$M_6 \subset M$,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

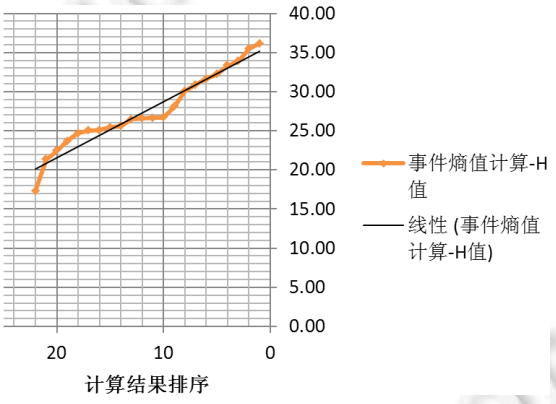

Fig.4 Verification of the calculation method rationality

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.

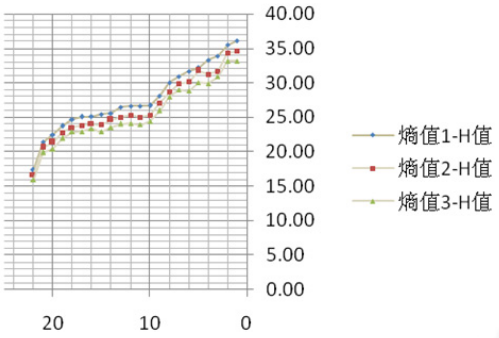

Fig.5 Experiment of contrast

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.