)$.

若$\langle y,z\rangle\in M$,则q_3值取得 1,即$q_3=1$;否则,$q_3=0$.

进行如下一阶谓词逻辑判断:

命题 E. X_3有一个取值,即,当$X_3=x_3$时,逻辑为真.

命题 F. 当$x_3\in M_1$,或$x_3\in M_2$,…,或$x_3\in M_4$其中一个成立时,逻辑为真.

这样,当 E∧F 的合取式为真时,表示q_3有一次取值,为 1.

当 E∧F 的合取式为假时,表示q_3有一次取值,为 0.这种情况下,对计算值无贡献.

4.1.4 分析随机变量X_4(主体所属机构名称或所属体系的名称)的取值范围.

设X_4为公众事件的机构名称,是离散型随机变量,我们建立X_4的取值集合M,并包含若干个子集$M_1,M_2,M_3,\ldots,M_6\subset M$,满足$M_1\cup M_2\cup M_3\cup\ldots\cup M_6=M$.

按机构的重要程度分为 5 级,细节略.

进行如下一阶谓词逻辑判断:

命题 G. X_4有一个取值,即,当$X_4=x_4$时,逻辑为真.

命题 H. 当$x_4\in M_1$,或$x_4\in M_2$,…,或$x_4\in M_6$其中一个成立时,逻辑为真.

这样,当 G∧H 的合取式为真时,表示q_4有一次取值,为 1.

当 G∧H 的合取式为假时,表示q_4有一次取值,为 0.这种情况下,对计算值无贡献.

4.1.5 分析随机变量X_5(时间信息)的取值范围

设X_5为公众事件的时间信息,是离散型随机变量,随机变量取值的所属时段作为集合M的元素,并包含若干个子集$M_1,M_2,M_3,\ldots,M_6\subset M$,满足$M_1\cup M_2\cup M_3\cup\ldots\cup M_6=M$.

按时段的重要程度由高到低,分为 6 级:

第 1 级M_1为灾害时期,如洪水、疾病暴发等时期,其子集为$L_1,L_2,\ldots$;

…;

第 5 级M_5为季节性时段,如春运期、休渔期、春播期、洪讯期、冰霜期、禁海期等;

第 6 级M_6为其他时段.

进行如下一阶谓词逻辑判断:

命题 I. X_5 有一个取值,即,当 $X_5=x_5$ 时,逻辑为真.

命题 J. 当 $x_5\in M_1$,或 $x_5\in M_2$,…,或 $x_5\in M_6$ 其中一个成立时,逻辑为真.

这样,当 I∧J 的合取式为真时,表示 q_5 有一次取值,为 1.

当 I∧J 的合取式为假时,表示 q_5 有一次取值,为 0.这种情况下,对计算值无贡献.

4.1.6 分析随机变量 X_6(社会(自然)行为)的取值范围

设 X_6 为舆情事件的社会行为,是离散型随机变量,我们建立 X_6 的取值集合 M,并包含若干个子集 $M_1,M_2,M_3,\ldots,M_6\subset M$,满足 $M_1\cup M_2\cup M_3\cup\ldots\cup M_6=M$.

我们依据社会学构建理论[10]对事件的行为进行划分:M_1 为自然灾害类社会行为;M_2 为邪教类行为、反人类行为、恶性刑事犯罪行为;M_3 为宗教类行为、群体性行为;M_4 为造谣中伤类行为、恶意商业攻击类事件、恶意人身攻击类事件、或普通犯罪行为等;M_5 普通个人意见表达、商业网络信息发布或讨论行为、普通民事纠纷等,M_6 为其他行为类型.

进行如下一阶谓词逻辑判断:

命题 K. X_6 有一个取值,即,当 $X_6=x_6$ 时,逻辑为真.

命题 L. 当 $x_6\in M_1$,或 $x_6\in M_2$,…,或 $x_6\in M_6$ 其中一个成立时,逻辑为真.

这样,当 K∧L 的合取式为真时,表示 q_6 有一次取值,为 1.

当 K∧L 的合取式为假时,表示 q_6 有一次取值,为 0.这种情况下,对计算值无贡献.

4.1.7 分析随机变量 X_7(事件发生的地址信息)的取值范围

设 X_7 为公众事件的地址信息,是离散型随机变量,我们建立 X_7 的取值集合 M,并包含若干个子集 $M_1,M_2,M_3,\ldots,M_{14}\subset M$,满足 $M_1\cup M_2\cup M_3\cup\ldots\cup M_{14}=M$.

M_1 为地址名称,M_2 为国家地名,M_3 为国家首都地名,M_4 为国家州省地名,M_5 为省会城市地名,M_6 为地市级城市地名集合,M_7 为县级地名,M_8 为乡镇以下级地名,M_9 为具有政治意义的地名集合,M_{10} 为文化名城集合,M_{11} 为著名风景区集合、M_{12} 为著名国家保护区集合,M_{13} 为娱乐场所,M_{14} 为其他地名集合.

进行如下一阶谓词逻辑判断:

命题 N. X_7 有一个取值,即,当 $X_7=x_7$ 时,逻辑为真.

命题 O. 当 $x_7\in M_1$,或 $x_7\in M_2$,…,或 $x_7\in M_{14}$ 其中一个成立时,逻辑为真.

这样,当 N∧O 的合取式为真时,表示 q_7 有一次取值,为 1.

当 N∧O 的合取式为假时,表示 q_7 有一次取值,为 0.这种情况下,对计算值无贡献.

4.1.8 分析随机变量 X_8(事件中数量信息)的取值范围

设 X_8 为公众事件社会行为的涉及数量,是离散型随机变量,我们建立 X_8 的取值集合 M,并包含若干个子集 $M_1,M_2,\ldots,M_5\subset M$,满足 $M_1\cup M_2\cup M_3\cup\ldots\cup M_5=M$.

按事件中数量的重要程度分 5 个级别:

第 1 级的数量,M_1 其子集为 $L_1,L_2,\ldots$.

L_1={地震级数 6 级以上,台风 8 级以上,…},L_2={死亡人数 10 人以上};

…;

第 5 级的数量,M_5 其子集为 $L_1,L_2,\ldots$

进行如下一阶谓词逻辑判断:

命题 P. X_8 有一个取值,即,当 $X_8=x_8$ 时,逻辑为真.

命题 Q. 当 $x_8\in M_1$,或 $x_8\in M_2$,…,或 $x_8\in M_5$ 其中一个成立时,逻辑为真.

这样,当 P∧Q 的合取式为真时,表示 q_8 有一次取值,为 1.

当 P∧Q 的合取式为假时,表示 q_8 有一次取值,为 0.这种情况下,对计算值无贡献.

4.1.9 分析随机变量 X_9(公众事件中的类别信息)的取值范围

设 X_9 为公众事件的类别名,是离散型随机变量,我们建立 X_9 的取值集合 M,并包含若干个子集 $M_1,M_2,M_3,\ldots,$

$M_6 \subset M$,满足 $M_1 \cup M_2 \cup M_3 \cup \ldots \cup M_6 = M$.

我们依据社会学构建理论[20]对事件的类别领域进行划分,此项与 X_9 项相对应.

M_1 为自然灾害类事件集合,M_2 为邪教类、反人类事件、恶性刑事犯罪事件集合,M_3 为宗教类、群体性事件、群体行为事件集合,M_4 为造谣中伤类事件集合、恶意商业攻击、人身攻击事件,M_5 为普通个人信息发布、商业网络信息发布或讨论类事件,M_6 为其他事件类别集合.

进行如下一阶谓词逻辑判断:

命题 R. X_9 有一个取值,即,当 $X_9=x_9$ 时,逻辑为真.

命题 S. 当 $x_9 \in M_1$,或 $x_9 \in M_2$,…,或 $x_9 \in M_6$ 其中一个成立时,逻辑为真.

这样,当 R∧S 的合取式为真时,表示 q_9 有一次取值,为 1.

当 R∧S 的合取式为假时,表示 q_9 有一次取值,为 0.这种情况下,对计算值无贡献.

4.1.10 X_{10} 为公众事件信息抽取过程中未抽取的信息

此随机变量是为了体现公众事件信息量定义的完备性,对事件的信息量计算没有贡献,不计算这一项.

9 个随机变量知识库的集合划分不是唯一的划分方法,这里所做的计算属于社会计算,要根据实际情况进行调整.

4.2 计算信息熵

当对事件进行信息抽取并进行知识库进行匹配计算后,可以得到 $q_1,q_2,\ldots,q_9$ 的值.根据第 3.3 节中公式(3)计算信息熵值,则 $H(X_1,X_2,\ldots,X_9)=\log(q_1,q_2,\ldots,q_9)$.

5 实 验

5.1 计算信息熵

计算信息抽取形式的“80 后清华硕士任副局长后受贿 1 600 万,被判无期事件”的信息熵值,如第 3.2 节中的形式.逐项匹配计算 q_i 值,见表 3,这里采用自然对数计算.

Table 3 Weight of X

表 3 X 的加权值

X	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
q_i值	104	342	15	8	6	41	10	10	3

H=ln(104×342×15×8×6×41×10×10×3)=26.48,取小数点后两位有效数字.

5.2 同类案例事件的熵值比较

以 2012 年第 4 季度公众事件为例,我们进行了繁琐的信息项信息抽取,并进行了相应的复杂计算,数据量和计算量都较大,这里选取“官员违纪类事件”进行了实验结果展示.

表 4 中熵值 1 的数据项显示为信息抽取后的计算值,此实验是为了验证计算方法的单调性,比较不同的事件包含的信息量,如图 4 所示.

我们根据表 4 的数据排序给出趋势图,熵值 1 列项为纵坐标.可以看到,得到了一个趋势性的单调关系.趋势线表明了我们计算方法的合理性,与理论分析第 3.4 节中单调性证明的结论相符合,是计算方法科学性的体现.

我们看到,其中最小的熵值事件为“涪陵艳照门事件当事者为执法干部 监察局立案调查”,值为 17.33,这是因为其文本事件描述很短,处于事件的爆发初期,内容所包含的信息较少的缘故;熵值最大的事件为“街道党工委书记受贿被审:732 万买景德镇瓷器”,因为事件已经调查完毕,并且已经由法院给出了详细的判决,其文本内容包含详细的内容,所以其信息量较大,这与我们的直觉接近.

Table 4 Ranking of calculation

表 4 计算结果排序

官员违纪类事件	排名	熵值1	熵值2	熵值3
街道党工委书记受贿被审:732万买景德镇瓷器	1	36.15	34.58	33.17
山西4妻10子村官人大代表资格被暂停已取保候审	2	35.46	34.31	33.18
杭州房管局副局长被指拥20多套房价值数亿	3	33.89	31.67	30.89
陕西“表哥”存款涉20多家银行	4	33.34	31.24	29.90
广州越秀区原城管局长涉嫌受贿178万受审	5	33.23	31.73	30.06
长沙副处级官员贪污7000余万被小三情妇揭发	6	31.64	30.13	28.79
新疆乌苏公安局长被指包养双胞胎当地纪委调查	7	30.91	29.86	29.04
太原市公安局局长被停职网传其子涉醉驾殴打交警	8	30.07	28.63	27.94
湖北一女县长被指持钞票炫富当地宣传部门否认	9	28.10	27.06	25.95
广西桂林一村委组长涉嫌贪污9万公款被判刑8年	10	26.68	25.31	24.37
广州一城管队长受贿400余万称怕得罪人才收钱	11	26.63	25.02	23.88
中纪委:李春城涉嫌严重违纪正接受组织调查	12	26.62	25.31	24.11
80后清华硕士任副局长后受贿1600万被判无期	13	26.48	24.98	24.00
山西价值2亿煤矿37万贱卖当地纪委介入调查	14	25.61	24.61	23.50
国家能源局回应局长被举报:纯属污蔑造谣正报案	15	25.47	23.88	22.84
湖北通山31岁女县长8年6次破格提拔被疑潜规则	16	25.10	24.07	23.35
山东临沂一副县级干部贪污19万受贿217万余元被判刑	17	25.05	23.80	22.93
长沙市规划局原高官拥16套房女儿过生日给20万	18	24.72	23.49	22.88
北京原朝阳副区长刘希泉之子受贿诈骗拆迁款477万获刑20年	19	23.72	22.75	21.93
中国党政机关255人因公务用车问题被处分	20	22.46	21.35	20.41
重庆南川人民医院骨科主任受贿逾356万获刑11年	21	21.36	20.60	19.86
涪陵艳照门事件当事者为执法干部监察局立案调查	22	17.33	16.71	15.87

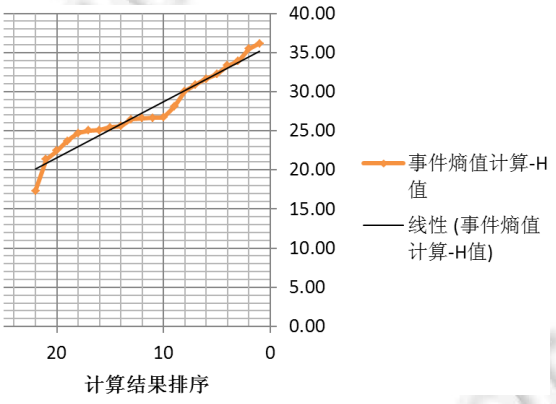

Fig.4 Verification of the calculation method rationality

图 4 计算方法的合理性验证

5.3 信息抽取方法对计算结果的影响

熵的计算值必然受到信息抽取方法[15]的影响,为了获得更为合理的计算值,往往需要对信息抽取项进行以下两步处理:

1) 重复项过滤:这个过程主要是过滤掉内容重复抽取的信息,计算结果如表 4 中熵值 2 列项所示.

2) 共指消解:过滤之后,进一步进行共指消解处理,消除掉具有共指关系的冗余信息抽取项,计算结果如表 4 中熵值 3 列项所示.

图 4 显示了进行信息抽取以后的计算结果,当进行重复项过滤与共指消解后,实验结果对比如图 5 所示,熵值比较接近的事件排序有些许的变化,但计算结果的单调性函数状态保持良好.

实验结果表明,经过滤与共指消解处理之后,对不同类型事件的计算结果影响类似,熵值在一定幅度上有所减小.

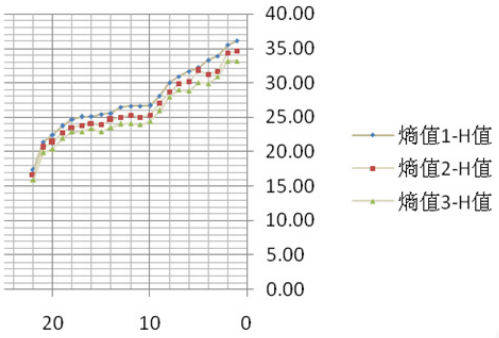

Fig.5 Experiment of contrast

图 5 对比实验

6 结束语

本文应用香农信息论和最大熵理论,给出了一个合理而且可行的计算方法,解决了互联网公众事件信息熵的定量化计算问题.文中所提到的计算方法是最大熵理论在社会计算中的一个直接应用,对于解决其他社会计算定量化问题应该有一定的借鉴意义.

文中所使用的计算方法仍然基于当前的社会计算理论基础,为了获得更加合理的计算结果,后续的研究工作可以探讨带有加权值的社会计算方法,这部分内容留待后续工作中单独进行阐述,并探讨社会计算的公理化体系问题[21].也希望其他研究人员关注该问题,共同促进这一领域的研究工作进展.

致谢 在此,我们向对本文的工作给予支持和建议的学者表示感谢.尤其是北京邮电大学的方滨兴院士,您提出的建议使我们在寻找单调函数的工作中得到启发,最终得以完成本文的工作,在此表示感谢.

References:

[1] Arab spring. https://en.wikipedia.org/wiki/Arab_Spring

[2] Public opinion. http://en.wikipedia.org/wiki/Public_opinion

[3] Key VO. Public Opinion and American Democarcy. New York: John Wiley, 2012.

[4] Mueller JE. War, Presidents, and Public Opinion. New York: Wiley, 1973.

[5] Lerman K, Gilder A, Dredze M, Pereira F. Reading the markets: Forecasting public opinion of political candidates by news analysis. In: Proc. of the 22nd Int'l Conf. on Computational Linguistics (Coling 2008). 2008. 473−480.

[6] Akcora CG, Bayir MA, Demirbas M, Ferhatosmanoglu H. Identifying Breakpoints in Public Opinion. In: Proc. of the 1st Workshop on Social Media Analytics (SOMA 2010). Washington: ACM Press, 2010. [doi: 10.1145/1964858.1964867]

[7] Li J, Zhou XG, Chen B. Research on analysis and monitoring of Internet public opinion. In: Proc. of the 2012 Int'l Conf. of Modern Computer Science and Applications Advances in Intelligent Systems and Computing, Vol.191. Berlin: Springer-Verlag, 2013. 449−453. [doi: 10.1007/978-3-642-33030-8_72]

[8] Social computing. http://en.wikipedia.org/wiki/Social_computing

[9] Lazer D, Pentland A, Adamic L, Aral S, Barabasi AL, Brewer D, Christakis NA, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Van Alstyne M. SOCIAL SCIENCE: Computational social science. Science, 2009,323(5915):721−723. [doi: 10.1126/science.1167742]

[10] Wang FY, Zeng DJ, Mao WJ. Social computing: Its significance, development and research status. e-Science, 2010,7:3−14 (in Chinese with English abstract).

[11] Chen H, Wang FY, Zeng D. Intelligence and security informatics for homeland security: Information, communication, and transportation. IEEE Trans. on Intelligent Transportation Systems, 2004,5(4):329−341. [doi: 10.1109/TITS.2004.837824]

[12] Wang FY. From Social Computing to Social Manufacturing: an upcoming industry revolution. Strategy & Policy Decision Research, 2012,27(6):658−669 (in Chinese). [doi: 10.3969/j.issn.1000-3045.2012.06.002]

[13] Wang FY, Zeng DJ, Cao ZD. Social computing methods for non-traditional security challenges enabled by the social media in cyberspace. Science & Technology Review, 2011,29(12):15−22 (in Chinese with English abstract). [doi: 10.3981/j.issn.1000-7857.2011.12.001]

[14] Wang FY. Social computing and dynamical state analysis of digitalized and networked societies. Science & Technology Review, 2005,23(9):4−6 (in Chinese with English abstract). [doi: 10.3321/j.issn:1000-7857.2005.09.002]

[15] Tan HY. Research on Chinese event extraction [Ph.D. Thesis]. Harbin: Harbin Institute of Technology, 2008 (in Chinese with English abstract).

[16] Yeung RW, Wrote; Cai N, *et al*., Trans. Information Theory and Network Coding. Beijing: Higher Education Press, 2011 (in Chinese).

[17] Jaynes ET. Information and statistical mechanics. Physical Review, 1957,106(4):620−630. [doi: 10.1103/PhysRev.106.620]

[18] Li XD. The method study about probability distribution based on the principle of maximum entropy [MS. Thesis]. Beijing: North China Electric Power University, 2008 (in Chinese with English abstract).

[19] Chen Y, Zhang HL. Overview of social computing in information security. Journal of Tsinghua University (Sci & Tech), 2011, 51(10):1323−1328 (in Chinese with English abstract).

[20] Waters M, Wrote; Yang SH, Trans. Modern Sociological Theory. Beijing: Huaxia Publishing House, 2000 (in Chinese).

[21] Zhao XS. I was in awe of the human society axiom. In: Ma XP, ed. The Humanities Reader. 2006 (in Chinese). http://www.teacherclub.com.cn/tresearch/blog/showArticle.jsp?ArticleCode=1390764846&CID=00001

附中文参考文献:

[10] 王飞跃,曾大军,毛文吉.社会计算的意义、发展与研究状况.e-Science,2010,7:3−14

[12] 王飞跃.从社会计算到社会制造:一场即将来临的产业革命.中国科学院战略与决策研究,2012,27(6):658−669. [doi: 10.3969/j.issn.1000-3045.2012.06.002]

[13] 王飞跃,曾大军,曹志冬.网络虚拟社会中非常规安全问题与社会计算方法.科技导报,2011,29(12):15−22. [doi: 10.3981/j.issn.1000-7857.2011.12.001]

[14] 王飞跃.社会计算与数字网络化社会的动态分析.科技导报,2005,23(9):4−6. [doi: 10.3321/j.issn:1000-7857.2005.09.002]

[15] 谭红叶.中文事件抽取关键技术研究[博士学位论文].哈尔滨:哈尔滨工业大学,2008.

[16] Yeung RW,著;蔡宁,等,译.信息论与网络编码.北京:高教出版社,2011.

[18] 李宪东.基于最大熵原理的确定概率分布的方法研究[硕士学位论文].北京:华北电力大学,2008.

[19] 陈昱,张慧琳.社会计算在信息安全中的应用.清华大学学报(自然科学版),2011,51(10):1323−1328.

[20] Waters M,著;杨善华,译.现代社会学理论.北京:华夏出版社,2000.

[21] 赵鑫珊.我对人类社会公理的敬畏.见:马小平,编.人文素养读本.2006. http://www.teacherclub.com.cn/tresearch/blog/showArticle.jsp?ArticleCode=1390764846&CID=00001

靳锐(1976－),男,黑龙江依安人,博士生,主要研究领域为网络与信息安全.

张玥(1975－),女,博士,讲师,CCF 专业会员,主要研究领域为网络与信息安全,社会计算与数据挖掘.

张宏莉(1973－),女,博士,教授,博士生导师,CCF 专业会员,主要研究领域为网络与信息安全,网络测量.

王星(1981－),男,博士,助理研究员,主要研究领域为网络与信息安全,网络舆情,机器学习,迁移学习,系统架构.