基于 SOM 聚类的微博话题发现*

宋莉娜[a]，冯旭鹏[b]，刘利军[a]，黄青松[a,c]
(昆明理工大学 a. 信息工程与自动化学院; b. 教育技术与网络中心; c. 云南省计算机技术应用重点实验室，昆明 650500)

摘 要： 随着微博用户的增多，微博平台的信息更新频繁。针对微博文本的数据稀疏性、新词多、用语不规范等特点，提出了基于 SOM 聚类的微博话题发现方法。从原始语料中对文本进行预处理，通过词向量模型对短文本进行特征提取，降低了向量维度过高带来的计算量繁重问题。采用改进的 SOM 对话题进行聚类，该算法改善了传统文本聚类的不足，进而能有效地发现话题。实验表明该算法较传统文本聚类算法的综合指标 F 值有明显提高。

关键词： 话题发现; 词向量模型; 文本相似度; 短文本; SOM 聚类
中图分类号： TP391 **文献标志码：** A **文章编号：** 1001-3695(2018)03-0671-04
doi:10.3969/j.issn.1001-3695.2018.03.007

Microblog topics detection based on SOM clustering

Song Lina[a], Feng Xupeng[b], Liu Lijun[a], Huang Qingsong[a,c]
(*a. Faculty of Information Engineering & Automation, b. Educational Technology & Network Center, c. Yunnan Provincial Key Laboratory of Computer Technology Applications, Kunming University of Science & Technology, Kunming* 650500, *China*)

Abstract: With the increase of microblog users, the information of microblog platform is updating frequently. This paper proposed microblog topics detection based on SOM clustering for the features of the microblog text data sparseness, new words and non-standard words. Firstly, it pretreated the short texts from the primitive text corpus, and extracted the features of the short texts by the word vector model which reduced the computational burden caused by the high vector dimension. In order to reduce the large amount of computation just to the high vector dimensions, this paper extracted the short text feature extraction by word vector model. Then, the topic clustering could be achieved by an improved SOM clustering. The algorithm improved the traditional texts clustering shortcoming. And the algorithm could find the topic effectively. Experimental results show that the algorithm's comprehensive index F value is improved obviously than the traditional methods.

Key words: topics detection; word vector model; texts similarity; short texts; SOM clustering

0 引言

近年来，随着微博用户的不断增多，微博平台已经广泛渗入到人们的生活中。微博，即微博客(microblog)的简称，是一个基于用户关系的信息分享、传播及获取平台。由于微博可以用来传播实时消息、发布新闻广告等，所以越来越受到人们的关注[1]。在信息多样化的今天，如何能够从海量信息中获取有用的信息并进行新的话题发现，是当今学者研究的热点之一。如今，微博信息数量以指数级的形式迅速增加，给大众带来实时消息的同时也增加了信息的冗余和噪声以及微博话题发现的难度[2]。因此，进行精确而快速的话题发现，不但能够对微博平台进行及时监管，营造良好的互联网氛围，还能够及时了解科研趋势并发现有用话题，为科研提供重要信息[3]。

话题发现的相关研究主要采用基于概率模型和基于文本聚类两大类方法。相关研究表明，基于概率模型的方法主要以 LDA (latent dirichlet allocation) 为代表。徐佳俊等人[4]提出了基于 LDA 模型的论坛热点话题识别和追踪，利用 LDA 模型对语料集进行建模，将话题从中抽取出来，根据生成的话题空间找到相应的话题支持文档，计算文档支持率作为话题强度。但是由于 LDA 话题模型作为一种非监督学习方法，不能够使词与词之间的语义信息融合到话题模型当中，从而使得文本聚类效果不佳。随着话题发现的研究不断深入，在基于文本聚类的方法中，由于增量算法不指定簇的约定而作为常用研究方法。刘星星等人[5]采用改进的 Single-Pass 算法设计了一个热点事件发现系统。格桑多吉等人[6]在传统的基于文本聚类的方法上进行改进，提出了一种改的 Single-Pass 聚类算法，该方法主要是利用 Single-Pass 聚类不受指定簇数的限制而进行降维。由于微博文本数据稀疏、噪声大、新词频繁出现等特点，以上方法并没有达到理想的聚类效果。为了快速有效地发现话题，杨菲等人[7]提出基于词共现网络的遗传聚类算法进行话题发现，该方法利用词共现网络从网络文档中提取热点话题。与之前的相关研究相比，该方法在进行话题聚类方面有所提高。

随着词向量的出现，于洁[8]在 Skip-Gram 模型融合词向量投影的微博新词发现一文中指出，采用分布式表示(distributed representation)的方法将文本中的词转换为词向量，其好处是

收稿日期：2016-11-16; 修回日期：2017-01-05 基金项目：国家自然科学基金资助项目(81360230,81560296)

作者简介：宋莉娜(1991-)，女，河南许昌人，硕士研究生，主要研究方向为机器学习、自然语言处理(1071397501@qq.com)；冯旭鹏(1986-)，男，河南郑州人，助理实验师，硕士，主要研究方向为信息检索；刘利军(1978-)，男，河南新乡人，讲师，硕士，主要研究方向为医疗信息服务；黄青松(1962-)，男，湖南长沙人，教授，博士，主要研究方向为智能信息系统、信息检索.

能够使相似的词在距离上更近，并能够充分体现出不同词之间的相关性，且词向量能有效克服自然语言处理中的数据稀疏和维数灾难问题。近年来，随着深度学习的不断发展，神经网络开始走进人们的视野，尝试用神经网络进行文本聚类以提高聚类效果。但是 SOM 网络初始权值选择复杂、聚类时间较长[9]。

针对以上提出的问题，本文提出一种基于 SOM 聚类的微博话题发现。该方法利用词向量模型进行文本特征提取，能有效解决传统聚类方法带来的维度过高等问题，采用极小值原理选择初始输入节点，通过对 SOM 聚类过程的改进，使得聚类效果更佳。

1 基于 SOM 聚类的微博话题发现方法

SOM 聚类算法的微博话题发现方法主要分为获取原始语料、SOM 话题聚类、聚类结果分析三个阶段。其流程如图 1 所示。

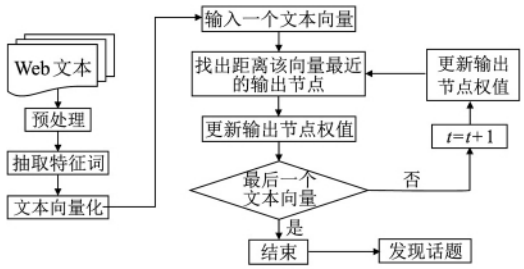

图1 话题发现的流程

获取原始语料部分包括：首先用爬虫方法获取微博文本，然后对文本进行预处理，最后采用分词系统进行分词。SOM 话题聚类部分首先采用词向量模型对文本进行向量化；然后选择符合的文本进行聚类，聚类过程中的文本相似度采用基于词向量的文本相似度进行计算；最后对聚类结果进行聚类分析，根据聚类分析的结果进行话题发现。

1.1 基于词向量的文本相似度计算

在自然语言处理中，文本相似度计算是十分基础的工作，首先需要将文本转换为一种数值型的数据结构，才能给计算机进行计算，词与词之间的相似性是通过词之间的距离来衡量，距离越小代表词越相似[10]。传统的聚类方法主要采用 TF-IDF 计算词与词之间的权重来衡量两者间的相似度，利用该方法进行计算时不能充分利用同义词或者近义词之间的关系，当词汇数目增大时，向量的维数会随着增高，从而加大计算量。词向量具有良好的语义特性，是表示词语特征的常用方式。其每一维的值代表一个具有一定的语义和语法上解释的特征，故可以将词向量的每一维称为一个词语特征。目前，常用于文本相似度计算的方法[11]有基于共现词的相似度计算、基于语义词典的相似度计算和编辑距离三种。微博文本除了内容短小、噪声大等特点，还出现了新词和缩写词等新的特点外，所以传统的单靠计算词汇之间的权值去衡量词汇之间的相似度是不够准确的。由于新词的不断出现，语义词典中不可能涵盖所有的新词，所以本文采用基于词向量和共现词的相似度计算、词向量和编辑距离的相似度计算，两者相结合的方法计算词汇之间的相似度。基于词向量和共现词的相似度计算公式如下：

$$\cos(\alpha,\beta) = \frac{\sum_{k=1}^{n}\alpha_k \times \beta_k}{\sqrt{\sum_{k=1}^{n}(\alpha_k)^2} \times \sqrt{\sum_{k=1}^{n}(\beta_k)^2}} \tag{1}$$

$$\text{sim1}(M,N) = \frac{\sum_{\alpha\in M}\min(\max(\rho\times\cos(\alpha,N)),1)}{|M|+|N|} + \frac{\sum_{\alpha\in N}\min(\max(\rho\times\cos(\alpha,M)),1)}{|M|+|N|} \tag{2}$$

其中：$\cos(\alpha,\beta)$ 表示词向量 α、β 夹角余弦值；n 为词向量长度；$\text{sim1}(M,N)$ 代表两个向量的相似程度；$\max(\rho\times\cos(\alpha,N))$ 是计算句子 N 中所有词对应词向量与词向量 α 的余弦相似度的最大值；M 是输入文档；N 是聚类中的文档；ρ 是调整词向量间余弦值的系数，目的是为了减少不同词向量计算的相似度浮动以减小神经元的误差，规定调整后的结果为 $[-1,1]$。基于词向量和编辑距离的相似度计算公式如下：

$$\text{sim2}(M,N) = 1 - \frac{\cos(\alpha,\beta)}{\max(|M|,|N|)} \tag{3}$$

$$\text{sim} = p\times\text{sim1}(M,N) + q\times\text{sim2}(M,N) \tag{4}$$

其中：p 和 q 为两个相似度值的概率系数，要求 p 和 q 满足 $p+q=1$。依据文献[12]，实验如表 1 所示，本文取“奥运会”和“运动员”为例进行相似度计算，相似度大的认为计算效果最好。由于“奥运会”和“运动员”是所属关系，然而在自动分词以后，计算两者相似度并不能达到设定的阈值要求，本文认为两者应当聚为一类，所以需要选择能使相似度最高的系数值。

表 1 概率系数系数 p 和 q 的取值调整

p	q	sim	p	q	sim	p	q	sim	p	q	sim
0.05	0.84	0.73	0.55	0.45	0.85	0.3	0.7	0.73	0.8	0.2	0.69
0.1	0.9	0.76	0.6	0.4	0.62	0.35	0.65	0.62	0.85	0.15	0.81
0.15	0.85	0.82	0.65	0.35	0.74	0.4	0.6	0.76	0.9	0.1	0.63
0.2	0.8	0.71	0.7	0.3	0.78	0.45	0.55	0.81	0.95	0.05	0.71
0.25	0.75	0.90	0.75	0.25	0.90	0.5	0.5	0.79	1.0	0	0.83

由表 1 可知，当 $p=0.75$，$q=0.25$ 时，或者 $p=0.25$，$q=0.75$ 时，能够达到预期的实验效果。

1.2 SOM 聚类算法

文本聚类就是通过计算文本间的相似度，将文本划分成若干个不同的类，使得同一类中文本尽可能相似，不同类中文本尽可能不同。SOM 网络是由 Kohonen 教授在 1981 年提出的对神经网络的数值模拟方法，是人工神经网络的重要分支之一[13]。自组织特征映射神经网络(self-organizing feature map)，简称 SOM 网络，主要用于解决模式识别类问题，属于无监督学习算法。传统的 SOM 网络只有两层，即输入层和输出层。输入层是一维神经元，其节点是样本维数；输出层的神经元处于二维平面网格节点上，构成一个二维矩阵。输入层通过权重向量将训练数据集传递到输出层各单元。输入层与输出层的组织结构如图 2 所示。

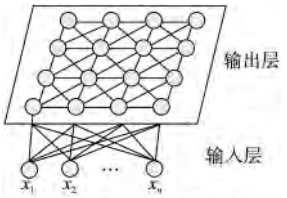

图2 SOM的组织结构

SOM 的训练过程分为竞争、合作和权值调整三个阶段。竞争即为寻找获胜节点的过程，找出判别函数值最大的神经元。传统的 SOM 训练过程中，初始聚点和初始权值的选择会对聚类结果产生一定的影响。本文在传统的 SOM 神经网络的输入层与输出层之间加一个过滤层，加入过滤层的目的是筛选符合要求的文本文档，对实验语料进行调整，减少文本之间的

数据冗余。其结构如图 3 所示。

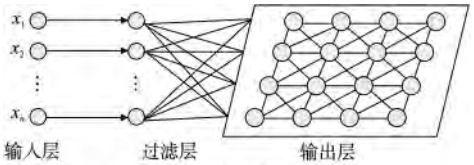

图3　改进的SOM组织结构

在 SOM 中，每个神经元对应一个与输入向量有相同维度的权重向量，根据神经元的输入节点找到获胜节点，根据这两个节点计算出获胜神经元的邻域。神经元的这种相邻关系反映了低维度中的空间聚类结构。根据学习率和输入神经元的值，SOM 能够自动调整输出神经元的加权向量，这样便能够使神经元之间的加权向量更加接近。但是由于输入神经元的初始向量值不同，从而使输出神经元之间具有不同的权重向量。近年来基于 SOM 聚类方法的改进层出不穷，改进的宗旨主要集中在网络拓扑和权重更新上。目前对 SOM 进行改进的方法主要有两种思路：一种是修改网络的神经元模型以适应复杂的数据结构[14]；另一种是将复杂的结构数据分成两个分量，然后对分离分量进行基于 SOM 的聚类分析[15]。本文认为基于神经拓扑距离的权重更新用于这种复杂的输入数据会导致数据计算冗余增大。例如，文本字符串之间的编辑距离可用于区分邻域，但是这样会得到新的权重致使权重方程进行不断的更新，加大了计算量。由文献[16]可知，大量的神经细胞可以在高维状态空间产生复杂的网络行为，神经网络的神经活动对初始值的依赖性很强，这种依赖性的原理涉及到一种最小化的函数取值，然而很难找到一种适用于全局化最优的最小值方法。因此，需要一种求解局部最小值的方法，以减少这种依赖性。为了解决以上问题，本文采用极小值法对 SOM 网络聚类的初始输入向量进行选择。由于聚类过程中具有相同极小值的两篇文章有可能不属于同一类，所以可采用极小值法寻找初始的输入向量。初始的输入文本向量选择原理是：首先选择 sim 值最小的两个向量 d_1 和 d_2 作为初始输入向量，剩下的输入文本向量选择依据式(5)作为选择。

$$d_{m+1} = \underset{d_r \notin S}{\arg\min}(\rho \times \cos(d_m, d_r)) \tag{5}$$

式(5)表示训练集中已经选择了 m 个输入向量。其中：d_{m+1} 表示第 $m+1$ 个输入向量的选择方式；S 代表已经选择过的向量；d_m 为已经选择过的向量；d_r 表示没有选择过的向量。

文献[17]认为 RBF 神经元模型具有局部性和高度非线性等特征，这些特性使 RBF 型神经元模型可以与无监督的聚类方法组合，使得自组织网络聚类效果更加理想。在自组织聚类方法中需要通过计算神经元的初始权重向量与输入向量之间的距离来找到最佳匹配神经元，从而更新权重。当使用 RBF 型神经元模型时，最佳匹配神经元被简单地定义为具有最大激活的神经元。从这个角度来看，大量的 RBF 神经元可被认为是输入数据的特征表示。因此，本文选择以 RBF 神经元模型与极小值原理相结合的方法选择符合条件的神经元进行聚类。相结合的原理如下：

a) 依据极小值原理选择初识神经元。

b) 第二个以及以后的神经元的选择有两种：

如果采用极小值原理选择的输入神经元不能使 SOM 聚类的聚类半径或者权值为发生改变，而采用 RBF 神经元模型选择的输入神经元时聚类半径发生了改变，本文选择 RBF 神经元模型进行输入神经元的选择；相反，如果采用极小值原理选择的输入神经元能使 SOM 聚类的聚类半径或者权值为发生改变，则采用极小值原理选择的输入神经元。

如果两种方法均能使聚类半径发生改变，本文取变化较大的方法进行输入神经元的选择；如果两种方法均不能使聚类半径发生改变，本文取能够与获胜神经元相似度最大的输入神经元方法。

SOM 算法的具体描述如下：

a) 在输入层上选择初始聚点作为初始输入节点。

b) 给输入节点 i 到输出节点赋一个初始权值，由随机函数产生。

c) 反复计算输入节点到所有输出节点之间的距离，找出最小值 d_i，该距离通常称为欧氏距离，该输出节点称为获胜节点 i。其计算公式为

$$d_i = \min \| x_n - w_i \| \tag{6}$$

其中：x_n 是输入向量，w_i 是节点 i 的权重向量。

d) 更新输出节点 i 的权值，更新权值的公式为

$$w_{ij}(t+1) = w_{ij}(t) + g_{ji^*}(t)(x_i(t) - w_{ij}(t)) \tag{7}$$

其中：$w_{ij}(t)$ 为获胜神经元的权值；$w_{ij}(t+1)$ 为更新的神经元权值；$g_{ji^*}(t)$ 是邻域函数，其计算公式为

$$g_{ji^*}(t) = \alpha(t) \times e^{\frac{-\|r_i - r_j\|^2}{2\partial^2}(t)} \tag{8}$$

其中：r_i 代表获胜节点的位置；r_j 代表神经网络上邻域节点 j 的位置；$\alpha(t)$ 是 t 时刻的邻域半径，是一个随着时间增加而变化的递减函数，其计算公式为

$$\alpha(t) = \alpha(0) \times (1 - \frac{t}{T}) \quad \alpha(0) \in (0,1) \tag{9}$$

e) 重复步骤 c) d)，直到输入样本为最后一个样本。

f) 算法结束。

2　实验设计分析

2.1　实验预料准备

为体现 SOM 聚类的实验效果，实验总体设计分为以下三个部分：a) 实验数据准备工作；b) 相似度系数 ρ 不同取值的设定实验；c) 为验证 SOM 聚类在微博话题发现中具有较好的效果，本文分别将文献[18]基于 LDA + K-means 的聚类方法与文献[19]提出的词向量 + Single-Pass 聚类方法进行对比实验。

本文采用网络爬虫方法爬取微博相关数据，对所爬取的博文进行预处理，包括去停用词、去噪、去除无关符号等处理。由文献[5]可知，采用 ICTCLAS 系统分词精度非常高，并且能够自动对词性进行标注。由于动词和名词对主题表达的贡献率最大[20]，所以只保留名词和动词作为特征词。由文献[5]可知，一个话题的发展历程必然要经历三个阶段，即上升、稳定和下降的过程。本文对 2016 年 8 月 10 ~ 15 号的微博进行爬虫，获取多于 30 000 条的博文数量。随机抽出 6 个话题共 6 778 条微博进行研究，抽取的微博话题和数量如表 2 所示。

2.2　实验评价标准

为了方便与相关研究进行实验对比，本文采用召回率 R、查准率 P、F-Measure 作为实验的评价指标，计算公式分别为

召回率：
$$R = \frac{n_i}{n_i + n_{ij}} \tag{10}$$

查准率：
$$P = \frac{n_i}{n_i + n_j} \tag{11}$$

综合指标:
$$F = \frac{2 \times R \times P}{R + P} \tag{12}$$

其中: n_i 表示与已检测到的与初始话题 i 相关的文档数; n_{ij} 表示没有检测出但与初始话题 i 相关的文档数; n_j 表示检测出的与初始话题 i 不相关的文档数。

表 2 实验语料

序号	话题	数量	序号	话题	数量
1	仙英座流星雨	29	4	王宝强离婚	1 812
2	洪荒之力	1 892	5	里约奥运会	1 952
3	使徒行者	997	6	青岛啤酒节	96

2.3 SOM 神经网络聚类算法的实验结果

2.3.1 相似度系数 ρ 的取值

输出的神经元向量与平均输出向量之间存在一定的误差,其误差公式为

$$\varepsilon_0 = \sum \| \bar{x} - x_i \| \tag{13}$$

其中: $\bar{x}$ 表示输入向量的平均值; x_i 表示活动的神经元的向量值。本文通过对比实验初始误差 ε_0 与调整后的误差 ε 的差值大小来确定相似度系数 ρ 的取值,实验结果如表 3 所示。

表 3 相似度系数 ρ 的取值不同时误差的差值变化

ρ	$\lvert\varepsilon_0 - \varepsilon\rvert$	ρ	$\lvert\varepsilon_0 - \varepsilon\rvert$
0.1	0.002 1	0.3	0.001 7
0.25	0.001 3	0.2	0.001 1
0.15	0.001 4		

由表 3 可知,当 ρ 的取值太大或者太小时,实验初始误差 ε_0 与调整后的误差 ε 的差值都比 ρ 取 0.2 时要大,达不到误差修正的目的,从而影响聚类效果。因此,为了减小实验误差,ρ 的取值应为 0.2。

由于微博文本自身的数据稀疏性,提供的信息量较少,所以 LDA + K-means 模型的主题数不同会对文本聚类造成不同的实验结果。本文对不同主题数设置实验对比,其实验结果如图 4 所示。图 4 所示为不同主题数的查准率、查全率和综合指标 F 值的变化。由图 4 可以看出,当主题数不一样时,查准率、查全率和综合指标 F 的值也不相同,但能稳定在一定的区间内。通过实验对比结果可看出,当 LDA 的话题数为 100 时整体性能最好,因此,LDA + K-means 模型的主题数采用 100 来进行实验。

2.3.2 改进的 SOM 的聚类实验结果

本文对基于 SOM 聚类的微博话题发现方法进行实验。首先对微博文本进行爬虫,然后对文档进行预处理,包括去噪、去停用词以及去除无关符号等处理。采用 ICTCLAS 系统对进行过预处理的文档进行分词,并对词性进行标注。将处理过的实验语料进行 SOM 神经网络聚类,并分别与 LDA + K-means 方法、词共现网络 + Single-Pass 方法进行对比,其实验结果如图 5、6 所示。图 5 所示为实验语料使用三种方法进行查准率的对比实验。从图 5 中可以看出,采用词向量的方法进行实验的效果明显优于 LDA 模型,由于词向量可以解决向量维度过高问题,不会因为话题数量的多少而使准确率发生太大的变化。本文采用改进的 SOM 网络对实验语料进行聚类的效果优于 K-means 和 Single-Pass 方法,原因是 K-means 进行聚类时需指定聚类中心,而 Single-Pass 方法进行聚类时则会受到文档输入顺序的影响。改进的 SOM 聚类时除了不受以上两种问题的干扰之外,还由于输入层后面的过滤层直接去掉了不符合要求的文档,提高了聚类准确度。图 6 所示为实验语料采用三种方法进行平均值查准率、查全率和综合指标 F 值的对比实验。由图 6 可以看出,采用基于 SOM 神经网络聚类的微博话题发现方法效果优于其他两种方法,该方法较 LDA + K-means 方法的综合指标 F 值高约 10.1%,较词向量 + Single-Pass 方法的综合指标 F 值高约 2.3%。

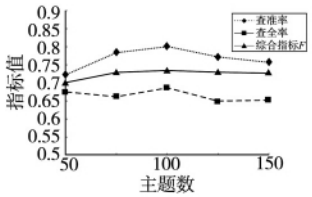

图4 不同主题数的实验对比

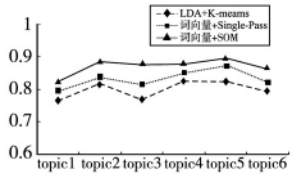

图5 话题查准率的对比实验

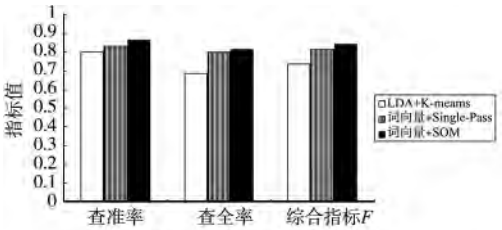

图6 三种方法的对比实验

3 结束语

本文提出的基于 SOM 聚类的微博话题发现方法,针对微博的数据稀疏性、用语不规范、新词出现频繁等特点,采用词向量模型进行文本特征提取,减少了向量维数过高带来的计算量繁重问题; 采用改进的 SOM 进行话题聚类,通过对初始输入向量的选择,提高了聚类的精度。后续工作是研究如何更好地改进算法质量来减少算法复杂度。

参考文献:

[1] Wang Yuan, Liu Jie, Huang Yalou, *et al*. Using hash tag graph-based topic model to connect semantically-related words without co-occurrence in microblogs [J]. IEEE Trans on Knowledge and Data Engineering, 2016, 28(7): 1919-1933.

[2] 贺敏, 王丽宏, 杜攀, 等. 基于有意义串聚类的微博热点话题发现方法 [J]. 通信学报, 2013, 34(z1): 256-262.

[3] 贺亮, 李芳. 基于话题模型的科技文献话题发现和趋势分析 [J]. 中文信息学报, 2012, 26(2): 109-115.

[4] 徐佳俊, 杨陋, 姚天昉, 等. 基于 LDA 模型的论坛热点话题识别和追踪 [J]. 中文信息学报, 2016, 30(1): 43-49.

[5] 刘星星, 何婷婷, 龚海军, 等. 网络热点事件发现系统的设计 [J]. 中文信息学报, 2008, 22(6): 80-85.

[6] 格桑多吉, 乔少杰, 韩楠, 等. 基于 Single-Pass 的网络舆情热点发现算法 [J]. 电子科技大学学报, 2015, 44(4): 599-604.

[7] 杨菲, 黄伯雄. 词共现网络的遗传算法在话题发现中的应用 [J]. 计算机工程与软件, 2013, 49(14): 126-129.

[8] 于洁. Skip-Gram 模型融合词向量投影的微博新词发现 [J]. 计算机系统应用, 2016, 25(7): 130-136.

[9] 刘铭, 刘秉权, 刘远超. 面向信息检索的快速聚类算法 [J]. 计算机研究与发展, 2013, 50(7): 1452 -1463.

[10] 方延风, 陈健. 基于词向量距离的相关词变迁研究——以《情报探索》杂志摘要为例 [J]. 情报探索, 2015(4): 5-7, 10.

[11] 郭胜国, 郭丹丹. 基于词向量的句子相似度计算及其应用研究 [J]. 现代电子技术, 2016, 38(13): 99-107.

[12] Zhao Jingling, Zhang Huiyun, Cui Baojiang. Sentence similarity based on semantic vector model [C] //Proc of the 9th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing. 2014: 499-503.

(下转第 679 页)

$$\overline{T}=\frac{\sum_{i=1}^{N}T_i}{N} \tag{10}$$

其中: N 表示实验次数,本文 N 取 10; T_i 表示每次独立实验所消耗的时间,为方便获取,本文直接以程序执行时间来表示。各类算法平均执行时间如表 4 所示。

表 4　10 次实验各类算法平均执行时间　/s

数据集	原始 K-means 串行算法	原始 K-means 并行算法	文献[6] PSO-Kmeans 并行算法	文献[19] ACS-Kmeans 串行算法	本文算法
A	13.53	29.42	27.92	12.98	28.67
B	48.89	132.52	121.53	44.32	123.56
C	1 325.12	782.23	688.32	1 252.34	652.89
D	2 892.56	1 428.31	1 203.71	2 366.38	1 138.33
E	N/A	9 721.66	7 892.76	N/A	7 523.99

由表 4 可知,对于五组样本数递增的随机数据集,当样本数较少时,Hadoop 分布式平台的并行算法处理效率比单机串行算法效率低; 但当样本数增多时,Hadoop 分布式平台的并行处理效率逐渐高于单机串行算法,特别是对于数据集 E,串行算法的处理效率过低,程序已不能在本次实验机器上正常执行(内存溢出),这是由于样本较少时,Hadoop 分布式平台需要不断地读写和传输数据占用较多时间,实际计算时间占比较小[20],而单机的串行算法不需要与其他机器交互所以效率更高; 但当样本数量达到一定规模时,单机系统资源有限,所以串行算法执行时间很长,而 Hadoop 分布式平台并行处理由于可利用的资源更多所以效率表现更好。并且与文献[6]中提出的 PSO-Kmeans 并行算法相比,在数据量较大时本文算法执行耗时更少。

4　结束语

本文提出了一种基于自适应布谷鸟搜索的并行 K-means 聚类算法,解决了原始 K-means 聚类算法全局搜索能力差,以及在样本数据量较大时单机串行环境下效率低等问题。通过在 Hadoop 分布式计算平台上进行实验对比分析,结果表明相对于原始 K-means 算法和基于粒子群优化算法的 K-means 算法,本文改进算法的聚类准确性和大数据情景下的执行效率均有所提高。但本文研究也存在一些局限性,样本数据间仅考虑了欧氏距离作为 K-means 聚类算法的测度,未考虑集群节点数量对算法效率的影响,初始聚类中心采用随机获取的方式会影响算法的稳定性,算法还有优化的空间。总体来说,布谷鸟搜索算法作为一种新的元启发式群体智能与仿生算法,不仅可以运用于 K-means 聚类算法的改进,而且为如何从海量数据中快速准确地发现有用信息提供了一种新的研究思路。

参考文献:

[1] Han Jiawei, Kamber M, Pei Jian, *et al*. Data mining concepts and techniques [M]. 3rd ed. San Francisco: Morgan Kaumann, 2011: 451-456.

[2] 汪中,刘贵全,陈恩红. 一种优化初始中心点的 K-means 算法 [J]. 模式识别与人工智能,2009,22(2):299-304.

[3] 李春生,王耀南. 聚类中心初始化的新方法 [J]. 控制理论与应用,2010,27(10):1435-1440.

[4] 陶新民,徐晶,杨立标,等. 一种改进的粒子群和 K-均值混合聚类算法 [J]. 电子与信息学报,2010,32(1):92-97.

[5] Lu Bin, Ju Fangyuan. An optimized genetic K-means clustering algorithm [C] //Proc of International Conference on Computer Science and Information Processing. Piscataway: IEEE Press, 2012: 1296-1299.

[6] 马汉达,郝晓宇,马仁庆. 基于 Hadoop 的并行 PSO-Kmeans 算法实现 Web 日志挖掘 [J]. 计算机科学, 2015,42(s1):470-473.

[7] Yang Xinshe, Deb S. Cuckoo search via Lévy flights [C] //Proc of World Congress on Nature & Biologically Inspired Computing. 2009: 210-214.

[8] 郑洪清,周永权. 一种自适应步长布谷鸟搜索算法 [J]. 计算机工程与应用,2013,49(10):68-71.

[9] Raveendra. DE based job scheduling in grid enviroments [J]. Journal of Computer Networks, 2013, 1(2): 28-31.

[10] 陈乐,龙文. 求解工程结构优化问题的改进布谷鸟搜索算法 [J]. 计算机应用研究,2014,31(3):679-683.

[11] Fister I, Yang Xinshe, Fister D, *et al*. Cuckoo search: a brief literature review [M] //Cuckoo Search and Firefly Algorithm, Volume 516 of the Series Studies in Computational Intelligence. Berlin: Springer-Verlag, 2013: 49-62.

[12] Yang Xinshe, Deb S. Cuckoo search: recent advances and applications [J]. Neural Computing and Applications, 2014, 24(1): 169-174.

[13] 欧阳喆,周永权. 自适应步长萤火虫优化算法 [J]. 计算机应用,2011,31(7):1804-1807.

[14] White T. Hadoop: the definitive guide [M]. 4th ed. Sebastopol: O'Reilly Media, 2015: 3-15.

[15] Ghemawat S, Gobioff H, Leung S T. The Google file system [C] //Proc of the 19th ACM Symposium on Operating Systems Principles. 2003: 19-43.

[16] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters [C] //Proc of Conference on Symposium on Opearting Systems Design & Implementation. 2004: 107-113.

[17] Chang F, Dean J, Ghemawat S, *et al*. Bigtable: a distributed storage system for structured data [C] //Proc of USENIX Symposium on Operating Systems Design and Implementation. [S.l.]: USENIX Association, 2006: 15.

[18] 喻金平,郑杰,梅宏标. 基于改进人工蜂群算法的 K-均值聚类算法 [J]. 计算机应用,2014,34(4):1065-1069,1088.

[19] 杨辉华,王克,李灵巧,等. 基于自适应布谷鸟搜索算法的 K-means 聚类算法及其应用 [J]. 计算机应用,2016,36(8):2066-2070.

[20] 周婷,张君瑛,罗成. 基于 Hadoop 的 K-means 聚类算法的实现 [J]. 计算机技术与发展,2013,23(7):18-20.

(上接第 674 页)

[13] 刘芳. 基于 SOM 聚类的可视化方法及应用研究 [J]. 计算机应用研究, 2012,29(4):1300-1303,1306.

[14] Gärtner T. A survey of kernrls for structured data [J]. ACM SIGKDD Explorations Newsletter, 2003,5(1):49-58.

[15] Hammer B, Micheli A, Sperduti A, *et al*. Recursive self-organizing network models [J]. Neural Networks, 2004,17(8):1061-1085.

[16] Tsutsumi K, Nakajima K. Maximum/minimum detection by a module-based neural network with redundant architecture [C] //Proc of International Joint Conference on Neural Networks. 1999: 558-561.

[17] Deng Zhidong, Mao Chengzhi, Chen Xiong. Deep self-organizing reservoir computing model for visual object recognition [C] //Proc of International Joint Conference on Neural Networks. 2016: 1325-1332.

[18] Qiu Lin, Xu Jungang. A Chinese word clustering method using latent dirichlet allocation and K-means [C] //Proc of the 2nd International Conference on Advances in Computer Science and Engineering. 2013: 267-270.

[19] Yan Danfeng, Hua Enzheng, Hu Bo. An improved single-pass algorithm for Chinese microblog topic detection and tracking [C] //Proc of IEEE International Congress on Big Data. 2016: 251-258.

[20] 郑飞,张蕾. 基于分类的中文微博热点话题发现方法研究 [C] //第 29 次全国计算机安全学术交流会论文集. 2014: 311-314.