

基于粒子群的舆情网络用户聚类模拟与仿真

马瑞新¹ 朱明¹ 孟宇²

(大连理工大学软件学院 大连 116621)¹ (大连外国语学院现代教育技术中心 大连 116621)²

摘要 当前对网络舆情的研究大多集中于突发事件的传播规律及预警分析,而忽视了用户在舆情传播中的主体位置。针对这一问题,引入“观念空间”的概念,使用粒子群算法对突发事件传播中用户的观念聚类过程进行模拟和仿真。根据用户观念的聚类结果分析事件的动态演化模型,识别热点事件。通过改变速度参数控制用户聚类收敛速度,进而协调事件的演化过程,同时实现对网络热点事件的识别和舆情预警。最后分别对基于基本 PSO 和基于物种遗传策略的 PSO(PSPO)算法的用户聚类行为进行了仿真,实验结果表明,PSPO 算法能够有效地模拟舆情网络中用户的聚类行为,同时发现多个用户聚类中心,有利于制定自适应的舆情预警应对策略。

关键词 观念空间,粒子群算法,物种形成策略,舆情预警

中图法分类号 TP391 **文献标识码** A

Modeling and Simulating of User Clustering on Network Based on Particle Swarm Optimization

MA Rui-xin¹ ZHU Ming¹ MENG Yu²

(School of Software, Dalian University of Technology, Dalian 116621, China)¹

(Modern Education Technology Center, Dalian University of Foreign Languages, Dalian 116621, China)²

Abstract Most of the current research about public opinion on the network focuses on the analysis of emergencies' spreading and early-warning process, ignores the user's main role on the procedure of opinion-spreading. In terms of this problem, we introduced the concept of "concept space", modeled and simulated the users' concept clustering process during the transmission of emergencies on the network by using particle swarm optimization algorithm. On the basis of users' clustering results, we analyzed the dynamic evolution model of network emergencies. Changing the parameter of velocity controls the convergence rate of user-clustering, and coordinates the evolution process of network emergencies, realizes the recognition of network hot events and early-warning for public opinion crisis. At last, we simulated the users' clustering behavior relatively based on basic PSO and speciation PSO (PSPO) algorithm. Simulation results show that PSPO algorithm is able to simulate the concept clustering more effectively. It is able to find multi-clustering center at the same time, and is good to set adaptively coping strategies for early warning of public opinion.

Keywords Concept space, Particle swarm optimization, Speciation algorithm, Public opinion crisis

1 引言

网络舆情是社会舆情的一种表现方式^[1],用户借助网络对现实生活中的某些热点、焦点问题表达所持观点和言论,因此,网络舆情是用户观念的表现空间。随着互联网技术的发展和普及,网络已然成为社会公众表达社情民意、参与政治生活和舆论监督的公共平台^[2]。近年来,越来越多的舆论源于网络、兴于网络,最终引发社会关注,形成热点问题。例如:郭美美炫富事件、“我爸是李刚”、药家鑫“激情杀人”事件等。面对这个没有法律约束和道德限制的舆情监督平台,如何及时发现新的热点问题、控制事件的波及范围、引导民意的走向进而减少网络对社会造成的负面影响,已成为政府及相关部门当前亟待解决的问题。

由事件引发的舆情网络的形成不同于以往的社会网络,

它在整个网络发展的宏观上来说是偶然得到的,而在微观上却是目标引导形成的。在事件消失之后,目标引导的网络演变轨迹将丧失其目标牵引的动力,成为偶遇者的一个表现场所。舆情网络遵循这样一种混合的网络发展过程,其演变轨迹以目标引导与偶得网络的交替式循环为特征。人们(彼此陌生的人)由于偶然的因素彼此关注,在突发事件的刺激下,将网络引到实现某一目标的方向上来。而在经历一段时间以后,该目标引导的网络实现了其初始目标,这一重新聚焦到特定目标上的网络会重新回到除满足成员社交的乐趣之外别无所图的松散的行动者联合体。这样,在某一特定点上我们会看到:网络中出现了若干具有明确目标导向的局部中心,且它们之间被某种力量联结在一起。而当事件消失之后,这些小社区也随之消散。

现有的舆情网络用户聚类算法大多基于 K-means 算

到稿日期:2012-02-11 返修日期:2012-07-26

马瑞新(1975—),男,博士,讲师,主要研究领域为电子商务、社会计算、群智能,E-mail:teacher_mrx@126.com;朱明(1981—),男,硕士,讲师,主要研究领域为电子商务、自适应系统;孟宇(1978—),男,硕士,实验师,主要研究领域为现代教育、电子商务。

法^[3]、支持向量机(SVM)^[4]等,在注重用户舆论倾向性分析的基础上却没有对用户性质的分析。而舆情网络中事件的传播是由用户的推动才得以形成热点事件,因此,用户对于某一特定事件的聚类行为决定了该事件是否能够成为热点事件。因此,本文提出使用基于基本粒子群^[5]和物种形成策略的粒子群算法^[6]对突发事件传播中用户的观念聚类进行模拟和仿真。

粒子群算法由于其快速收敛、无全局控制、自适应的特点在近年来得到了广大研究学者的认可,被广泛地应用于分布式网络^[7]、图像处理^[8]、信号处理^[9]等方面,并取得了不错的效果。在本文中,网络舆情被看作是网民的世界观、价值观、道德观的体现,因此,舆情网络在一定程度上代表了网络用户的信念空间;将舆情网络中的用户看作是在信念空间中不断运动的粒子,其根据自身的经验和社会学习的过程来寻找最适合自己的位置。

2 粒子群算法与舆情网络

粒子群算法是群智能算法^[10]的一个分支,群智能算法是近年来新兴的一种仿生类优化算法^[11]。它受社会性生物行为的启发,指出:自然界中的单个个体并不表现智能特征,但是整个生物种群却表现出处理复杂问题的能力。对于没有集中控制且不提供全局模型的问题,群智能提供了一种高效的分布式解决方案^[7]。它的最大特点是不需要建立问题本身的精确模型,适合于解决难以建立有效的形式化模型并且用传统技术难以有效解决甚至无法解决的问题。

在舆情网络中,事件的发生导致了情绪化意见的产生,当个体间意见一致时,便产生了意见的汇总,形成公众意见;当个体之间意见冲突时,舆情网络被分化为不同的意见中心。本文将舆情网络作为用户的观念空间,则用户可看作运动在空间中的粒子,粒子受到不同的价值观、社会观的驱动,寻找最符合自身性质的观念位置。

2.1 基本 PSO 与 SPSO 算法思想

基本 PSO 算法最初由 Kennedy 和 Eberhart 于 1995 年提出^[12]。通过对鸟群、鱼群行为的模拟,两人发现,虽然单个个体并不表现智能特征,但是整个生物种群却表现出处理复杂问题的能力。其基本思想为,假设有 N 个粒子在一个 D 维空间中运动,每个粒子 i 有一个 D 维的空间位置向量 $X_i = (x_{i,1}, x_{i,2}, \dots, x_{i,D})$ 和速度向量 $V_i = (v_{i,1}, v_{i,2}, \dots, v_{i,D})$ 。在搜索解空间时,粒子 i 根据两个元素不断地调整自己飞向全局最优解的位置;粒子本身找到的最优解 p_{best} ,记作 $P_i = (p_{i,1}, p_{i,2}, \dots, p_{i,D})$;目前种群所找到的最优解 g_{best} ,用 $P_g = (p_{g,1}, p_{g,2}, \dots, p_{g,D})$ 来表示。粒子 i 在第 t 次迭代中的速度用 $V_i^t = (v_{i,1}^t, v_{i,2}^t, \dots, v_{i,D}^t)$ 来表示,那么 i 在 $t+1$ 次迭代中的位置根据如下公式计算:

$$V_i^{(t+1)} = w \cdot V_i^{(t)} + c_1 \cdot rand() \cdot (P_i - X_i^{(t)}) + c_2 \cdot rand() \cdot (P_g - X_i^{(t)}) \quad (1)$$

$$X_i^{(t+1)} = X_i^{(t)} + V_i^{(t+1)} \quad (2)$$

式中, c_1 和 c_2 是两个正数常量,分别被称为认知学习因子和社会学习因子; $rand()$ 产生 $[0, 1]$ 之间的随机数; $w > 0$, 被称为

惯性因子。迭代终止条件根据具体问题一般选为最大迭代次数或者粒子迄今为止搜索到的最优位置满足预定最小适应阈值。

速度更新过程受到 3 个因素的影响:先前的速度,该因素阐释了粒子当前的状态;自身认知部分,代表了粒子的当前位置与其历史最好位置的距离;社会学习部分,反映了粒子所处的环境对其行为的引导作用。这 3 部分共同决定了粒子的空间搜索能力。

SPSO 算法在基本 PSO 算法的基础上增加了对多峰函数的搜索^[13],其将粒子群系统中相似的粒子并行地划分成不同的群体,一个群体就是一类具有相似性质的粒子集合。每个群体的粒子都围绕在本群体中具有最高适应值的粒子周围,该粒子被称为种群的“种子”。利用粒子之间的距离来判断它们之间的相似程度:距离越近,相似度越大;反之,距离越远,相似度越低。基于物种形成原理的 PSO 算法 (SPSO) 具有能够同时搜索到多个局部最优优点的能力。

假设两个粒子的位置分别为 $X_i = (X_{i,1}, X_{i,2}, \dots, X_{i,D})$, $X_j = (X_{j,1}, X_{j,2}, \dots, X_{j,D})$, 则两粒子之间的距离定义为:

$$d(X_i, X_j) = \sqrt{\sum_{k=1}^D (X_{i,k} - X_{j,k})^2} \quad (3)$$

根据粒子间距离的定义,进而设定子群 s 的半径 r_s ,当距离小于阈值 r_s 时,两个粒子属于一个种群;反之,它们不在同一个种群中。当一个粒子同时属于两个种群时,它将被分配到适应值更好的那个种群中。

2.2 舆情网络中的用户聚类

舆情网络中用户的观念形成过程与粒子群中鸟群的觅食行为类似。

热点事件在形成发展过程中具有以下特点:围绕特定的目标,网络被迅速建立起来;以突发事件为中心,形成一个暂时的、特殊的网络群体;生命周期比较短,随着事件的消亡,网络也会解散;典型的核心边缘结构;同一社会网络中用户的观念更加相似。

鸟群的觅食过程遵循以下规则:围绕当前的觅食目标,种群迅速形成;根据全局最优点和局部最优点的信息不断地调整自己的位置;种群的生命周期短,觅食行动结束,种群解散;典型的核心边缘结构;同一种群中的粒子行为更加相似。

从中可以看出,热点事件的形成过程与鸟群的觅食行为具有很大的相似之处,两者都会围绕特定目标迅速集聚,形成新的群体,而当目标实现之后群体迅速解散。舆情网络中,网民是推动事件传播扩散的主体,网络舆情是对网民政治态度的反映。已有的研究指出,网民对待突发事件的态度是一个由感性向理性的过渡阶段。在突发事件的初始阶段,由于信息的不对称性、事件缺乏真实信息,导致网络中的舆情态度不明显,大部分网民处于一个寻找真实答案的过程;而随着时间的推移,事件变得愈加明朗,网民的态度也随着信息透明度的增大而变得越加清晰,最终形成赞成或反对的群体。该过程与鸟群觅食的过程类似,因此,本文提出应用粒子群算法对网民对于某一事件的舆情态度变化进行模拟和仿真。

本文引入矢量空间模型^[14]的概念对用户的特征向量进行表示,用户 i 的特征向量可表示为: $U_i = (\langle keyword_1, w_1 \rangle,$

$\langle keyword_2, w_2 \rangle, \dots, \langle keyword_m, w_m \rangle$ 。其中,每一个关键字的权重通过式(4)进行计算。

$$w_i^k(T) = \frac{Num_{keyword(i)}^k(T)}{\sum_{k=1}^m Num_{keyword(k)}^k(T)} \quad (4)$$

关键字的权重表示用户对于某个事件的观念的倾向性。在事件扩散的过程中,具有相似的舆论倾向性的用户聚集到一起,抱成一团,形成用户聚类。因此,可以通过用户倾向性之间的欧氏距离来判断用户之间的相似性,如式(5)所示。相似性大的用户聚集到同一子群中,而相似性很小的用户则会出现在不同的种群里。

$$Similarity(i, j) = \sqrt{\sum_{k=1}^m (w_{i,k} - w_{j,k})^2} \quad (5)$$

2.3 构建适应度函数

在 PSO 算法中,适应度函数用于判断粒子的位置好坏,而在舆情网络中,粒子的位置标志其活跃度和影响范围。活跃度越高、影响范围越广的粒子,其位置越好。使用粒子群算法进行舆情态度的模拟仿真,首先需要解决如何设置适应度函数^[15]的问题,即如何寻找局部最优粒子。在舆情网络中,用户的活跃程度决定了用户的影响范围,因此,本文使用用户发表评论的数量和发表评论的频率来表示用户的活跃程度,即用户在舆情网络中的适应度。设定粒子群算法的迭代周期为每天一次,使用 $TC_i^u(T)$ 表示截止至周期 T 用户 i 关于事件 u 所发表的所有评论, $DC_i^u(T)$ 表示用户 i 关于事件 u 在第 T 个周期内发表的评论数,则用户 i 在事件 u 的传播过程中的适应度函数如式(6)所示。

$$F_i^u(T) = \alpha \times \frac{TC_i^u(T)}{\sum_{k \in N} TC_k^u(T)} + \beta \times \frac{DC_i^u(T)}{\sum_{k \in N} DC_k^u(T)} \quad (6)$$

$$\alpha \geq 0, \beta \geq 0, \alpha + \beta = 1$$

$$DC_i^u(t) = TC_i^u(T) - TC_i^u(T-1) \quad (7)$$

3 基于粒子群的用户聚类过程模拟与仿真

3.1 用户聚类

在事件的传播过程中,随着事态的逐渐明朗,用户对于事件的定位愈加清晰,则用户的态度愈加明确。因此,由事件引导的舆情网络会逐渐趋向于稳定。在初始的舆情网络中,网络中的初始成员提出自己对于事件的观点,并且围绕着这一观点“招募”成员,但不可避免地会面临一群行动者围绕着某一与之不同的观点而互动的情景。在舆情网络中,观念态度成为网络发展的基本组织原则,在对观点意愿的追求驱动下,当前网络往往会形成两个基于其各自目标的差异化的群体。每个群体都可能具有自身的层级结构,从而整个网络就会呈现出多个权力中心。如果继续将用户群体进行细化,则会出现更多的小的聚群现象。

3.2 用户聚类模拟与仿真

(1) 基于 PSO 的用户聚类流程

应用基本 PSO 算法对网络中用户的社会行为进行模拟,当用户的观点一致时,他们共同移动到信念空间的同一点(等价的点);当用户之间的观念不一致时,他们在信念空间中的距离会增加。

算法流程如下所示:

- ① 初始化一个规模为 n 的粒子群,设定初始位置和速度;
- ② 根据式(4)计算每个粒子的适应度值;
- ③ 根据粒子的运行情况,更新粒子本身的最优位置 p_{best} ;
- ④ 更新全局最优位置 g_{best} ;
- ⑤ 根据式(1)、式(2)分别更新粒子的速度和位置;
- ⑥ 若满足终止条件,则输出最终解,否则转至第②步。

(2) 基于 SPSO 的用户聚类流程

SPSO 算法能够将整个用户群体分为多个小的子群,用以同时发现舆情网络中的多个用户聚类中心,其算法流程如下所示。

- ① 初始化规模为 n 的粒子群,设定其初始位置和速度,并将每个粒子的个体最好位置设定为当前位置;
- ② 计算所有粒子当前位置的适应值,并且更新其个体最好位置;
- ③ 根据粒子的个体最好位置 p_{best} 进行降序排列;
- ④ 选取子粒子群的“种子”;
- ⑤ 将“种子”的个体最好位置 l_{best} 设定为其所在子粒子群中所有粒子的全局最好位置 g_{best} ;
- ⑥ 根据式(1)、式(2)分别更新粒子的速度和位置;
- ⑦ 若满足最终条件,则输出 3,否则转到步骤②。

(3) 用户聚类模拟

设置用户的特征向量长度为 20,为了简化仿真,使用数字表示每个用户的特征向量,设置空间维数为 2,即每个用户选择一个二元有序对作为自己的观念表示。首先对粒子进行初始化处理:设置种群的规模为 1000,迭代次数设置为 2000。

分别使用基本 PSO 算法和 SPSO 算法对舆情网络中的用户聚类行为进行模拟。图 1 为粒子的初始分布情况,图 2 为基于基本 PSO 算法的用户聚类模拟的输出,图 3 为基于 SPSO 算法的用户聚类模拟的输出。图中横坐标和纵坐标均表示粒子位置的横纵坐标。

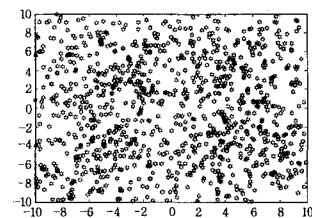


图 1 初始粒子分布图

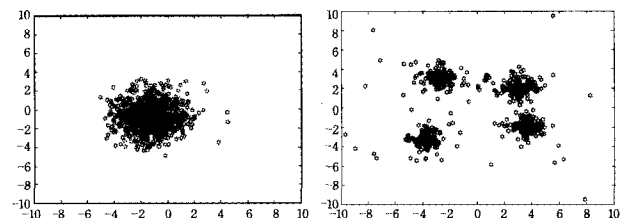


图 2 基于基本 PSO 的舆情网络用户聚类模拟结果 图 3 基于 SPSO 的舆情网络用户聚类模拟结果

从图 1 可以看出,初始阶段,粒子均匀地分布在空间中,随着事件的出现及演化,舆情网络中的用户观念逐渐极化,具有相似观念的用户聚集到一起,出现抱团现象。但是基

本 PSO 算法只能发现全局最优点,引导其他所有用户形成一个大的社团。而 SPSO 却能够发现具有不同意见倾向的用户聚类中心,形象而逼真地模拟舆情网络中用户聚类的整个过程。同时存在独立的情感分明的用户在网络上,不参与任何一个集团,游离于各个社团之外,这部分用户通常被称为“听者”。

4 自适应的预警分析策略

在粒子群算法中,通过控制粒子的运动速度可以控制群体的收敛^[16]。而在实际的舆情网络中,通过统计分析用户的观念变化对用户粒子群模型中的位置进行调整,同时根据用户的倾向性变化对事件的扩散演化模型进行分析,进而预测事件是否能够对整个社会造成影响。同时,粒子的速度也可以根据用户的观念变化动态进行设置。

若群中粒子的数量在不断增加,并且能够快速达到收敛,即舆情网络中出现“极化”现象,则事件可能引起很大的社会反响;若关于当前事件的粒子群数量一直没有显著的变化,并不存在收敛现象,则可推断当前事件并不能引起用户的普遍反映,不属于热点事件。

基于 SPSO 算法对舆情网络中的用户聚类进行模拟,同时通过跟踪子群半径和粒子相似度的阈值对舆情的发展演化过程进行判断,进而提供自适应的舆情预警分析策略。设定子群半径 $r_s=2$,粒子相似度阈值 $\epsilon=0.5$,舆情分析流程图如图 4 所示。

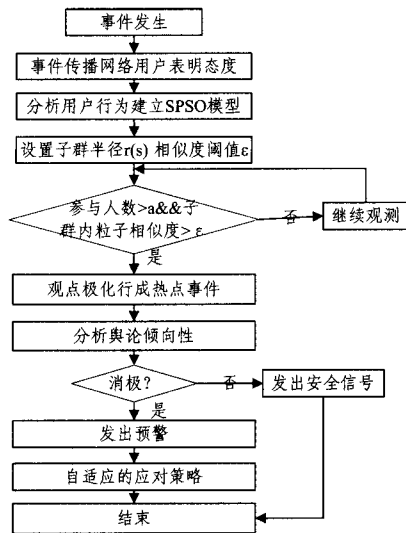


图 4 事件识别及预警分析流程

当网络上出现关于某一事件的相关报道时,系统进行数据采集,提取舆情网络中的参与用户数量及用户观念,建立 SPSO 模型;设定 SPSO 模型的子群半径和相似度阈值;当种群内粒子观念相似度达到一定程度时,形成观点极化,则热点事件形成;然后对参与用户的舆论倾向性进行分析,如果是消极的态度,则发出舆情警告,并提供自适应的舆情应对策略。

结束语 本文基于基本 PSO 算法和 SPSO 算法对舆情网络中面对突发事件时用户的聚类行为进行模拟和仿真,证明 SPSO 算法对于舆情网络中用户聚类行为具有良好的模拟

结果。在此基础上分析了用户聚类行为对事件发展的推动作用,并提出了自适应的舆情预警分析策略,为构建合理的舆情网络中突发事件的识别及预警机制提供了有效的理论依据。

参考文献

- [1] 麦林. 虚拟社区热点话题意见挖掘模型研究[M]. 合肥: 中国科学技术大学, 2009
- [2] 方付建. 突发事件网络舆情演变研究[D]. 武汉: 华中科技大学, 2011
- [3] Cover T, Hart P. Nearest Neighbor pattern classification [J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27
- [4] Cortes C, Vapnik V. Support vector networks [J]. Machine Learning, 1995, 20: 273-297
- [5] Hu Xiao-hui, Eberhart R C, S Yu-hui. Engineering optimization with particle swarm[C]//Proceedings of the 2003 IEEE Swarm Intelligence Symposium, 2003: 53-57
- [6] Li X. Adaptively choosing neighborhood bests using species in a particle swarm optimizer for multimodal function optimization [C]//Del K, et al., eds. Proc. Genetic Evol. Comput. Conf. 2004: 105-116
- [7] Vlachogiannis J, Lee K. Determining generator contributions to transmission system using parallel vector evaluated particle swarm optimization[J]. IEEE Transactions on Power Systems, 2005, 20: 1765-1774
- [8] Yin P Y. A discrete particle swarm algorithm for optimal polygonal approximation of digital curves[J]. Journal of Visual Communication and Image Representation, 2004, 15(2): 241-260
- [9] Venayagamoorthy G K, Zha W. Comparison of nonuniform optimal quantizer designs for speech coding with adaptive critics and particle swarm [J]. IEEE Transactions on Industry Applications, 2007, 43(1): 238-244
- [10] 张青, 康立夫, 李大农. 群智能算法及其应用[J]. 黄冈师范学院学报, 2008, 28(6): 44-48
- [11] 胡玉成. 面向动态环境的粒子群算法研究[D]. 武汉: 华中科技大学, 2010
- [12] Colnari A, Dorigo M, Maniezzo V. An investigation of some properties of an ant algorithm [C]//Proc. of Parallel Problem Solving from Nature (PPSN). France: Elsevier, 1992: 509-520
- [13] Eberhart R C, Kennedy J. A new optimizer using particle swarm theory [C]//Proceedings of the Sixth International Symposium on Micromachine and Human Science, Nagoya, Japan, 1995: 39-43
- [14] Gevrey J, Ruder S. Link-based approaches for text retrieval [C]//Proceeding of TREC-10, NIST. NIST Special Publication, 2002
- [15] Shan S M, Deng G S. Improved adaptive particle swarm optimizer in dynamic environment [J]. Systems Engineering-Theory & Practice, 2006, 26(3): 39-44
- [16] Parsopoulos K E, Vrahatis M N. Particle Swarm Optimization Method for Constrained Optimization Problems [C]//Proceeding of the Euro-International Symposium on Computational Intelligence 2002. Intelligent Technologies - Theory and Applications: New Trends in Intelligent Technologies, 2002: 214-220