doi: 10.3969/j.issn.1674-8425(z).2019.09.015

本文引用格式: 韩楠,乔少杰,黄萍,等. 基于群体智能的跨语言网络舆情文本聚类模型[J]. 重庆理工大学学报(自然科学),2019,33(9):99-108.

Citation format: HAN Nan, QIAO Shaojie, HUANG Ping, et al. Multi-Language Text Clustering Model for Internet Public Opinion Based on Swarm Intelligence[J]. Journal of Chongqing University of Technology(Natural Science),2019,33(9):99-108.

基于群体智能的跨语言网络舆情文本聚类模型

韩 楠[1a],乔少杰[1b],黄 萍[1a],彭 京[2],周 凯[2]

(1. 成都信息工程大学 a. 管理学院; b. 软件工程学院,成都 610225;
2. 四川省公安厅,成都 610014)

摘 要: 跨语言的互联网文本信息在中国多个民族构成中非常普遍,但当前文本聚类模型主要针对单一语言,跨语言文本挖掘的研究较少。群体智能算法具有自组织、启发式、自适应和鲁棒性的特点,提出一种基于群体智能的跨语言网络舆情文本聚类模型 SI-Cluster(swarm-intelligence-based text clustering model),应用3种优化策略。梯度下降法弱化智能体拾取文本的能力,避免陷入局部最优解,添加信息素引导智能体移动并有效避免信息素挥发过快的问题,智能体从当前位置选择下一位置考虑信息素感应浓度和方向权重因子。在中文、英文和藏文文本数据集上进行实验,从聚类准确性上看应用优化策略的 SI*-Cluster 算法的 F-measure 值达到0.862,相比于 k-means 算法提高44.1%;从收敛性上看 SI*-Cluster 算法在聚类效果明显的前提下迭代500次收敛,相比 SI-Cluster 算法900次收敛,具有更快的收敛速度。模拟展示了 SI-Cluster 和 SI*-Cluster 算法进行文本聚类的迭代过程,证明所提优化策略的有效性。

关 键 词: 群体智能; 跨语言; 文本聚类; 网络舆情

中图分类号: TP311 **文献标识码**: A **文章编号**: 1674-8425(2019)09-0099-10

Multi-Language Text Clustering Model for Internet Public Opinion Based on Swarm Intelligence

HAN Nan[1a], QIAO Shaojie[1b], HUANG Ping[1a], PENG Jing[2], ZHOU Kai[2]

(1. a. School of Management; b. School of Software Engineering,
Chengdu University of Information Technology, Chengdu 610225, China;
2. Sichuan Provincial Department of Public Security, Chengdu 610014, China)

收稿日期: 2019-02-12

基金项目: 国家自然科学基金资助项目(61802035,61772091,61962006);四川省科技计划项目(2019YFG0106,2018JY0448,2019YFS0067);四川高校科研创新团队建设计划(18TD0027);成都市软科学研究项目(2017-RK00-00053-ZF);广西自然科学基金项目(2018GXNSFDA138005);成都信息工程大学中青年学术带头人科研基金项目(J201701);成都信息工程大学科研基金项目(KYTZ201715,KYTZ201750)

作者简介: 韩楠,女,博士,副教授,主要从事网络舆情分析研究;通讯作者 乔少杰,男,博士,教授,主要从事数据库、人工智能研究,E-mail: sjqiao@cuit.edu.cn。

Abstract: Multi-language text from the Internet is ubiquitous in China which is a very huge country composed of many nationalities. Existing text clustering models is mainly applied for one single language, and there are few studies on multi-language text mining. Swarm intelligence algorithms have the characteristics of self-organizing, heuristic, adaptive and robust. A multi-language text clustering model for Internet public opinion based on swarm intelligence is proposed, which is called SI-Cluster (swarm-intelligence-based text clustering model). Three optimization strategies are applied: a gradient descent method is applied to degrade agents' capability of picking up texts in order to avoid falling into the local optimal solution; the pheromone is used to guide agents to move, which can effectively avoid the problem of excessive volatilization of pheromones; the agent selects the next position from the current position by taking into consideration the pheromone concentration of sensing and the weight factor of directions. Experiments were conducted on Chinese, English and Tibetan text datasets. In terms of clustering accuracy, the F-measure of the improved SI*-Cluster algorithm with optimization strategies can reach to 0.862, which is 44.1% higher than that of the k-means algorithm. In terms of convergence, SI*-Cluster can converge after 500 times of iterations with obviously good clustering results, which is faster than that of the SI-Cluster algorithm converging after 900 times of iterations. Simulation shows the iterative process of SI-Cluster and SI*-Cluster for text clustering, and the results prove the effectiveness of the proposed optimization strategies.

Key words: swarm intelligence; multi-language; text clustering; Internet public opinion; optimization

网络舆情分析需要对网络社会中群体所表达的情绪、态度、信念等信息进行实时准确的分析和挖掘,以掌握社情民意,应对网络突发的公共事件等[1]。随着互联网的飞速发展,跨语言(多语言)互联网文本信息在我国非常普遍,如藏族同胞经常采用藏文和中文混合使用的方式在网络上发表自己的观点。然而,目前的舆情分析系统主要针对单一语言进行挖掘和预测,跨语言的网络舆情分析的研究成果在国内外鲜有报道,国内目前没有同类技术和相关应用软件出现。因此,研发跨语言的网络舆情分析系统对于分析处理当前快速增长的大规模网络信息具有重要的科学价值及应用意义,也是迫切需要投入人力和物力开发的。

跨语言舆情分析系统旨在满足客户检测网络舆情动态并获得敏感信息的需求,主要功能包括:提取半结构化和无结构化网页、博客、论坛等文本中的主题信息,存储到数据库中;实现跨语言文本信息的预处理,包括分词、提取特征词、建立存储结构等;对提取的半结构化信息进行聚类或分类处理;分析海量的数据中发现主题、检测热点、追踪专题等。

分词是网络舆情分析系统的关键环节,在数据存储和数据分析中起着承前启后的作用。近年来,分词方法的研究备受学者关注,出现了多种具有应用前景的分词方法,其中基于字典的分词方法[2]应用十分广泛,尤其适用于多语言环境下的文本挖掘研究。基于字典的分词具有3个要素:分词字典、文本扫描顺序和匹配原则。这3个要素互相组合生成了许多种分词方法,包括正向最大匹配法、逆向最大匹配法、双向扫描法、逐词遍历法、最佳匹配法等。综合上述方法的优缺点及通用性,本文采用的是正向最大匹配法对爬取的半结构和无结构的文本进行分词,然后进行文本聚类。

文本聚类分析是网络舆情分析的关键步骤,旨在将散乱的抽象对象的集合划分为多个分组,生成的每个分组由相似的对象组成。本文研究的目的在于设计一种高效、准确、支持多语言环境的网络舆情文本聚类模型,为下一步的舆情分析做准备。

为了实现大规模半结构和无结构文本数据的聚类分析,各种聚类算法被广泛提出。随着大数

据时代的来临,聚类算法不但要考虑结果的准确性,更重要的是算法如何更加高效地适应实时更新的网络环境。简单传统的聚类算法,如 k-means[3]、k-mediods 聚类算法,在局部环境的聚类中表现出较好的聚类效果,但运用于复杂和大规模文本数据中效率低下[4]。基于群体智能的聚类算法以其自身的自适、启发等特点,具有分布式处理大数据的优势[5]。

群体智能(swarm intelligence,SI)[6]作为一个新兴的研究领域,最早来自于对自然界中昆虫群居生活的观察和模拟。群居性生物就个体而言并不具有较高的智能,而是通过种群个体之间相互协作,共同完成复杂的工作,群体智能算法成功地解决了组合优化、计算机、机器人、电力系统等众多领域问题。目前,常见的群体智能算法有遗传算法(genetic algorithm,GA)[7]、蚁群算法(ant colony optimization,ACO)[8]、细菌觅食算法(bacterial foraging algorithm,BFA)[9]、人工鱼群算法(artificial fish swarm algorithm,AFSA)[10]、粒子群算法(particle swarm optimization,PSO)[11]、人工蜂群算法(artificial bee colony,ABC)[12]等。

本文借鉴蚁群算法的工作原理,生成智能体并将其随机放置于二维文本空间中,根据信息素的存量指定移动朝向进而实现文本聚类。此外,提出了多种优化策略,改进群体智能算法运行的网格环境和智能体(本文称为 agent)。所提的新型群体智能算法具备启发式和自适性的特点,可以实现大规模跨语言文本聚类。

1　基本概念

本节首先给出模型中使用的重要定义,并给出基于群体智能文本聚类的概念。

定义 1　文本向量。文本被表示成一个多维向量,向量的每一维对应文本中的一个特征词,向量每一维的取值对应特征词在文本中的权重值,即该特征词在文本中的重要程度。一个文本 d 可以转换成一个多维文本向量,如式(1)所示。

$$d = \langle (t_1, w_1), (t_2, w_2), \cdots, (t_i, w_i), \cdots, (t_n, w_n) \rangle \tag{1}$$

其中: t_i 表示向量空间中的特征词; w_i 表示在文档 d 中 t_i 对应的权重。将所有的文本文档转化为文本向量,所有的文本向量构成一个特征向量空间。

问题描述: 文本聚类定义为对给定的文本集合 $D = \{d_1, d_2, \cdots, d_n\}$,通过求取目标函数的最优值得到类簇集合 $C = \{c_1, c_2, \cdots, c_k\}$, $\sum_{j=1}^{k} c_j = D(i = 1, 2, \cdots, k)$,其中,$n$ 表示文本数量,k 为聚类后簇的数量,$c_p \cap c_q = \varnothing, p \neq q$。对于每个 $d_i(d_i \in D)$,$c_j(c_j \in C)$,$d_i \in c_j$,使得目标函数 $Q(C)$ 达到最优值。目标函数的度量指标很多,常用的如均方根误差的和。

定义 2　基于群体智能的文本聚类。假设具有 n 维向量文本数据集的 m 个对象 $\{t_1, t_2, \cdots, t_m\}$ 以及 L 个智能体,基于群体智能文本聚类的基本思想是: L 个智能体根据簇内相似度高、簇间相似度低的原则,将这 m 个对象划分到 k 个簇中,使得评价函数值最小。

利用智能体对文本进行聚类,生成智能体并将其随机放置于二维文本空间中,根据信息素的存量指定移动朝向,若当前网格的信息素存量相同,则随机指定移动朝向。这里给出文本拾起和放下的概念[5]。

定义 3　文本拾起概率。智能体利用式(2)计算拾起所在区域文本对象的概率

$$p_{\text{pick}}(t_i) = \left(\frac{k_p}{k_p + f(t_i)}\right)^2 \tag{2}$$

其中: k_p 表示智能体拾起文本对象的概率值; $f(t_i)$ 表示文本对象 t_i 与邻近文本相似度的平均值。

定义 4　文本放下概率。智能体利用式(2)计算放下所在区域文本对象的概率:

$$p_{\text{drop}}(t_i) = \left(\frac{f(t_i)}{k_d + f(t_i)}\right)^2 \tag{3}$$

其中: k_d 表示智能体放下样本对象的概率; $f(t_i)$ 表示文本对象 t_i 与邻近文本相似度的平均值。

定义 5　文本相似度。文本相似度定义为

$$f(t_i) = \frac{1}{s^2}\sum_{t_j \in \mathrm{Neigh}_{(s\times s)}(t_i)}\left[1 - \frac{d(t_i,t_j)}{\alpha}\right] \quad (4)$$

其中: s^2 **代表局部环境范围;** $d(t_i,t_j)$ **表示文本对象** t_i **和** t_j**之间的距离;** α **因子用来控制环境相似度的取值范围,决定了相邻对象之间的辨识度。** α **取值过小会导致簇过于紧凑且空间中出现多个小簇。**

为了实现跨语言文本聚类,需要将分词后文本进行特征词提取生成多维文本向量,然后计算所有文本向量之间的相似度生成文本相似度矩阵,计算文本向量相似度利用夹角余弦公式。本文将式(4)修改为

$$f(t_i) = \frac{1}{s^2}\sum_{t_j \in \mathrm{Neigh}_{(s\times s)}(t_i)}\left[\frac{\mathrm{sim}(t_i,t_j)}{\alpha}\right] \quad (5)$$

将 $1 - d(t_i, t_j)/\alpha$ **修改成为** $\mathrm{sim}(t_i, t_j)/\alpha$**。标准的群体智能聚类算法是根据对象之间的距离算出相似度,修改后的公式适应于文本聚类过程中文本相似度的计算。通过式(5)可以发现,文本向量** t_i **与相邻的文本向量** t_j **之间的相似度越高,说明** t_i **的局部相似度越高。**

2　基于群体智能的文本聚类算法

基于群体智能的跨语言网络舆情文本聚类模型的工作原理如图 1 所示: ① 参数初始化,主要包括文本空间网格大小、信息素、拾起和放下文本的概率值 k_p **和** k_d **等参数的设置。② 计算智能体所在网格内文本对象相对于其他文本对象的相似度,将其作为文本拾起的概率,当智能体拾起文本的概率大于随机概率阈值,则拾起该文本。如果智能体本身拾有文本,计算当前文本与其所在网格相邻的 8 个不同区域,即: 上、下、左、右、左上、左下、右上、右下中文本的相似性,作为文本对象放下的概率值; 如果其值大于随机概率值,智能体便选择放下这一文本。大量的智能体在网格中向不同区域移动,重复上述步骤最终使得评估函数代价最小,完成文本向量聚类的任务。**

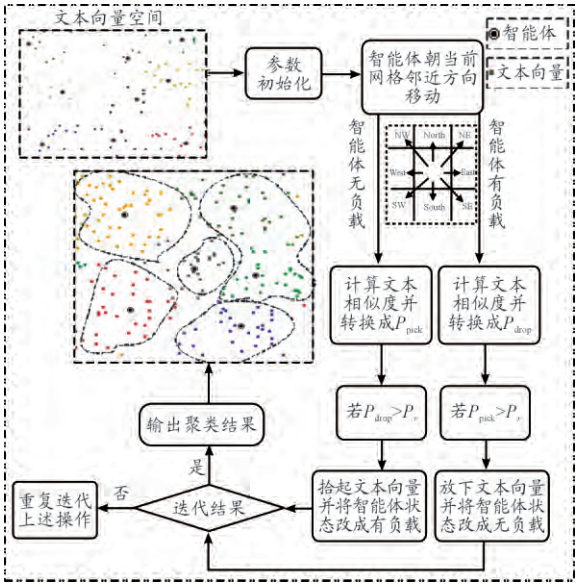

图 1　基于群体智能的文本聚类算法工作原理

算法 1　基于群体智能的跨语言文本聚类算法

输入: 多语言文本特征向量集合 T

输出: 文本类簇集合 $C = \{c_1, c_2, \cdots, c_p\}$

```
initialTex();   //将文本向量随机分布到网格中
initialAgent();   //智能体及网格环境初始化
for i = 1 to N do   //迭代 N 次
    for j←—agentNum do   //遍历每个智能体
        r = agent. getLocation();   //r 获取智能体所在位置
        for t_k ∈ T
            if(a_j is unloaded && r is occupied by t_k)
            //智能体 a_j 无负载且 r 处有文本对象 t_k
                compute f(t_k) and p_pick(t_k);
            //计算群体相似度 f(t_k) 和拾起概率 p_pick(t_k)
                if(p_pick(t_k) > p_r)   //p_r 表示随机概率值
                    a_j picks up t_k; //智能体拾起文本对象
                else if(a_j has t_k and r is empty)
                //智能体负载对象 t_k 且 r 处为空
                    compute f(t_k) and p_drop(t_k);
                //计算群体相似度和放下概率 p_drop(t_k)
                    if(p_drop(t_k) > p_r)  //p_r 表示随机概率值
                        a_j drops off t_k; //放下文本对象
                    end if
            end for
        a_j. randomMove();   //智能体选择随机移动
    end for
end for
output C = {c_1, c_2, ···, c_p};   //输出聚类结果
```

算法复杂性分析: 通过分析可以得到算法的

时间复杂度为 $O(A \times N \times M)$,其中: A 表示智能体的数量; N 表示算法的迭代次数; M 表示文本的数量。

3　优化策略

标准的基于群体智能的文本聚类算法主要存在如下几个方面的不足:

1) 算法 1 是基于网格空间的聚类算法,文本被随机放置于网格空间中,算法经若干次迭代后,文本向量在网格空间中的分布呈现出聚类结构,即相似的文本向量对象被聚到一起。但是网格空间中的聚类结构并不稳定,智能体在移动的过程中不断形成新的聚类结构,同时也在不停地破坏智能群体的分布。

2) 在标准的群体智能本文聚类算法中引入信息素机制,借助信息素引导和控制智能体的移动,可以加快聚类收敛,极大地提高算法的收敛速度,但同时也带来了局部对象过多、智能体容易陷入局部最优等问题。智能体会被引导到信息素浓度较高的区域,由于智能体与环境的正反馈作用,局部的文本向量集中于该区域。

3) 利用群体智能实现文本聚类的优越性表现在自组织、启发式和鲁棒性等方面,聚类结果需要从网格空间结构中迭代获取,网格中的聚类效果越明显,越容易获取较好的结果。

优化标准群体智能算法,需要改进算法运行的网格环境和智能体,保证群体智能算法具备启发式和自适应性的特点。本文提出如下优化策略:

策略 1　在算法运行过程中,逐渐弱化智能体拾取文本对象的能力,以达到网格中聚类结构趋于稳定的目的。具体做法是在迭代的过程中利用梯度下降法逐步降低 k_p 值,使得智能体拾取概率 p_{pick} 降低,当 k_p 缩小到智能体不再具有拾取对象的能力时,网格中的聚类结构趋于稳定,k_p 值也无需进一步降低。

策略 2　添加信息素引导智能体的移动,信息素能够提高智能体与环境的交互能力[13],基本思路为: 在智能体搬运和搜索食物时,每个智能体会在所经过路径上留下一定量的信息素。智能体在选择下一步之前会感应周围环境中的信息素以指导下一步走向,同时信息素以一定速度挥发,最终的结果是环境中形成一条由高浓度信息素构成的路径,最终找到最优路径。智能体在网格环境中移动,会在网格环境中留下大小为 σ 的信息素,信息素会随着时间的流逝而挥发,挥发因子为 η。算法每经过一次迭代,网格环境中的信息素都会按照挥发因子消失一部分。第 $k+1$ 次迭代后,网格环境中某个位置的信息素由 σ_k 变为 σ_{k+1},计算方法如式(6)所示。

$$\sigma_{k+1} = (1-\eta) * \sigma_k \tag{6}$$

式中 η 的取值范围为 $[0,1]$。

针对信息素挥发过快的问题,本文对式(6)进行改进,智能体在经过某一位置时,该位置的信息素 σ_l 变为 σ_{l+1},计算方法如式(7)所示。

$$\sigma_{l+1} = \sigma_l + (\eta + N/A) \tag{7}$$

式中: N 表示算法执行的次数; A 表示智能体数量。式(7)中增加了执行次数和智能体数量的比值,可以有效地避免信息素挥发过快的问题。

策略 3　智能体根据当前位置选择下一位置需要考虑两个因素: 信息素感应浓度和方向权重因子。

智能体移动的步长均为 1,因此智能体的下一步移动方向为 $(1,\Delta\theta)$,其向各个方位移动的概率是不同的,每个方位的权重值都是由智能体当前移动方向决定的,上一步的位置为 (x^*,y^*),经过移动方向 $(1,\Delta\theta)$,移动到了 (x,y),可以利用公式(8)计算得到智能体在 (x,y) 处的移动方向状态 θ 值。

$$\begin{cases} \cos(\theta) = \dfrac{x-x^*}{r} \\ \sin(\theta) = \dfrac{y-y^*}{r} \\ r = \sqrt{(x-x^*)^2+(y-y^*)^2} \end{cases} \tag{8}$$

智能体在网格环境中移动时,共有 8 个方向,

分别与当前移动方向形成 8 个不同角度 $\Delta\theta$。相对于智能体当前方向，每个移动方向的权重因子 $\bar{\omega}(\Delta\theta)$ 各不相同。每个方向的权重因子 $\bar{\omega}(\Delta\theta)$ 各不相同。本文采用的权重因子为：$(0°)=1$，$(45°)=1/2$，$(90°)=1/4$，$(135°)=1/8$，$(180°)=1/12$，$(225°)=1/8$，$(270°)=1/4$，$(315°)=1/2$。

网格环境中信息素响应浓度利用如下公式求取：

$$W(\sigma)=\left(1+\frac{\sigma}{1+\gamma\sigma}\right)^{\beta} \tag{9}$$

其中参数 β 表示对信息素刺激的敏感程度，用来决定响应函数 $W(\sigma)$。β 控制随机移动的程度，依据这个参数，智能体沿着信息素梯度方向移动。低于 β 值的信息素浓度并不会很大程度地影响智能体的移动概率，而高于 β 值的信息素浓度会使智能体更确定地沿着信息素浓度的梯度方向移动。$1/\gamma$ 描述了智能体对信息素的感知能力，即在高浓度情况下对信息素减少的感知能力。参数 β 和 $1/\gamma$ 共同决定了信息素的痕迹。因此，信息素浓度的响应函数由 σ、β 和 $1/\gamma$ 三个参数共同决定。

因此，感应信息素浓度和方向权重因子决定了智能体的下一步移动方向。智能体移动下一步选择临近网格信息素为 σ 的概率 p 可用式(10)计算。

$$p=\frac{W(\sigma_i)\,\bar{\omega}(\Delta\theta_i)}{\sum_{j/k}W(\sigma_j)\,\bar{\omega}(\Delta\theta_j)} \tag{10}$$

式中：j/k 表明了网格 k 周围 8 个相邻网格空间；$\bar{\omega}(\Delta\theta)$ 表示智能体移动到不同方向的权重因子。

通过上述分析可以发现：实现基于信息素的群体智能文本聚类算法，仅需要将算法 1 中第 17 行的智能体随机移动方法改为基于式(6)～(10)的感应信息素移动方法，进而实现对本文所提基于群体智能的文本聚类算法的优化。

4 实验及算法性能分析

4.1 实验数据集及参数设置

实验数据为利用笔者开发的跨语言网络爬取器从中文、英文及藏文门户网站抓取的半结构化和无结构化的文本，并保存到数据库中。对存入数据库的文本信息利用正向最大匹配法进行分词，去噪和提取关键字等操作后得到初始数据集。藏文数据如图 2 所示。

id	title	content	url	words	num	date
1	ཅི་བྱེད་པ་ · 中国藏	རྒྱ19ཀྱི་བྱིན་ཐུང་	http://ti.tibet3.com/	འབྲི་རོགས་རིན་ཆེན་སྐྱིད་	19,14,12,11,9,8	2012-05-19
3	ཀུན་གསོད་བྱིད་པ་ལས་	ཀུན་གསོད་བྱིད་པའི་འཕྲ	http://ti.tibet3.com/n	ཞེས་འཚུག་བྱིས་པའམ་	9,9,9,9,8,7	2012-05-21
4	ཀུན་གསོད་བྱིད་པ་ཕྱུག	ཁ་བའི་རོགས་བཙན་བྱུང་ཞེས	http://ti.tibet3.com/n	བྱིས་གཤན་ཤེད་བྱང་རོགས་བོང་	21,17,14,11,11,10	2012-05-21
5	ཀུན་གསོད་བྱིད་པ་ཕྱུག	ཁ་བའི་རོགས་བཙན་བྱུང་ཞེས	http://ti.tibet3.com/n	བྱིས་གཤན་ཤེད་བྱང་རོགས་བོང་	21,17,14,11,11,10	2012-05-21
6	ཀུན་གསོད་བྱིད་པ་ཅི་ཞེ	རོགས19ཀྱི་ ཀུན་བྱུང་བྱང་	http://ti.tibet3.com/n	འཕྲོ་བྱིས་ཤེས་ཞེན་ཤེན་འཕྲོ	24,15,13,12,10,10	2012-05-21
7	ཀུན་གསོད་བྱིད་པ་ཕྱུག	རྒྱལ་པོའི་རོགས་བཙན་བྱུང་	http://ti.tibet3.com/n	བྱིས་པོའི་ཤུར་རྒྱུ་རིགས་ཀྱོང	20,14,14,14,12,12	2012-05-21
8	ཀུན་གསོད་བྱིད་པ་ཁང་	རྒྱལ་པོའི་རོགས18ཀྱི་བྱིན	http://ti.tibet3.com/n	བྱིས་པོའི་ཤུར་རྒྱུ་རིགས་ཀྱོང	73,69,54,44,43,40	2012-05-20
9	ཀུན་གསོད་བྱིད་པ་བང་	རྒྱལ་པོའི་རོགས18ཀྱི་བྱིན	http://ti.tibet3.com/n	བྱིས་བྱིན་ཁང་ཚ་ཤས་	14,8,8,7,7,5	2012-05-20
10	ཀུན་གསོད་བྱིད་པ་རྒྱུས	ཀུན་གསོད་བྱིད་པའི་འཕྲ	http://ti.tibet3.com/n	ཞེས་རོགས་ཤུར་ཚ་ཤས་	4,4,3,3,2,2	2012-05-19
11	ཀུན་གསོད་བྱིད་པ་ཕྱུག	ཁ་བའི་རོགས་བཙན་བྱུང་ཞེས	http://ti.tibet3.com/n	བྱིས་གཤན་ཤེད་བྱང་རོགས་བོང་	21,17,14,11,11,10	2012-05-21
12	ཀུན་གསོད་བྱིད་པ་ཅི་ཞེ	རོགས19ཀྱི་ ཀུན་བྱུང་བྱང་	http://ti.tibet3.com/n	འཕྲོ་བྱིས་ཤེས་ཞེན་ཤེན་འཕྲོ	24,15,13,12,10,10	2012-05-21
13	ཀུན་གསོད་བྱིད་པ་ཕྱུག	རྒྱལ་པོའི་རོགས་བཙན་བྱུང་	http://ti.tibet3.com/n	བྱིས་པོའི་ཤུར་རྒྱུ་རིགས་ཀྱོང	20,14,14,14,12,12	2012-05-21
14	ཀུན་གསོད་བྱིད་པ་ཁང་	རྒྱལ་པོའི་རོགས18ཀྱི་བྱིན	http://ti.tibet3.com/n	བྱིས་པོའི་ཤུར་རྒྱུ་རིགས་ཀྱོང	73,69,54,44,43,40	2012-05-20
15	ཀུན་གསོད་བྱིད་པ་བང་	རྒྱལ་པོའི་རོགས18ཀྱི་བྱིན	http://ti.tibet3.com/n	བྱིས་བྱིན་ཁང་ཚ་ཤས་	14,8,8,7,7,5	2012-05-20
16	ཀུན་གསོད་བྱིད་པ་ཚང་	རོགས19ཀྱི་ཚའི་སྐྱང་བང་	http://ti.tibet3.com/n	རྒྱལ་ཚན་ཚའི་སྐྱིད་ཁལ་ཁ་	16,15,11,8,8,7	2012-05-21
17	ཀུན་གསོད་བྱིད་པ་ཚང་	རོགས19ཀྱི་ཚའི་སྐྱང་བང་	http://ti.tibet3.com/n	རྒྱལ་ཚན་ཚའི་སྐྱིད་ཁལ་ཁ་	16,15,11,8,8,7	2012-05-21

图 2 藏文数据表结构示例

图 2 为输入数据表，包括 7 个属性：id，title，content，url，words，num，date。其中：id 表示文本编号；title 是文本主题；content 是文本内容；url 是文本的爬取链接；words 是经过分词和提取关键字后的关键字序列；num 是对应关键字在文本中出现的频数。本文进行跨语言(如藏文和中英文)文本聚类时，主要利用了图 2 中 title、content 和 words 三个涉及文本内容的属性。

本文的实验环境为：奔腾双核 2.0 GHz CPU，2 GB内存，开发语言为 Java，使用 Eclipse SDK1.6。

实验中分别使用 k-means、标准群体智能文本聚类算法(swarm-intelligence-based text clustering model，简称 SI-Cluster)和应用第 3 节介绍的优化策略的群体智能文本聚类算法 SI*-Cluster 算法对多语言文本数据集进行聚类。通过比较聚类算法的准确性和收敛速度，证明本文所提算法的性能优势。

SI-Cluster 算法参数设置如表 1 所示。在如表 1 所示参数设置下，群体智能算法性能较好。实验中当文本数量增加时，仅需改变网格空间大小和智能体数量即可，无需进行其他人为设置。

表 1　基于群体智能的文本聚类算法参数设置

参数及说明	初始值
k_p: 影响拾起概率的参数	0.1
$k_p^{\min}$: k_p 的最小值	0.001
k_d: 影响放下概率的参数	0.06
α: 控制环境相似度的因子	0.7
η: 环境信息素挥发控制参数	0.05
γ: 影响感应信息素的参数	2.0
β: 对信息素刺激的敏感程度	9
网格空间	45 × 45
A: 智能体数量	18

4.2　聚类准确性对比

实验中对聚类的有效性利用查准率 P、查全率 R 和 F-measure 值[14] 进行评价,定义如下:

已知 T_P 表示利用聚类算法正确聚类出的文本; F_P 为不在本簇被错误聚类出的文本; F_N 表示包含在本簇但未被聚类到这个簇中的文本; T_N 表示不在本簇且未被聚类到本簇的文本,于是:

$$P = T_P/(T_P + F_P) \tag{11}$$

$$R = T_P/(T_P + F_N) \tag{12}$$

$$F\text{-measure} = 2PR/(P + R) \tag{13}$$

实验选用 1 200 条数据作为基本测试数据集。这些数据集主要从新浪中英文网站中的 3 个栏目中爬取的,即体育(sports)、新闻(news)和视频(video),并自动加上标签用于聚类准确性的判别。对比 SI-Cluster 和 SI*-Cluster 算法,聚类准确性结果如表 2 ~ 4 所示。

表 2　k-means 算法聚类准确性

类别	文本数量	P	R	F-measure
体育(sports)	375	0.534	0.641	0.583
新闻(news)	397	0.682	0.508	0.582
视频(video)	428	0.579	0.685	0.628
合计	1 200	F-measure 平均值		0.598

表 3　SI-Cluster 算法聚类准确性

类别	文本数量	P	R	F-measure
体育(sports)	375	0.676	0.862	0.758
新闻(news)	397	0.800	0.692	0.742
视频(video)	428	0.737	0.823	0.778
合计	1 200	F-measure 平均值		0.759

表 4　SI*-Cluster 算法聚类准确性

类别	文本数量	P	R	F-measure
体育(sports)	375	0.810	0.931	0.866
新闻(news)	397	0.947	0.692	0.800
视频(video)	428	0.918	0.923	0.920
合计	1 200	F-measure 平均值		0.862

通过表 2 ~ 4 可以发现: 应用优化策略的群体智能文本聚类算法 SI*-Cluster 的 F-measure 平均值相比于标准算法 SI-Cluster 提高 13.6%,相比于 k-means 算法提高 44.1%,证明了本文所提出应用优化策略的基于群体智能的文本聚类算法的准确率较传统算法有显著提升,主要原因在于:

1) 传统 k-means 算法需要根据初始聚类中心来确定一个初始划分,然后对初始划分进行优化。这个初始聚类中心的选择对聚类结果有较大的影响,一旦初始值选择不合适,可能无法得到有效的聚类结果。此外,k-means 算法中 k 是事先给定的,k 值的选定非常难以确定。而本文提出的基于群体智能的文本聚类算法,一旦初始参数确定,算法可以自适应训练,通过一定的改进和优化,可以保证找到最优聚类结果。

2) 改进的群体智能文本聚类算法 SI*-Cluster 明显优于标准的 SI-Cluster 算法,原因在于 SI*-Cluster 算法通过改进网格环境和智能体,保证了群体智能算法具有启发式和自适应性的特点。在算法运行过程中,逐渐弱化智能体拾取文本对象的能力,以达到网格中聚类结构趋于稳定的目的。此外,添加信息素引导智能体的移动,信息素能够提高智能体与环境的交互能力。

为了进一步验证本文所提算法可以处理多语言文本，如藏文，实验中对藏文文本数据集进行实验，验证 k-means、SI-Cluster 和 SI*-Cluster 算法的聚类准确性。实验文本数量分别选取 150、300、450、600、750、800、950、1 100。由于采集的数据集分类不均衡，因此利用 F-measure 方法来评价聚类效果，结果如图 3 所示。

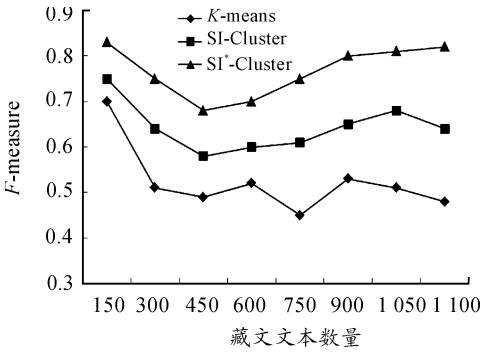

图 3 藏文文本聚类准确性 F-measure 值对比

通过图 3 可以发现：对于藏文文本，本文提出的群体智能文本聚类算法明显优于 k-means 算法。从 F-measure 值的变化可以看出，3 种算法在文本数量较少的情况下，聚类效果都比较好。随着文本数量增加到 450，算法的 F-measure 值不断减小。然而当文本数量增加到一定的程度（当文本数量为 600 时），SI-Cluster 算法和 SI*-Cluster 算法的聚类准确性不断提升，说明基于群体智能的文本聚类算法不会受到文本数量的影响，具有比较高的稳定性和自适应性。

4.3 聚类运行收敛性及运行时间对比

本节比较 SI-Cluster 和 SI*-Cluster 的算法收敛性并对比收敛时的运行时间，进而说明本文所提出应用优化策略的群体智能文本聚类算法的性能优势。实验中选用 4.2 节中 1 200 条中英文半结构化文本数据，为了保证实验结果的客观性和可比性，每组实验对 1 200 条文本数据集训练 10 次取 F-measure 的平均值。实验参数设置如表 1 所示，SI*-Cluster 算法应用第 3 节优化策略 1 将 k_p 值设置为 0.01 ~ 0.001 变化，下降梯度设置为 0.000 1，实验结果如图 4 所示。

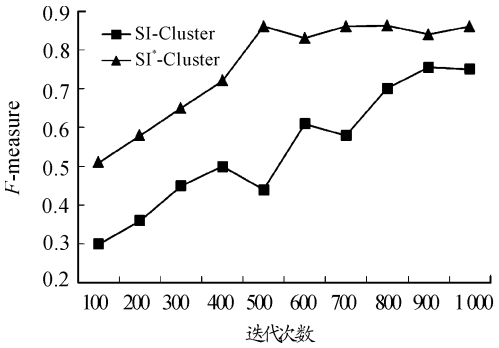

图 4 基于群体智能的本文聚类算法收敛性对比

表 5 基于群体智能的文本聚类算法收敛次数及时间

算法	执行次数	F-measure 值	运行时间/ms
SI-Cluster	900	0.755	208
SI*-Cluster	500	0.862	71

通过图 4 和表 5 可以发现：

1）基于群体智能的文本聚类算法会在运行若干步之后达到收敛，即 F-measure 值达到一定的水平。SI-Cluster 算法随着迭代次数的增加，F-measure 值的变化波动较大，而 SI*-Cluster 算法的 F-measure 曲线相对平滑，而且稳定时的 F-measure 值明显高于 SI-Cluster 算法。主要原因在于 SI*-Cluster 算法应用了第 3 节优化策略 3，智能体移动时根据当前位置选择下一位置结合了信息素感应浓度和方向权重因子，智能体稳定地沿着信息素梯度增加的方向移动，避免了陷入局部最优的问题，具有较高的稳定性。

2）SI*-Cluster 算法的 F-measure 曲线能够平滑且较快地收敛并达到稳定值。原因在于 SI*-Cluster 算法应用第 3 节优化策略 2，通过添加信息素并有效避免信息素挥发过快的问题，引导智能体快速地找到最优解，具有较快的收敛速度。

5 基于群体智能的文本聚类过程模拟

为了进一步观察应用本文提出 3 种优化策略前后的文本聚类效果，笔者开发了一个基于群体智能的跨语言文本聚类系统。系统模拟过程中采用第 4.2 节介绍的 3 个栏目的中英文文本共 90 条

数据作为待聚类数据,每类含有 30 个文本向量(相同的颜色表示属于相同的类),被随机放置在网格环境中。SI-Cluster 和 SI* -Cluster 算法运行结果如图 5 ~ 6 所示。

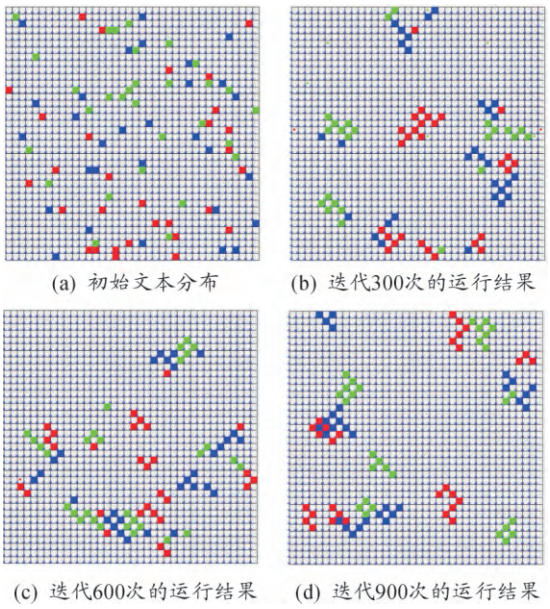

图 5　SI-Cluster 算法不同迭代次数下聚类结果展示

图 5(a)为初始化数据后的网格结构图,图中网格表示网格环境,不同颜色的正方形代表不同类的文本向量。图 5(b)是 SI-Cluster 算法迭代 300 次之后的网格布局。图 5(c)是 SI-Cluster 算法迭代 600 次之后的网格布局。可以看出,算法在迭代 300 次之后所有的文本向量已经被大致划分为 11 个类,有一定的聚集效果。再次迭代 300 次之后网格中又出现了一个全新的聚类结构。算法由于没有应用优化策略,所以经过 900 次迭代后的聚类效果并不好,没有达到收敛,很多不同颜色的文本被错分到一个类簇中。

图 6(a)为初始化数据后的网格结构图,图 6(b)是 SI* -Cluster 算法迭代 300 次之后的网格布局,图 6(c)是 SI* -Cluster 算法迭代 600 次之后的网格布局,图 6(d)是 SI* -Cluster 算法迭代 900 次之后的网格布局。可以看出,算法在迭代了 300 次之后所有的文本向量已经被大致分为 11 个类,相比于 SI-Cluster 算法,聚类效果具有显著的提升。迭代了 600 次之后聚类结构与迭代 300 次时的聚类结构布局相似且相对收敛,证明了本文所提的优化策略具有较好的稳定性。算法迭代 900 次之后类簇结构比较明显,网格中构建了 9 个簇,不同的类之间的重合区域比较少,只需要将聚类结构继续聚合,便可方便地得出最终的聚类结构,进一步说明应用本文提出的优化策略,算法聚类效果好,具有较高的收敛速度。

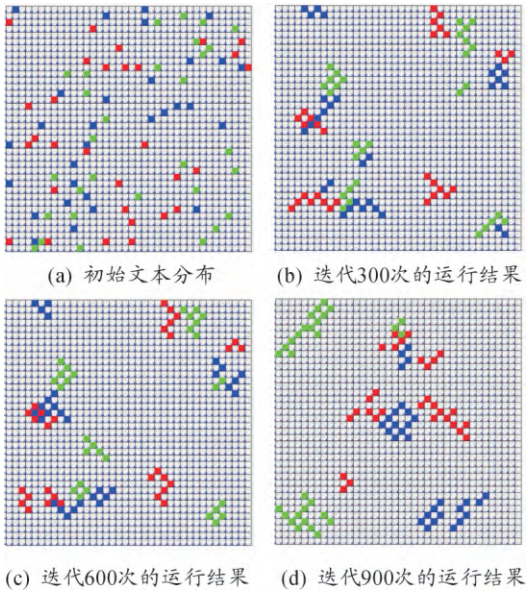

图 6　SI* -Cluster 算法不同迭代次数下聚类结果展示

6　结束语

随着 Web2.0 技术快速发展,Internet 网上信息交流变得更加方便和快捷,不同语言和民族的个体可以通过论坛、微博、社交网站及时发表个人观点,研发跨语言的网络舆情文本挖掘技术对于分析和处理当前日益增长的大规模网络本文信息具有重要的科学价值及应用意义。本文充分考虑群体智能算法的启发性、自适应性和鲁棒性,提出了一种基于群体智能的跨语言文本聚类模型。为了克服传统群体智能算法的不足,所提算法应用了 3 种策略,即: 应用梯度下降法弱化智能体拾取文本对象的能力避免陷入局部最优解,添加信息素引导智能体移动并有效避免信息素挥发过快的问题,智能体从当前位置选择下一位置考虑信息素感应浓度和方向权重因子。大量实验证明本文

所提算法具有较高的聚类准确性和较快的收敛速度。

未来工作将从以下几个方面展开研究：① 将所提的文本聚类方法应用于大规模网络舆情文本挖掘，通过分析网络文本预测可能潜在的突发事件；② 借鉴蜂群等其他更加智能的群体算法，进一步提高文本聚类的准确性；③ 对算法进一步优化，提高群体智能寻找最优解的速度，及算法收敛速度。

参考文献：

[1] 周楠，杜攀，靳小龙，等. 面向舆情事件的子话题标签生成模型 ET-TAG [J]. 计算机学报，2018，41(7)：1490－1503.

[2] 孔雪娜，孙红. 中文微博文本采集与预处理综述 [J]. 软件导刊，2017，16(2)：186－189.

[3] GARCIA J，CRAWFORD B，SOTO R，et al. A *k*-means binarization framework applied to multidimensional knapsack problem [J]. Applied Intelligence，2018，48(2)：357－380.

[4] GAN H，HUANG R，LUO Z，et al. On using supervised clustering analysis to improve classification performance [J]. Information Sciences，2018，454－455：216－228.

[5] 乔少杰，韩楠，金澈清，等. 基于 Multi-agent 的分布式文本聚类模型 [J]. 计算机学报，2018，41(8)：1709－1721.

[6] FERNANDES C M，MORA A M，MERELO J J，et al. KANTS：A stigmergic ant algorithm for cluster analysis and swarm art [J]. IEEE Transactions on Cybernetics，2017，44(6)：843－856.

[7] LIN C T，PRASAD M，SAXENA A. An improved polynomial neural network classifier using real-coded genetic algorithm [J]. IEEE Transactions on Systems，Man，and Cybernetics：Systems，2015，45(11)：1389－1401.

[8] YANG Q，CHEM W N，YU Z，et al. Adaptive multimodal continuous ant colony optimization [J]. IEEE Transactions on Evolutionary Computation，2017，21(2)：191－205.

[9] VERMA O P，PARIHAR A S. An optimal fuzzy system for edge detection in color images using bacterial foraging algorithm [J]. IEEE Transactions on Fuzzy Systems，2017，25(1)：114－127.

[10] ZHU X，NI Z，CHENG M，et al. Selective ensemble based on extreme learning machine and improved discrete artificial fish swarm algorithm for haze forecast [J]. Applied Intelligence，2018，48(7)：1757－1775.

[11] COLLOTTA M，PAU G，MANISCALCO V. A fuzzy logic approach by using particle swarm optimization for effective energy management in IWSNs [J]. IEEE Transactions on Industrial Electronics，2017，64(12)：9496－9506.

[12] SINGH A，BANDA J. Hybrid artificial bee colony algorithm based approaches for two ring loading problems [J]. Applied Intelligence，2017，47(4)：1157－1168.

[13] LI G，BOUKHATEM L，WU J. Adaptive quality-of-service-based routing for vehicular Ad Hoc networks with ant colony optimization [J]. IEEE Transactions on Vehicular Technology，2017，66(4)：3249－3264.

[14] QIAO S，HAN N，ZHOU J，et al. SocialMix：a familiarity-based and preference-aware location suggestion approach [J]. Engineering Applications of Artificial Intelligence，2018，68(2018)：192－204.

（责任编辑　陈　艳）