基于 Multi-agent 的分布式文本聚类模型

乔少杰[1] 韩 楠[2] 金澈清[3] 高云君[4] 李天瑞[5] 唐常杰[6] 康 健[5]

[1](成都信息工程大学网络空间安全学院 成都 610225)

[2](成都信息工程大学管理学院 成都 610103)

[3](华东师范大学计算机科学与软件工程学院 上海 200062)

[4](浙江大学计算机科学与技术学院 杭州 310027)

[5](西南交通大学信息科学与技术学院 成都 611756)

[6](四川大学计算机学院 成都 610065)

摘 要 Internet 网络大数据与日俱增，当前亟需设计能够处理大规模半结构化和无结构化文本数据的新型聚类方法。现有工作的不足体现在：应用的文本集较为单一，对半结构和无结构的 Web 文本进行聚类的准确性较低，当文档规模较大时聚类的时效性无法得到保证。针对上述不足，提出新的基于群体智能的文本聚类模型 Switch (a **Sw**arm **i**ntelligence based **t**ext **c**lustering algorit**h**m)，支持包括藏文、中文、英文等多语言的文本聚类。基本思想为：构建文本的向量空间模型，借助自然语言处理和数据预处理技术得到由特征向量构成的文本集合；对群体智能文本聚类算法的参数进行初始化，不同智能体可以在二维文本空间上任意移动，计算其所在网格区域文本与其他样本的相似度，利用概率转换函数求取智能体拿起和放下样本的概率，进而实现文本聚类。提出分布式动态文本流聚类的 multi-agent 架构，将这一架构应用于群体智能文本聚类算法中，分布式工作环境被设计成相互通信的软 agents 集合，设计了相似度计算，智能体状态感知，文本解析三类智能体。通过解决智能体状态同步、处理器负载均衡和处理器之间通信的代价问题，将计算任务分成不同子任务，在多处理器上分布执行。此外，阐述了基于 multi-agent 的分布式群体智能文本聚类方法的工作原理，给出一种分布式通信架构，各种智能体相互通信，相互协作完成文本聚类工作。基于 multi-agent 通过 JADE 中间件实现集群上的分布式文本聚类，优势在于：分布式计算和大内存处理较单机具有更好的处理能力，借助 JADE 中间件能够使智能体间相互通信及协作，实现高效的文本聚类。在大量真实的半结构化包含藏文、中文和英文多语言的 Web 文本数据集上进行实验，以藏文为例，实验结果表明：相比于 *k*-means 和单节点上的群体智能聚类算法，提出的分布式架构下文本聚类算法准确性平均高出 12.2%和 3.8%，时间代价平均缩减了 73.0%和 50.6%。在 *n* 个节点集群下 agents 数量介于 150-250 之间时，文本聚类时间代价近似可以达到单节点的 1/*n*。

关键词 multi-agent；分布式架构；群体智能；文本聚类

中图法分类号 TP311

论文引用格式：

乔少杰，韩 楠，金澈清，高云君，李天瑞，唐常杰，康 健, 基于 Multi-agent 的分布式文本聚类模型,2017, Vol.40, 在线出版号 No.35

QIAO Shao-Jie，HAN Nan，JIN Che-Qing，GAO Yun-Jun，LI Tian-Rui，TANG Chang-Jie，KANG Jian, A Distributed Text Clustering Model Based on Multi-agent, 2017, Vol.40,Online Publishing No. 35

本课题得到国家自然科学基金(61100045, 61165013, 61363037)、教育部人文社会科学研究规划基金(15YJAZH058)、教育部人文社会科学研究青年基金(14YJCZH046)、四川省教育厅资助科研项目(14ZB0458)、成都市软科学项目(2015-RK00-00059-ZF)资助. **乔少杰**，男，1981年生，博士，教授，计算机学会(CCF)高级会员(E200013959S)，主要研究领域为大数据、数据挖掘、移动对象数据库，E-mail: sjqiao@cuit.edu.cn. **韩 楠(通讯作者)**，女，1984年生，博士，讲师，主要研究领域为数据挖掘，E-mail: hannan@cuit.edu.cn. **金澈清**，男，1977年生，博士，教授，主要研究领域为不确定数据管理. **高云君**，男，1977年生，博士，教授，主要研究领域为数据库. **李天瑞**，男，1969年生，博士，教授，主要研究领域为智能信息处理. **唐常杰**，男，1946年生，硕士，教授，主要研究领域为数据库. **康 健**，男，1986年生，硕士，主要研究领域为Web数据挖掘.

A Distributed Text Clustering Model Based on Multi-agent

QIAO Shao-Jie[1)] HAN Nan[2)] JIN Che-Qing[3)] GAO Yun-Jun[4)] LI Tian-Rui[5)]

TANG Chang-Jie[6)] KANG Jian[5)]

[1)] (School of Cybersecurity, Chengdu University of Information Technology, Chengdu 610225)

[2)] (School of Management, Chengdu University of Information Technology, Chengdu 610103)

[3)] (School of Computer Science and Software Engineering, East China Normal University, Shanghai 200062)

[4)] (College of Computer Science and Technology, Zhejiang University, Hangzhou 310027)

[5)] (School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756)

[6)] (College of Computer Science, Sichuan University, Chengdu 610065)

Abstract As the Internet big data grow rapidly, it urgently needs us to design new clustering approaches that can handle large-scale semi-structured and unstructured text data. The existing methods have the following disadvantages: the commonly used text datasets are very monotonous, the accuracy of text clustering on semi-structured and unstructured Web texts is very low, and the efficiency of clustering can not be guaranteed when the cardinality of documents is very large. Aiming to cope with these drawbacks in existing methods, a new clustering model based on swarm intelligence was proposed, called Switch (a **Sw**arm **i**ntelligence based **t**ext **c**lustering algorit**h**m), which can support multiple languages including Tibetan, Chinese, and English as well. The basic idea of the proposed method is that: it first constructs the vector space model and then obtains the feature vector set of texts by employing the natural language processing and data preprocessing techniques. The parameters of the proposed swarm intelligence based text clustering algorithm are initialized, and the agents can randomly move in a two dimensional text space. The agents calculate the similarity of texts in the grids where they currently stay in to other texts, and use the probability transition function to calculate the probability of picking up and dropping down texts. A distributed dynamic text stream clustering architecture based on multi-agent was proposed, and the proposed distributed architecture was applied to the swarm intelligence based text clustering approach. The distributed working environment of swarm intelligence is designed to be a set of soft agents through communication. Three agents were proposed, including similarity calculation agents, state awareness agents and text parsing agents. By coping with the problems of agent states synchronization, the cost of communication between processors, and load balancing of processors, the calculation tasks are partitioned into different subtasks and the processors perform these tasks in a distributed fashion. In addition, the working mechanism of the proposed distributed swarm intelligent clustering approach based on multi-agent was introduced and the distributed communication schema was given, by which the agents can communicate with others and collaborate with each other to complete the task of text clustering. The distributed clustering on computer clusters can be achieved by the middleware of JADE based on multi-agents, and its advantages include: it has better distributed computing power and large memory processing capability than the stand-alone processing, and employs JADE middleware to perform communication and cooperation among agents in order to complete text clustering efficiently. Experiments were conducted on real semi-structured Web text datasets including Tibetan, Chinese and English. By taking Tibetan as an example, the results show that: the clustering accuracy of the proposed distributed clustering approach is averagely improved by 12.2% and 3.8% and the time cost is reduced by 73.0% and 50.6% on average by comparing to the *k*-means and stand-alone single node cluster. The results show that when the number of agents is between 150 and 250 in the computer cluster with n nodes, the time cost of text clustering might approximate to $1/n$ time cost with regard to a stand-alone node.

Key words multi-agents; distributed architecture; swarm intelligence; text clustering

1 引 言

互联网上的海量文本数据呈爆炸式增长，这些信息普遍存在于门户网站、论坛、微博及信息系统等。高效和有效地处理半结构化和无结构化的文本数据面临巨大挑战，大数据的产生需要更强大的信息处理能力。数据聚类分析是数据挖掘中重要的任务，其本质是将数据集聚类划分，找出数据间的内在关联[1]，数据类簇内的文本相似度极高，不同类簇间的相似度极低[2]。文档聚类常用于无监督的文档划分操作，自动话题的抽取，信息检索等。*k*-means 算法[3]为经典的聚类划分方法，思想简单、直接，然而其聚类效果特别依赖输入的初始化的聚类中心点个数，可能聚类的簇的数量等先验知识。相对于单机计算，分布式技术适合于对大规模数据的聚类划分和识别[4]，引起了国内外学者的广泛关注。

本文提出的新型文本聚类算法受智能群体生物的启发，用来解决 Web 数据挖掘问题。群体智能算法的特性是大量智能个体间的相互协作，遵循一定的规则，且涌现出的特性具有单个个体所不能实现的群体特性。文中提出的基于群体智能生物启发式文本聚类算法，将分布式 multi-agent 和群体智能的文本聚类技术应用于更新文本流的聚类。

2 相关工作

近年来，群体智能模型，如：蚁群聚类算法，粒子群优化算法，得到学者的广泛关注和深入研究。其中，蚁群聚类算法的优势在于不需要先验知识，即：样本初始的划分，文本分类数量等信息。Alsulami等人[5]提出了基于multi-agent系统的语义聚类方法，加强Web信息搜索能力，不足之处在于文档不能分配到多个集群上。Forestiero[6]采用生物启发技术构建基于multi-agent的分布式推荐系统。为了有效地对内存进行有效管理，文献[7]介绍了一种分布式群体智能模型，所提方法的不足之处在于某些路径信息被分享时，不能更准确地控制这些信息。为了解决协作移动学习中的资源共享问题，Iglesia等人[8]提出了一种自适应multi-agent系统，具有个体交互学习的自治性，群组协同学习的资源共享性及学习系统的鲁棒性。为了降低智能体之间异步通信的代价，Ilie 等人[9]设计分布式群体智能模型求解问题。康健等人[10]提出了基于群体智能的半结构化Web文本聚类算法，寻找最相似文本过程中通过信息素形成正反馈机制来完成，提高了算法收敛速度。文献[11]在面向服务的multi-agent系统中引入同质性来创建有效的分散和自组织结构，不足之处在于系统在特殊情况下，如破坏事件发生，才具有更好的鲁棒性。

通过分析上述工作，可以发现当前基于群体智能的文本聚类方法的不足主要体现在：使用的文本集比较简单且规模较小，对半结构和无结构的 Web 文本进行聚类的准确性较低，当文档规模较大时聚类的时效性无法得到保证。为了解决上述问题，本文将 multi-agent(称为多智能体)模型应用于半结构化和无结构化 Web 文本的聚类，提出了一种分布式计算框架并应用于文本聚类，不但适用于不同类型语言文本集的聚集操作，而且由于采用了多智能体的分布式处理架构，可以针对大规模 Web 文本集进行挖掘，处理效率非常高。此外，对传统的基于蚁群的聚类算法进行改进，多智能体协调工作，计算文本相似度，使得结果更加准确。

3 基于智能体的 Web 文本聚类模型

本节介绍一种基于智能体的新型文本聚类模型，对大规模半结构化和无结构 Web 文本进行聚合。算法的核心思想是：利用群体智能算法计算二维文本空间上不同 Web 样本之间的相似度，进而对文本进行拾起和放下操作，实现高效准确的大规模文本聚类分析。下面首先给出算法中的主要概念。

定义 1 (文本聚类模型). 已知预处理后的 Web 文本集合 $T=\{t_1, t_2, \cdots, t_n\}$，聚类得到集合 $C=\{C_1, C_2, \cdots, C_m\}$，$\sum_{i=1}^{m} C_i=T(i=1, 2, \cdots, m)$，对于每个 t_i ($t_i \in T$)，$C_j(C_j \in C)$，$t_i \in C_j$，使得目标函数 $Q(C)$达到最大值(目标函数的指标很多，常用到的如误差平方和)。其中，n 表示文本数目，k 为聚类后类簇数量，$C_j \cap C_l = \varnothing$，$j \neq l$。

定义 2 (群体智能文本聚类). 假设具有 n 维数据集的 m 个文本对象$\{O_1, O_2, \ldots, O_m\}$，利用群体智能进行文本聚类定义为：多智能体协同工作，将 m 个文本对象中簇内相似度较高，并且不同簇之间相似度较低的对象划分到同一个聚簇中，满足评估函数最小的要求。

本文所提的基于智能体的 Web 文本聚类模型的主要流程如下图所示[10]：首先，完成群体智能文本聚类算法的参数初始化工作，如信息素，网格大

小的设置。然后，计算智能体所在网格内文本对象相对于其他文本的相似度，并将其作为对象被选中并拿起的概率。当智能体拿起这一文本的概率大于某一概率随机值，则拿起这一文本向量。如果智能体本身携带有样本对象，计算已有样本对象与其所在网格相邻的如图 2 所示 8 个不同区域上样本对象的相似性，作为文本放下的概率值，如果其值大于某一概率随机值，智能体便选择放下这一文本，不同智能体在文本空间中移动，重复上述操作使评估函数达到最小，进而实现无标签文本的聚集操作。

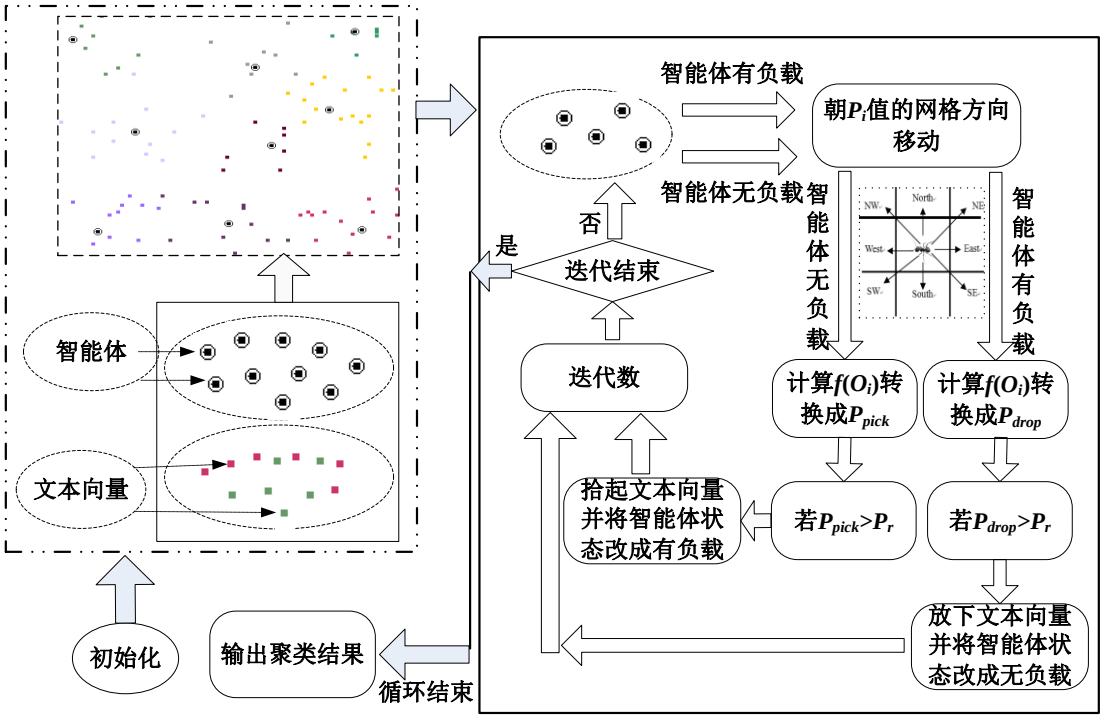


图 1 基于智能体的 Web 文本聚类方法内部原理图

如图 2 所示，在进行 Web 文本聚类过程中，智能体被创建并随机放置于二维网格(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。

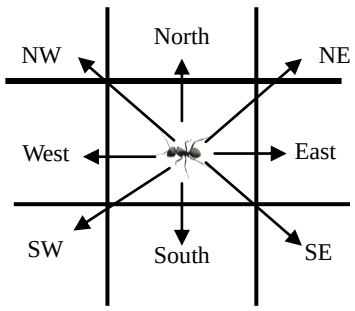


图 2 智能体在网格中可能的移动方向

图 2 展示了 9 个网格、箭头代表相邻网格的朝向和中心为网格中的智能体，表示智能体在当前网格中下一步可能移动的方向。智能体被放置到网格，可能会朝相邻 8 个网格移动，指定朝向分别是 North、South、East、West、NorthEast (NE)、SouthEast (SE)、NorthWest (NW)、SouthWest (SW)，智能体朝信息素存留较大的网格方向移动。

定义 3 (群体相似度)[10]. 智能体的群体相似度利用如下公式求取：

$$f(O_i)=\sum_{O_j\in E_{l\times l}(r)}1-\frac{d(O_i,O_j)}{\alpha} \tag{1}$$

其中，$f(O_i)$表示智能体相似密度，$O_j\in E_{l\times l}(r)$表示文本对象O_j所在的区域，邻域面积为$l\times l$，$d(O_i,O_j)$表示文本对象 O_i和 O_j的差异度，α 因子是一个参数，用于度量文本之间的相似度。

α 因子会影响算法的收敛性能和文本类簇的多样性。本文通过多次实验获得最佳的参数值，具体取值参见表 2。

定义 4 (文本拾起概率). 智能体拿起所在区域样本对象的概率通过如下公式计算：

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

其中，k_p 表示智能体拿起样本对象的概率值，$f(O_i)$表示文本对象 O_i 与邻近其他文本相似度的平均值。

定义 5 (文本放下概率). 智能体放下所在区域样本对象的概率利用如下公式计算：

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

其中，k_d 表示智能体放下样本对象的概率，$f(O_i)$ 表示文本对象 O_i 与邻近其他文本相似度的平均值。

算法 1.　基于群体智能的文本聚类算法.

输入：智能体列表

输出：聚类后的类簇, 文本关系, 智能体位置

```
1.  initialize the agents list L and a 2D grid;
2.  FOR (iteration number i∈N) {
3.      FOR (each agent a∈ L) {
4.          IF ((load == null) && (v != null) )
5.              { compute f(Oi);
6.                  Ppick←ComputePick(Oi);
7.                  IF (Ppick > r) THEN
8.                      { pick up Oi; }
9.              }
10.         ELSE IF ((load != null)&&(v == null)){
11.             compute f(Oi);
12.             Pdrop←ComputeDrop(Oi);
13.             IF (Pdrop > r) THEN
14.                 { drop Oi; }
15.         }
16.         RandomMove();
17.     }
18.     END FOR
19. END FOR
```

如算法 1 所示，基于群体智能的文本聚类算法 Switch(a **sw**arm **i**ntelligence based **t**ext **c**lustering algorit**h**m)主要包含如下操作：

1) 算法第 1 行首先完成群体智能文本聚类的参数初始化工作，如设置网格的大小及信息素值。

2) 算法第 4-9 行计算文本相似度，如果智能体未携带文本且移动到的网格上含有文本信息，计算智能体携带文本的相似性和拾起文本的概率。如果拿起文本的概率大于随机概率值，则执行拾起操作。其中，*load* 表示智能体是否加载有文本，*v* 表示智能体所在网格区域是否有文本向量。

3) 算法第 10-14 行利用公式 2 计算智能体拿起所在网格上样本与其他样本的相似度，并利用公式 3 计算智能体放下样本的概率，判断是否放下这一样本。

3) 智能体随机移动到下一个网格(第 16 行)。

算法性能：通过分析可以得到算法的时间性能为 $O(i*a)$，其中，a 表示智能体的数量，i 表示算法的运行次数。

本节提出的基于智能体的文本聚类算法虽然可以通过智能体之间信息素的激励作用，自发地寻找到最优解，但是面对大规模海量的半结构化和无结构化的 Web 文本，算法的运行时间较长，效率极低。因此，需要借助多智能体技术实现分布式文本聚类，这样可以极大地减少算法的运行时间。

4　基于 Multi-agent 分布式文本聚类

上一节介绍的基于群体智能的文本聚类算法较传统的基于距离的聚类方法具有更高的准确性和时间性能优势，可以实现半结构化的文本数据进行准确的聚集处理。然而，实时大规模文本数据的聚类对计算要求很高，单处理机已经无法满足巨大的计算开销。考虑到群体智能聚类的智能特性，使用分布式方法可以提高文本聚类的时间性能。本节提出新型基于多智能体的分布式架构，对基于智能体的文本聚类算法进行改进，达到高效准确文本聚类效果，本文将基于多智能体的分布式群体智能算法简称为 DSwitch。

本文提出的基于 multi-agent 的分布式架构由互相协调工作的 agent(称为代理)构成，其中代理按照功能分成两类[12]：自主决策代理和独立控制代理。将分布式架构应用于基于智能体的文本聚类模型的基本思路是将复杂的工作划分成不同的小工作，均匀地分配到不同的处理器上进行协同工作。利用智能体对文本进行聚类，其中计算代价最高的操作包括：文本相似度的计算，智能体状态的感知、及文本的解析。针对上述三项子任务，本文设计实现了三类不同的智能体，实现分布式架构将上述子任务分配给不同的处理器去执行，实现了文本的分布式聚类运算，极大地提高了处理文本的时间效率。

定义 6 (分布式系统). 分布式系统由计算机网络将地理上分散的各逻辑单位连接起来而组成，被连接的逻辑单位称为节点，节点是指物理上或逻辑上的单机系统。

定义 7 (负载感知度). 负载感知度定义为 $L(i)$，已知节点 i 能处理的最多 agent 个数是 m_i，θ_i 表示节点 i 上的阈值常数，n_i 表示节点 i 上智能体的数量，节点 i 的负载感知度 $L(i)$定义为：

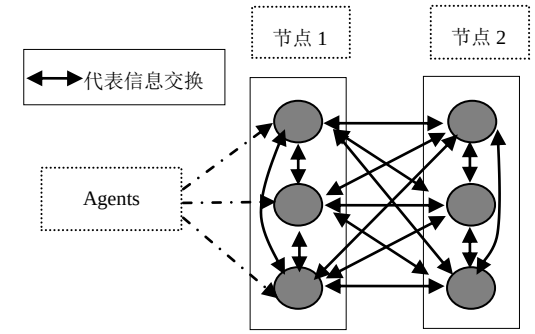
$$L(i)=\begin{cases}\dfrac{(1-c)n_i}{m_i-\theta_i}+c\times m_i-\theta_i, \theta_i<n_i<m_i, c\in(0,1) \\ 0 \quad \text{其他}\end{cases} \tag{4}$$

其中，$L(i)$表示主机感知agent负载大小的能力，仅与节点 i 上当前智能体个数 n 相关，每台主机容器都有一个最大可承受的 agent 数量 m_i，θ_i 是阈值常数，阈值越大说明能支持 agent 的数量越多，c 为常数。

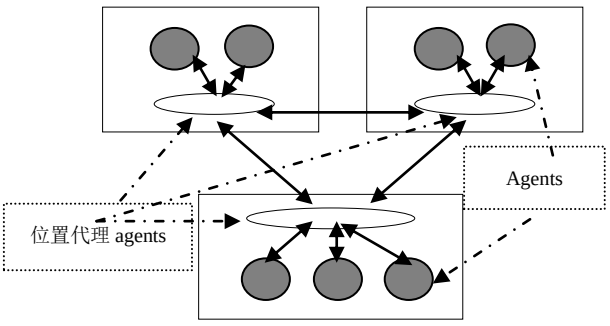
在本文提出的基于多智能体的文本聚类模型中，完整的任务被表示成三类子任务，即：$W=\{W_1, W_2, W_3\}$，其中，W_1 表示文本相似度计算子任务，W_2 表示智能体状态感知子任务，W_3 表示文档解析子任务。

于是，分配的子任务对应三种 agent 实体： 相似度计算实体用于计算文本的相似度，智能体状态感知实体用于获得智能体的移动速度和移动方向等状态信息，文本解析实体用于解析不同文本信息。此外，DSwitch 算法在聚类过程中，还会应用聚类实体对文本进行聚集操作。

在本文提出的基于多智能体的分布式架构中，当 agent 数量发生变化，如：增加或者减少，不会对其他智能体产生影响，具有可伸缩性，可以应对工作负载的变化。



(a) 简单广播通信



(b) 基于位置代理协同工作

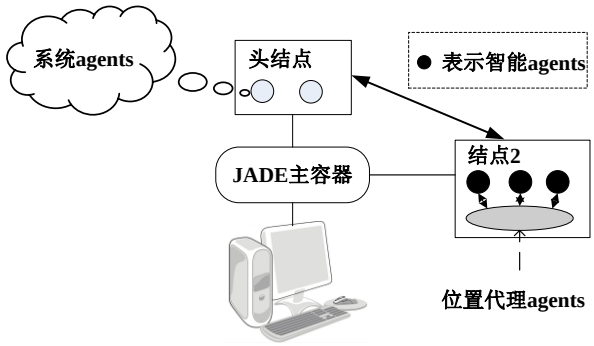
图 3 不同场景下智能体协同通信模型

分布式multi-agent的文本聚类模型中需要解决的主要问题包括：(1) 分布式系统主要由完成不同子任务的处理器构成，当 agent 交换数据或协同工作时需要保证智能体之间的同步性；(2) 分布式工作环境中，智能体并不知道其他智能个体的信息，若与其他智能体进行通信，必须首先知道其他智能体的位置信息，然后将信息传送给其他智能体。因此，分布式架构中的每个智能体都应该能够感知周围智能体的移动速度和方向等状态信息。本文采用如图 3 所示的不同场景下 agent 协同通信模型，实现不同处理其上智能体的状态更新。如图 3(a)所示的广播式通信模型，每个 agent 将状态信息通过发送广播的方式给传递给其他的 agent，并且利用广播信息探寻到邻近的智能体，获取包括移动速度和方向等状态信息。这一模式的不足在于广播需要占用大量的通信带宽。

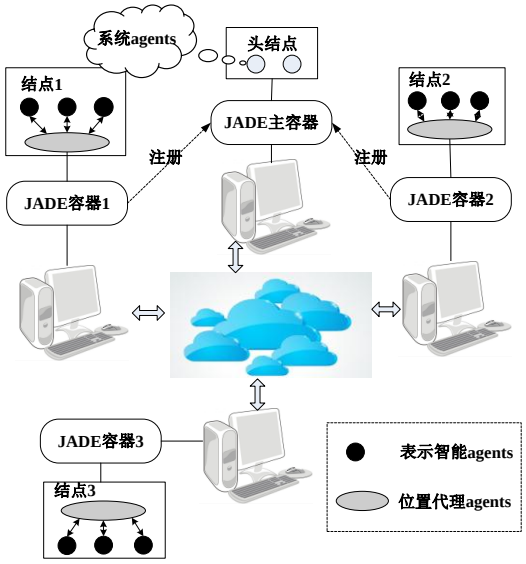
针对上述问题，如图 3(b)所示，增加了具有分布式环境信息的位置代理 agent，每个处理器上设置一个代理，分布式计算过程中智能体仅需将状态信息传递给位置代理。此外，智能体也可以利用位置代理获得邻近智能体的位置信息。因为位置代理在每次收集智能体的状态信息后，会将这些状态信息发送给其他智能体，所以每个处理器上的位置代理都记录了整个分布式系统的环境信息。当位置代理得到当前处理器上 agent 的状态信息后，会将这些 agent 状态发送给其他节点的位置代理，同理另外两个节点也会以此方式告知节点内的 agent 状态，以此达到每个代理 agent 都具有整个系统的全局状态信息的目的。

基于 JADE (Java Agent DEvelopment Framework)分布式计算平台[13]，本文将群体智能文本聚类算法进行分布式实现，将智能体按照执行任务的不同分配到不同的处理器上。在本文所提出的分布式文本聚类模型中，不同智能体按照执行任务的不同被均匀分配到 3 类处理器上，且每类处理器利用自己的位置代理收集智能体执行操作后的位置信息。单处理器模型和分布式集群架构图示分别见图 4(a)-(b)。如图 4(b)所示，在分布式集群模型中，主机之间是通过网络连接的，每个主机上都有一个 JADE 容器，其中有一台作为 JADE 主容器，其他的 JADE 容器加入要注册服务，JADE 主容器中有系统 agent 主要用于负责控制平台内 agent 的活动、生存周期及外部应用程序与平台的交互。位置代理用于收集节点内的智能单体的状态，与其他位置代

理进行交互，获得全局的智能单体的状态信息。



(a) 单处理器模型



(b) 分布式集群模型

图 4　JADE 平台不同计算模型架构

多个 JADE 容器可以同时运行在同一个处理器上，当智能体被分配到不同的 JADE 容器中执行任务时，在同一个 JADE 容器中比分配到不同容器中通信效率更高。基于这一考虑，本文将智能体部署在同一个处理器的相同 JADE 容器中。此外，算法将 JADE 的主容器作为处理器集群的头结点，这样可以减少 JADE 系统的计算开销。

图 5 给出了基于 JADE 的多智能体分布式文本聚类模型主要工作流程。在基于 JADE 的分布式文本聚类模型中，智能体之间通过信号触发的方式实现直接通信，利用位置代理得到其他处理机上智能体的位置、速度和方向等状态信息，并触发间接通信。在任务分配层，JADE 主容器将主任务划分成三类子任务，即：W_1，W_2，W_3，分别表示文本相似度计算，智能体状态感知和文本解析子任务。在任务执行层，不同的子任务被分配给由 A 代表的多个 agents 构成的 JADE 容器分布式执行。

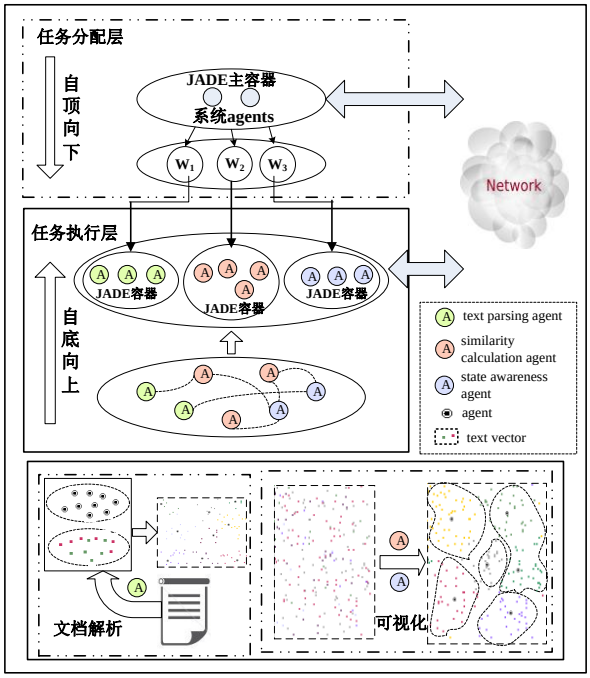


图 5　基于 JADE 的分布式文本聚类模型工作流程

本文提出的分布式计算结构描述如下：

算法 2.　分布式通信架构.

1. **IF** *type* = “single” **THEN** *addTask*();
2. **IF** *p* = “send” **THEN** add “accept”;
3. **ELSE IF** *p* = “receive” **THEN** add “send”;
4. **ELSE** throw Exception;
5. **ELSE IF** the node has three subnodes
6. *subtask*(1,2,3)←*TaskPartition*();
7. *clu*(1,2,3)←*Clustering*(*subtask*(1,2,3));
8. *result*←*Merge*(*clu*(1,2,3));
9. **ELSE IF** the node has *n* subnodes
10. register these *n* agents;
11. *subtask*(1,2,…,*n*)←*TaskPartition*();
12. *clu*(1,2, …,*n*)←*Clustering*(*subtask*(1,2, …,*n*));
13. *result*←Merge(*clu*(1, 2,…,*n*);

分布式通信架构的主要描述如下：

(1) 判断当前节点类型是单节点处理，直接添加一个简单任务，方法内实现基于群体智能的文本聚类算法(第 1 行)。

(2) 如果通信参数是“send”则添加 accept 消息；否则，如果“receive”则添加 send 消息，并将接收到的信息编码；否则抛出异常(第 2-4 行)。

(3) 如果节点为 3 个，将任务分割成 3 个子任务，调用 *Clustering* 函数对子任务进行处理，获得 3 个子聚类结果，对聚类结果进行聚合(第 5-8 行)。

(4) 若节点有 *n* 个，则将其注册到 agent 管理系

统中(第 9-10 行)。

(5) 调用 *TaskPartition* 函数分割成 *n* 个子任务，分别聚类，对聚类结果合成，得到最终聚类结果集合 *result*(第 11-13 行)。

5 实验及算法性能分析

5.1 实验环境及数据集描述

本文所提算法主要应用于包括藏文、中文和英文等多语言的文本聚类，实验中的数据集来源于主要门户网站 Tibet3，新浪，搜狐等主流门户网站抓取的 Web 样本，利用自然语言处理和文本挖掘技术对 Web 数据进行特征向量的抽取，并构建文本向量的 TF-IDF 模型，完成文本数据的预处理。此外，实验采用来自于 UCI 机器学习数据库的文本数据，即：Iris、Wine、Glass 数据集，这部分数据的内聚性较强，能够更好地测试算法性能。

实验中所有算法利用 Java 程序设计语言实现。实验硬件平台为双核 2GHz CPU, 2GB 内存，操作系统为 Windows 7，系统部分参数设置如表 1 所示。

表 1 系统参数设置

参数	值
文档数据集数目	19289
文档训练集(带标注)	8871
文档测试集(不带标注)	10418
智能体数量	30
二维平面行、列数	30
拾起概率 k_p	0.1
放下概率 k_d	0.06

如表 1 所示，实验中使用的训练数据样本数据包含 8871 篇 Web 样本，包含不同类别栏目下的新闻信息，通过降维和特征词约减操作，保留了 2120 个特征词；不带标注的 10418 篇文本信息作为测试数据样本，包括不同类别栏目的新闻数据，保留了 3219 个特征词。

群体智能聚类算法 Switch 详细参数设置如表 2 所示，为通过多次实验获得的最佳参数值。

表 2 Switch 算法参数设置

参数	说明	取值
pheromone	信息素，决定智能体下一步移动方向	5.0
α	计算群体相似度因子	0.7
lines	网格行数	20
columns	网格列数	20
number	网格上的智能体数量	10
step	步长范围	1.0
life_time	智能体生命周期	50
sigma	计算感应信息素的参数	2.0

5.2 文本聚类算法准确性评价

评价文本聚类算法的正确性的指标包括准确率(用 *P* 表示)，召回率(用 *R* 表示)，*F*-measure 等。准确率 *P* 定义为聚类得到的文档数量除以所有文档数量的总和。召回率 *R* 定义为聚类得到的文档数量除以样本库中与这一类簇相关文档的数量。

实验中文本聚类算法的准确性主要通过 *F*-measure 反映出来，其取值越大说明聚类算法的准确性越高。

已知 TP 为被正确聚类出的相关文本；FP 表示不相关的被聚出的文本；FN 表示相关的但未被聚类到的文本；TN 表示不相关的且未被聚类的文本，于是：

$$P=\mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \tag{4}$$

$$R=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \tag{5}$$

F-measure 定义如下：

$$F\text{-measure}=2*P*R/(P+R) \tag{6}$$

本节将分别对比基于群体智能的文本聚类算法 Switch，*k*-means 和基于 multi-agent 的群体智能文本聚类算法 DSwitch 三种算法对藏文、中文和英文进行本文聚类的准确率、召回率和 *F*-measure 值。

表 3 基于 Switch 算法的藏语言 Web 文本聚类结果准确性

类标	主要词汇	样本数量	正确识别数	错误识别数	*R*	*P*	*F*-measure
1	རྩ་བའི་མ་དཔུལ (水利投资), དཀའ་ངལ་ཞི་གསལ (困难补助)等 attention	410	306	235	0.427	0.566	0.487
2	སྐྱེ་འཚས་ཚོན་རྩལ (生态产业), འཕྲིན་འབྲུལ་ཀྱི་བའི་འཇགས (交通安全)等 tibet	6645	4530	1894	0.405	0.705	0.515
3	ཡུལ་ཀྲུལ་ག་ཡོས (玉树地震) ས་ཁའི་གོག་གཅོད་ཞེས (抗震救灾)等 inland	485	375	125	0.436	0.75	0.551
4	འཇིགས་སྐྲག་འཇངས་རྩོལ (恐怖袭击)等 international	1103	676	411	0.380	0.622	0.472

5	དེང་རབས་སློབ་གསོའི	（现代教育），ཀྱི་རྒྱམ་རིག	（民间文学）等 culture	228	169	150	0.426	0.530	0.472
		R, *P*, *F*-Measure 的平均取值					0.414	0.634	0.499

表 4　基于 *k*-means 的藏语言 Web 文本聚类结果准确性

类标	主要词汇	样本数量	正确识别数	错误识别数	*R*	*P*	*F*-measure
1	རྩ་བའི་མ་དཏུལ	（水利投资），དཀའ་ངལ་ཞ་གསལ（困难补助）等 attention	410	263	223	0.391	0.541	0.454
2	སྐྱེ་གཞས་ཚོན་ལས	（生态产业），འཕྲིན་ལས་ཀྱི་བའི་འཇགས	6645	4105	2653	0.382	0.607	0.469
		（交通安全）等 tibet						
3	ཡུལ་རྒྱལ་ས་ཡོན	（玉树地震）ཡོན་འགོག་གནོད་སེལ	(抗震救灾)等 inland	485	296	205	0.379	0.591	0.462
4	འཇིགས་སྐྲག་འཇའ་རྩོལ	（恐怖袭击）等 international		1103	569	373	0.340	0.604	0.435
5	དེང་རབས་སློབ་གསོའི	（现代教育），ཀྱི་རྒྱམ་རིག	（民间文学）等 culture	228	165	95	0.420	0.635	0.505
		R, *P*, *F*-Measure 的平均取值					0.382	0.569	0.465

表 5 基于 DSwitch 算法的藏语言 Web 文本聚类结果准确性

类标	主要词汇	样本数量	正确识别数	错误识别数	*R*	*P*	*F*-measure
1	རྩ་བའི་མ་དཏུལ	（水利投资），དཀའ་ངལ་ཞ་གསལ（困难补助）等 attention	410	326	170	0.443	0.657	0.529
2	སྐྱེ་གཞས་ཚོན་ལས	（生态产业），འཕྲིན་ལས་ཀྱི་བའི་འཇགས	6645	4795	1721	0.419	0.736	0.534
		（交通安全）等 tibet						
3	ཡུལ་རྒྱལ་ས་ཡོན	（玉树地震）ཡོན་འགོག་གནོད་སེལ	(抗震救灾)等 inland	485	418	149	0.463	0.737	0.569
4	འཇིགས་སྐྲག་འཇའ་རྩོལ	（恐怖袭击）等 international		1103	729	322	0.398	0.694	0.506
5	དེང་རབས་སློབ་གསོའི	（现代教育），ཀྱི་རྒྱམ་རིག	（民间文学）等 culture	228	185	56	0.448	0.768	0.566
		R, *P*, *F*-Measure 的平均取值					0.434	0.718	0.541

表 6　基于 Switch 算法的中文 Web 文本聚类结果准确性

类标	主要词汇	样本数量	正确识别数	错误识别数	*R*	*P*	*F*-measure
1	Mh370 失联，搜救，坠海等马航客机失联	2586	1656	836	0.390	0.665	0.492
2	鲁甸，6.5 级，地震，等鲁甸县发生 6.5 级地震	1550	1018	268	0.396	0.792	0.528
3	韩国客轮，沉没等韩国客轮在西南海域沉没	1253	761	525	0.378	0.592	0.461
4	叙利亚，联合国，化武等叙利亚化武疑云	852	496	341	0.368	0.593	0.454
5	全能神，招远，张帆等 6 人在山东招远快餐厅打死女顾客	961	624	393	0.394	0.614	0.480
6	玉林，狗肉节，爱心人士，热议等广西玉林狗肉节引争议	220	102	132	0.317	0.436	0.367
7	军训，教官殴打师生，教官，冲突等湖南军训教官与师生冲突	675	410	361	0.378	0.532	0.442
8	UFO，黑龙江，坠入黑龙江境内等 5 个不明飞行物坠入黑龙江	421	362	233	0.462	0.608	0.525
	R, *P*, *F*-Measure 的平均取值				0.385	0.604	0.469

表 7　基于 *k*-means 算法的中文 Web 文本聚类结果准确性

类标	主要词汇	样本数量	正确识别数	错误识别数	*R*	*P*	*F*-measure
1	Mh370 失联，搜救，坠海等马航客机失联	2586	1274	1015	0.330	0.557	0.414
2	鲁甸，地震，6.5 级地震等鲁甸县发生 6.5 级地震	1550	859	368	0.356	0.699	0.472
3	韩国客轮，沉没等韩国客轮在西南海域沉没	1253	561	786	0.309	0.416	0.355
4	叙利亚，联合国，化武等叙利亚化武疑云	852	467	420	0.354	0.526	0.423

5	全能神，招远，张帆等 6 人在山东招远快餐厅打死女顾客	961	623	385	0.393	0.618	0.481
6	玉林，狗肉节，爱心人士，热议等广西玉林狗肉节引争议	220	95	162	0.302	0.370	0.332
7	军训，教官殴打师生，教官，冲突等湖南军训教官与师生冲突	675	432	412	0.390	0.512	0.443
8	UFO，黑龙江，坠入黑龙江境内等 5 个不明飞行物坠入黑龙江	421	362	297	0.462	0.549	0.502
	R, *P*, *F*-Measure 的平均取值				0.362	0.531	0.428

表 8 基于 DSwitch 算法的中文 Web 文本聚类结果准确性

类标	主要词汇	样本数量	正确识别数	错误识别数	*R*	*P*	*F*-measure
1	Mh370 失联，搜救，坠海等马航客机失联	2586	1957	718	0.431	0.732	0.542
2	鲁甸，地震，6.5 级地震等鲁甸县发生 6.5 级地震	1550	1159	425	0.428	0.732	0.540
3	韩国客轮，沉没等韩国客轮在西南海域沉没	1253	842	331	0.402	0.718	0.515
4	叙利亚，联合国，化武等叙利亚化武疑云	852	557	297	0.395	0.652	0.482
5	全能神，招远，张帆等 6 人在山东招远快餐厅打死女顾客	961	569	130	0.372	0.814	0.511
6	玉林，狗肉节，爱心人士，热议等广西玉林狗肉节引争议	220	101	113	0.315	0.472	0.378
7	军训，教官殴打师生，教官，冲突等湖南军训教官与师生冲突	675	467	292	0.409	0.615	0.491
8	UFO，黑龙江，坠入黑龙江境内等 5 个不明飞行物坠入黑龙江	421	384	176	0.477	0.686	0.563
	R, *P*, *F*-Measure 的平均取值				0.404	0.678	0.504

表 9 基于 Switch 算法的英文 Web 文本聚类结果准确性

类标	主要词汇	样本数量	正确识别数	错误识别数	*R*	*P*	*F*-measure
1	UFO, fall etc. Multiple UFO fall in heilongjiang	112	62	135	0.356	0.315	0.334
2	Mh370，lost etc. MH370 search discovers "interesting" objects	2453	1462	543	0.373	0.729	0.494
3	earthquake etc. Yunnan earthquake	325	276	221	0.459	0.555	0.503
4	Ebola fears grow, with Europe and Asia on alert	238	165	311	0.409	0.347	0.375
5	7 tourists killed by falling rocks in SW China	156	89	174	0.363	0.338	0.350
6	Yulin, loving people etc. dog festival	312	213	66	0.406	0.763	0.530
7	explosion ,bus etc. Guangzhou bus explosion	469	312	96	0.399	0.765	0.525
8	Vietnam , hit, enterprise etc. Vietnam hit foreign companies	384	243	81	0.388	0.750	0.511
	R, *P*, *F*-Measure 的平均取值				0.394	0.570	0.453

表 10 基于 *k*-means 算法的英文 Web 文本聚类结果准确性

类标	主要词汇	样本数量	正确识别数	错误识别数	*R*	*P*	*F*-measure
1	UFO, fall etc. Multiple UFO fall in heilongjiang	112	30	184	0.211	0.140	0.169
2	Mh370，lost etc. MH370 search discovers "interesting" objects	2453	953	965	0.280	0.497	0.358
3	earthquake etc. Yunnan earthquake	325	176	245	0.351	0.418	0.382
4	Ebola fears grow, with Europe and Asia on alert	238	118	189	0.331	0.384	0.356
5	7 tourists killed by falling rocks in SW China	156	120	235	0.435	0.338	0.380
6	Yulin, loving people etc. dog festival	312	171	186	0.354	0.479	0.407
7	explosion ,bus etc. Guangzhou bus explosion	469	286	208	0.379	0.579	0.458
8	Vietnam , hit, enterprise etc. Vietnam hit foreign companies	384	197	186	0.339	0.514	0.409
	R, *P*, *F*-Measure 的平均取值				0.335	0.419	0.365

表 11 基于 DSwitch 算法的英文 Web 文本聚类结果准确性

类标	主要词汇	样本数量	正确识别数	错误识别数	*R*	*P*	*F*-measure
1	UFO, fall etc. Multiple UFO fall in heilongjiang	112	89	43	0.443	0.674	0.535
2	Mh370，lost etc. MH370 search discovers "interesting" objects	2453	1682	374	0.407	0.818	0.543
3	earthquake etc. Yunnan earthquake	325	283	164	0.465	0.633	0.536

4	Ebola fears grow, with Europe and Asia on alert	238	172	142	0.420	0.548	0.475
5	7 tourists killed by falling rocks in SW China	156	113	98	0.420	0.536	0.471
6	Yulin, loving people etc. dog festival	312	223	87	0.417	0.719	0.528
7	explosion ,bus etc. Guangzhou bus explosion	469	320	276	0.406	0.537	0.462
8	Vietnam , hit, enterprise etc. Vietnam hit foreign companies	384	207	176	0.350	0.540	0.425
	R, P, F-Measure 的平均取值				0.416	0.626	0.497

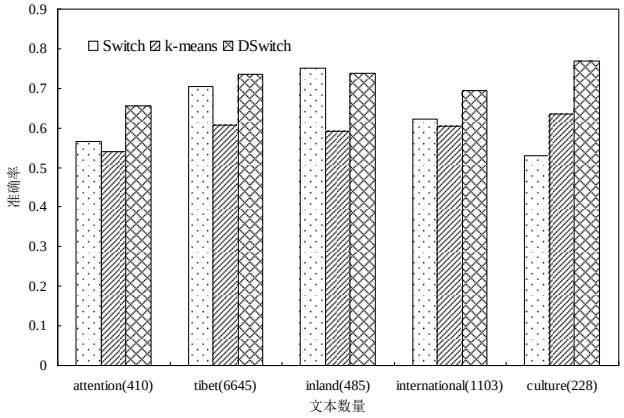


图 6　不同算法 Web 文本聚类的准确率 *P* 比较

通过图 6 可以发现：Switch 和 DSwitch 算法的准确率明显好于 *k*-means 算法，DSwitch 算法略优于 Switch 算法。其中，横轴表示类标+文本数，如 tibet(6645)表示 tibet 栏目新闻包含 6645 个文本。

不同栏目的 Web 样本数据聚类过程中选取不同维度的属性本文，样本本身的内聚性不同，其聚类效果也是不一样的。通过 inland 样本集上的实验结果可以发现：基于 *k*-means 的文本算法的准确性略高于 Switch 算法。因为 culture 样本的文本差异较大，且样本规模较小。而在其他几种样本数据集上进行聚类，可以获得比较高的聚类准确性。此外，分布式文本聚类算法 DSwitch 性能要明显优于 Switch 和 *k*-means 算法。

5.2　文本聚类时间性能对比

为了进一步验证本文所提算法的性能优势，本节实验观察随着不同类型语言文本数量的增加，不同算法运行时间的变化情况，实验结果如图 7-9 所示。可以发现：算法的时间性能曲线与算法时间复杂性一致，Switch 和 DSwitch 算法的时间效率优于基于 *k*-means 的聚类方法。尤其是当 Web 文本规模非常大的时候，基于群体智能的文本聚类算法耗时明显优于 *k*-means 算法。以藏文文本为例，DSwitch 算法相对于 *k*-means 和 Switch 算法时间代价平均缩减了 73.0%和 50.6%。主要原因在于：基于群体智能的多语言文本聚类算法本质是通过信息素形成正反馈机制聚合文本，算法运行效率较高。此外，DSwitch 算法时间性能优于 Switch 算法，因为分布式 DSwitch 算法将单处理机上的任务分配到 3 个节点上进行分布式计算，聚类消耗时间明显减少。

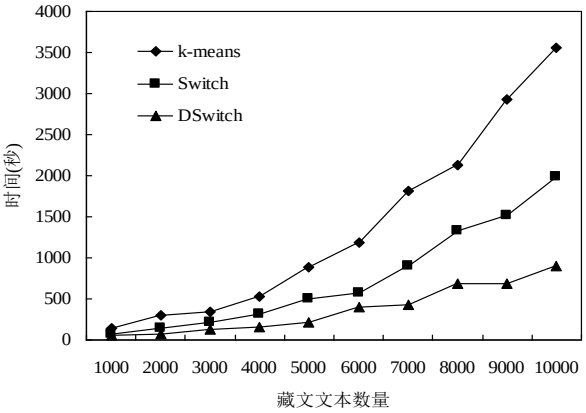


图 7　不同算法藏文文本聚类时间对比

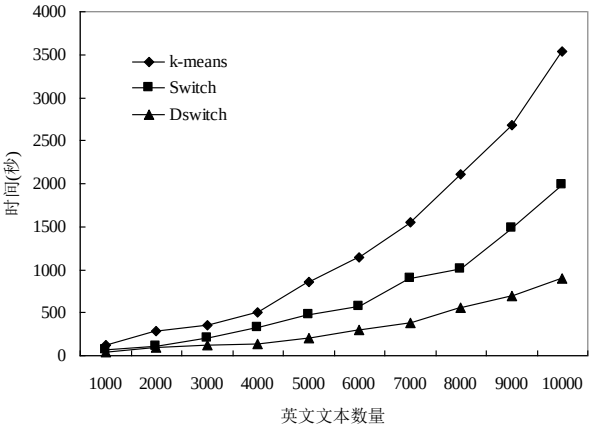


图 8　不同算法英文文本聚类时间对比

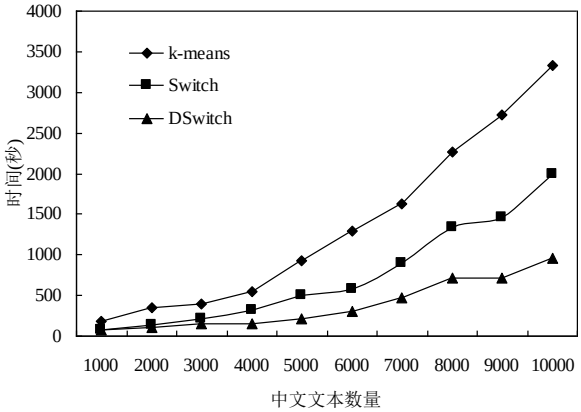
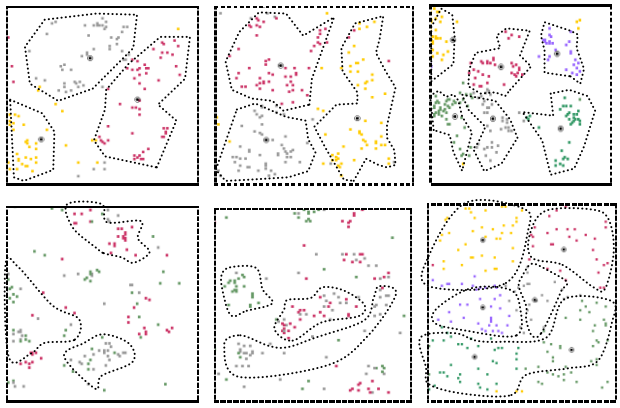


图 9 不同算法中文文本聚类时间对比

5.3 基于群体智能的文本聚类效果分析

实验中采集到的离散文本的内聚性较弱，为了更好地对比基于群体智能的 Switch 和分布式 DSwitch 算法，本节实验引入 UCI 机器学习样本集 Wine、Iris、Glass，三类数据集分别包含 178 个，150 个和 214 个样本，每类数据包含的属性个数分别是：3 个，4 个和 9 个。

DSwitch 和 Switch 算法在上述 3 组数据集上的文本聚类结果的如下图所示，从左向右分别表示 Iris，Wine 和 Glass 三个样本集，同一列的两张效果图表示相同样本集上的文本聚类结果。上面三组是 DSwitch 算法聚类效果，相同颜色代表一个类簇，可以发现标出的部分很明显是一类。下面三组是 Swtich 算法聚类效果，与改进后的 DSwitch 的聚类效果对比，相对较差。



(*a*) Iris (*b*) Wine (*c*) Glass

图 10 DSwitch 和 Swtich 算法聚类效果对比图

5.4 基于群体智能的文本聚类迭代次数对比

评价群体智能算法的一个很重要的指标是算法的收敛速度，即聚类完成的迭代次数。本节主要观察 Switch 和 DSwitch 算法在取不同样本数据时算法收敛时的迭代次数。DSwitch 和 Switch 算法在藏文本文聚类上的迭代次数对比结果如图 11 所示。

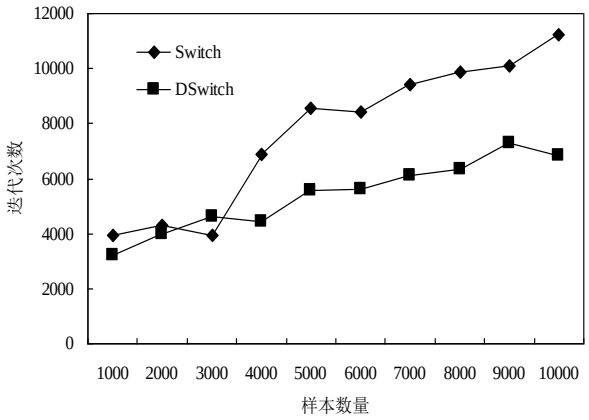


图 11 DSwtich 和 Switch 算法迭代次数比较

实验结果表明：分布式群体智能文本聚类算法 DSwitch 收敛的迭代次数明显小于 Switch 算法。随着 Web 文本数据不断增加，两种基于群体智能的文本聚类算法的收敛迭代次数均持续增加。原因在于：DSwitch 算法在聚类划分过程中，多节点分布式运算，节点之间相互协作，异步交换数据，共同寻求最优解，此过程相比单节点下聚类寻优从时间性能和聚类迭代次数上均表现出了较大的优势。此外，可以发现当文本数量高于 4000 时，DSwitch 算法的迭代次数明显少于 Switch 算法。

5.5 分布式计算时间性能分析

本节观察单节点处理器模型和分布式集群架构在不同数量的 agents 下分别迭代 1000 次的运行时间，结果如图 12 所示。

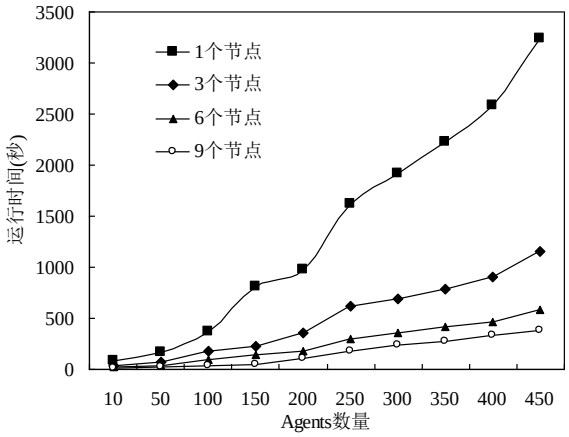


图 12 不同节点集群和单处理器上的运行时间对比

如图 12 所示：当 agents 数量为 10 时，在多个节点集群和单主机节点上聚类处理的时间消耗没有明显的区别；当 agents 的数量大于 120 时，3 个节点以上集群上花费的时间要明显少于单主机节点上花费的时间；当 agents 数量为 150 个或更多时，3 个节点集群上花费的时间为单主机节点花费时间

的 1/3。6 个节点和 9 个节点集群运算结果类似，近似为单节点花费时间的 1/6 和 1/9。在单节点处理器上，当多于 250 个 agents 运行时，可能需要的内存要大于实际内存，此时可能借助虚拟内存，在聚类完成时消耗的时间会有所增加。通过实验可以得到这样的结论：本文所提 DSwitch 算法在 n 个节点集群下 agents 数量介于 150-250 之间时，文本聚类时间代价近似可以达到单节点的 $1/n$。原因在于：在分布式集群架构上，agents 被均等地分配给 n 个不同的集群节点，每个节点仅需要 $1/n$ 的内存需求，明显少于单节点机器上内存的需求，这样就有效地避免了超出节点物理内存限制的问题。

6 结论

针对海量半结构化文本数据，本文提出一种新的分布式架构并应用于基于群体智能的 Web 文本聚类，此外，可以支持不同语言的 Web 文本聚类。所提方法不需要先验知识，结合群体智能和 multi-agent 分布式架构，能高效和高质量地处理在线海量文本流。大量真实的半结构化 Web 文本数据上的实验结果表明，所提基于群体智能文本聚类方法无论在算法的准确性还是运行时间性能上均明显优于 *k*-means 算法和传统的基于群体智能的文本聚类算法。提出的分布式聚类划分方法为线上话题发现和舆情监控等研究提供了较好的技术支持。

未来工作包括：进一步降低分布式环境下节点之间的通信代价，提高算法效率；将所提分布式文本聚类算法应用于在线话题发现和舆情监控等真实应用场景中。

参 考 文 献

[1] Ferrari D G, Castro L N D. Clustering algorithm selection by meta-learning systems: a new distance-based problem characterization and ranking combination methods. Information Sciences, 2015, 301(2015): 181-194

[2] Zhou J, Chen C L P, Chen L, et al. A collaborative fuzzy clustering algorithm in distributed network environments. IEEE Transactions on Fuzzy Systems, 2014, 22(6): 1443-1456

[3] Hartigan J A, Wong M A. A k-means clustering algorithm. Applied Statistics, 1979, 28(1): 100-108

[4] Zhao X, Sayed A H. Distributed clustering and learning over networks. IEEE Transactions on Signal Processing, 2014, 63(13): 3285-3300

[5] Alsulami B S, Abulkhair M F, Essa F A. Semantic clustering approach based multi-agent system for information retrieval on web. International Journal of Computer Science and Network Security, 2012, 12(1): 41-46

[6] Forestiero A. AIRS: Ant-Inspired Recommendation System//Proceedings of the 7th IEEE International Conference Intelligent Systems IS'2014. Warsaw, Poland, 2015: 213-224

[7] Collings J, Kim E. A distributed and decentralized approach for ant colony optimization with fuzzy parameter adaptation in traveling salesman problem//Proceedings of 2014 IEEE Symposium on Swarm Intelligence. Orlando, Florida, USA, 2014: 1-9

[8] Iglesia D G D L, Calderon J F, Weyns D, et al. A self-adaptive multi-agent system approach for collaborative mobile learning. IEEE Transactions on Learning Technologies, 2015, 8(2): 158-172

[9] Ilie S, Bădică C. Multi-agent approach to distributed ant colony optimization. Science of Computer Programming, 2013, 78(6): 762-774

[10] Kang Jian, Qiao Shao-Jie, Ge Sangduoji, et al. A semi-structured Tibetan text clustering algorithm based on swarm intelligence. Pattern Recognition and Artificial Intelligence, 2014, 7(8): 663-670 (in Chinese)
(康健, 乔少杰, 格桑多吉等. 基于群体智能的半结构化藏文文本聚类算法. 模式识别与人工智能, 2014, 27(8): 663-670)

[11] Del Val E, Rebollo M, Botti V. Enhancing decentralized service discovery in open service-oriented multi-agent systems. Autonomous agents and multi-agent systems, 2014, 28(1): 1-30

[12] Zhu Q, Shun Y Q, Cen Y. A clustering algorithm based on multi-agent meta-heuristic architecture. International Journal of Hybrid Information Technology, 2014, 7(2): 227-234

[13] Chmiel K, Gawinecki M, Kaczmarek P, Szymczak M, Paprzycki M. Efficiency of JADE Agent Platform. Scientific Programming, 2005, 13(2): 49-56

QIAO Shao-Jie, born in 1981, Ph.D., professor. His current research interests include big data, data mining and moving objects databases.

HAN Nan, born in 1984, Ph.D., lecturer. Her current research interests include data mining.

JIN Che-qing, born in 1977, Ph.D., professor. His current research interests include uncertain data management.

GAO Yun-jun, born in 1977, Ph.D., professor. His current research interests include databases.

LI Tian-rui, born in 1969, Ph.D., professor. His current research interests include intelligent information processing.

TANG Chang-Jie, born in 1946, master, professor. His current research interests include databases.

KANG Jian, born in 1986, master. His current research interests include Web data mining.

Background

This work is a part of the “Research on Tibetan Online Public Opinion Analysis and Emergency Alert Based on Emergence of Swarm Intelligence”, which is mainly supported by the National Natural Science Foundation of China under Grant No. 61165013. This project treats the websites, bbs, blog and microblog as the subjects, employs Tibetan text mining and natural language processing techniques to retrieve essential factors including the time, location, subjects and objects of the interested events, and designs the attention and attitude models with regard to events and creates the attention and attitude model base with difference levels. This project aims to propose effective and efficient Tibetan public opinion analysis algorithm based on swarm intelligence emergence, including attention of topics analysis, hot topics analysis, focus analysis of public opinions, sensitive topics analysis, break point analysis, frequent point analysis, and key point analysis, by which to predict potential emergencies. Moreover, this project can set up scientific architecture for predicting the emergencies occurring in Tibetan websites, monitor these emergencies breaking out by virtual swarms, and provide solutions corresponding to different levels of warning based on the results of public opinion analysis by swarms.