DOI:10.19789/j.1004-9398.2015.04.004

基于改进 K-means 算法的 BBS 热点话题发现

马国栋 李 慧
(首都师范大学教育技术系,北京 100048)

摘 要

详细介绍了经典划分式聚类算法 K-means 的特点,针对该算法中"聚类中心难以确定"的不足提出"选择最优聚类质心"的改进方案,改进后的算法得到了可靠的聚类质心. 最后,基于"中华网 BBS"文本数据对改进后的算法进行验证、分析,并与经典 K-means 算法聚类结果进行对比,得到了更为理想的实验结果,表明算法可行有效.

关键词: K-means,文本聚类,热点话题发现.
中图分类号: G35

0 引 言

BBS(Bulletin Board System,也称论坛)已经成为普通民众表达个人观点和意愿的主流媒体. 与传统的传播媒介(电视、报纸)相比,BBS 具有更强的开放性和互动胜:用户通过注册 ID 可以表达自己的观点、参与讨论并能获取多元化信息,还能找到其他达成共识的用户. 由于用户的注册 ID 具有一定的匿名性,因此其言论一般都大胆、率真,往往最能表达民众的真实观点[1]. 但由于 BBS 通常缺少有效的言论限制和舆论监督,一些对社会问题认识偏激的用户可能在 BBS 中发表帖子来宣泄不满或表达错误的价值观,如果不及时发现这些不安全舆论并对其合理引导,一个社会热点事件的出现加上这些片面、偏执、煽动人心的不良言论,就可以成为点燃一起舆论战争的导火索,轻则对青少年网民造成负面影响,重则可能破坏民族团结、危害国家安全[3].

话题检测与跟踪(Topic Detection and Tracking,简写为 TDT)由美国国防高级研究计划署(DARPA)发起倡议,主要用于在新闻报道流中发现和跟踪新生事件,从事新闻报道的自动识别、锁定和发现突发性新闻话题、跟踪话题发生发展过程等相关任务. 随着网络信息爆炸时代的到来,TDT 的思想和技术被应用到网络热点话题发现、跟踪领域[2].

收稿日期:2014-04-24

本文针对经典聚类算法 K-means 的不足之处提出改进方案,基于 BBS 数据对改进前后的 K-means 算法做出对比,并详细分析了实验结果.

1 K-means 聚类算法

1.1 经典 K-means 算法

K-means 算法是一种最常用的划分式聚类算法,算法须预先指定聚类个数 k,且每一个聚类有一个聚类中心[4].

对于给定文档对象集合 $D=\{d_1,d_2\cdots d_n\}$ 和聚类数 k,K-means 算法从文本中任意选择 k 个文档对象作为初始的聚类中心 $C=\{C_1,C_2,C_3\cdots C_k\}$. 对文档集合 D 中的每个文档对象 di,依次计算它与各个聚类中心 Ci 的相似度 $\mathrm{sim}(di,cj)$,并将 di 赋给与其最相似的聚类 $\arg\max \mathrm{sim}(di,cj)$. 然后重新计算每个聚类中文档对象的平均值,更新聚类中心. 若用向量空间模型来表示文档对象,文档对象的平均值表示文档所对应的向量上所有维的平均值. 循环前二步,直到聚类结果不再改变或变化小于指定阈值.

K-means 算法是一个简单高效的聚类算法,通常只需迭代数次就可达到收敛. 时间复杂度为 $O(n)$. K-means 算法受聚类数 k 和初始聚类中心的影响较大,聚类数的选取直接决定了识别的话题数;初始聚类中心的选取直接影响到 K-means 算法的聚类结果[5]. K-means 算法的基本流程如图 1 所示.

K-means 算法非常易于理解,算法时间复杂度

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

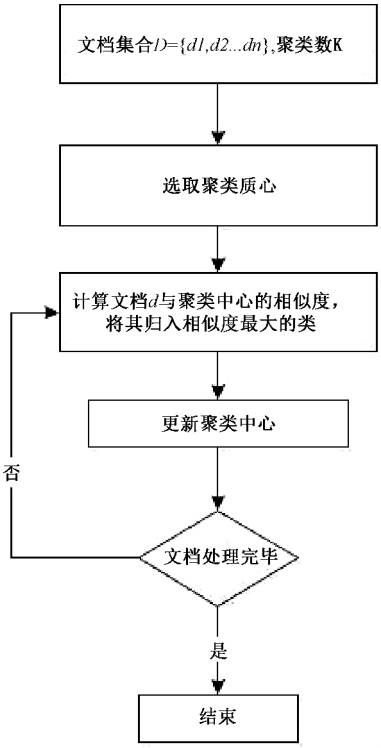


图 1　K-means 聚类算法流程图

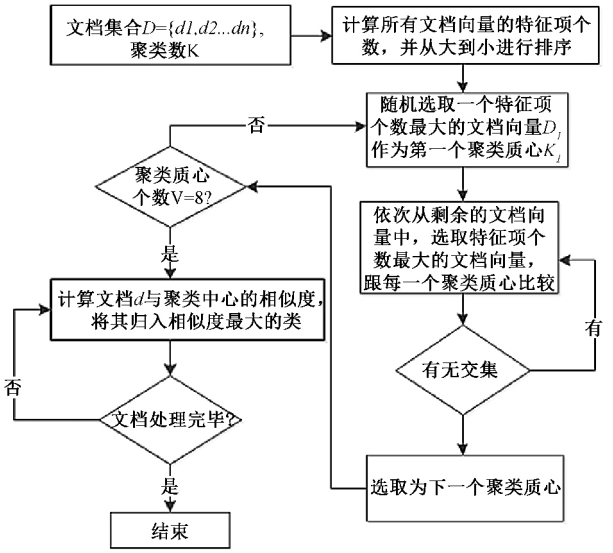


图 2　改进后的 K-means 算法流程图

较低,但算法存在一定缺陷:

1. K-means 算法属于划分算法,算法要求在运行前给定聚类的个数; 这个问题至今也没有好的解决方案,不过 K-means 算法本身应用于已知聚类个数的文档集,如果聚类个数未知,那么一般采用 Singlg-Pass 或者其他增量聚类算法.

2. 在指定聚类个数 k 的同时,要求指定聚类内的质心,其他未划分聚类的文档必须和指定的聚类质心作相似度比较,才能得到聚类结果. 这样人为干预的因素过大,而且聚类质心的选取很大程度上影响了聚类的结果.

1.2　改进的 K-means 算法

K-means 算法受聚类数 k 和初始聚类中心的影响较大,聚类数的选取直接决定了识别的话题数; 初始聚类中心的选取直接影响到 K-means 算法的聚类结果. 本文利用初始给定的聚类数 k,解决了 K-means 算法中聚类中心的选取问题,算法流程图如图 2.

改进的 K-means 聚类中心选取,主要过程有以下几步.

1. 根据特征词的数目,从大到小对文档向量排序,排序后的文档向量存储在集合 $S=\{v_1,v_2,v_3\cdots v_n\}$; 存放聚类质心的集合 $C=\{\}$;

2. 选取 v_1 作为第一个聚类的中心,$C=\{v_1\}$;

3. 设定一个值 $n=0$. 从 v_2 开始,依次比较集合 S 中每一个元素与 C 中的每一个元素是否存在相同特征词; 如果存在相同的特征词,则从 S 中删除该元素,如果不存在,则以此元素为第二个聚类中心放入集合 C 中,则同时 $n=n+1$; 直到 $n=k$(聚类个数) 为止.

4. 当 $n=k$ 时,聚类质心的集合 $C=\{v_1,v_2,v_3\cdots v_k\}$.

2　基于聚类的 BBS 热点话题发现

2.1　评估指标

对于文本分类算法,测试时常用的评测指标包括丢失率、错误率、准确率、召回率、F1 值等,其定义如下:

丢失率(Miss, P_{miss}): 指应该属于某聚类,却没有被归类到该聚类中的数据.

$$P_{miss}=\frac{n^i_{miss}}{n^i}$$

错误率(False, P_{false}): 是指本来不属于某聚类,却被归类到该聚类中的数据.

$$P_{false}=\frac{n^i_{false}}{n-n^i}$$

本实验中, n^i 代表语料库中话题n^i 实际相关帖子数,n 代表语料库中所有帖子数. 在话题识别算法后,可以得到关于话题 i 的相关帖子数,从而得知系统未识别出话题 i 的帖子数n^i_{miss} .

准确率(Precision, P) : 是指系统正确的结果占

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

所有抽取结果的比例.

$$P = \frac{\text{算法发现的相关文档数}}{\text{系统返回的总文档数}} * 100\%$$

召回率(Recall,R):是指系统正确的结果占所有可能正确结果的比例.

$$R = \frac{\text{算法发现的相关文档数}}{\text{相关总文档数}} * 100\%$$

Fl(Test,T):是准确率和召回率充分考虑后,找到一个折中的评测标准. 因为准确率和召回率针对不同方面进行评测,这两个评测结果都是小于 1 的小数.

$$F1 = \frac{P * R * 2}{P + R}$$

2.2 聚类实验结果

本文 BBS 实验数据集选取了"中华网论坛"中 8 个主题类别的发帖回帖,8 个主题分别是"郭美美与红十字会"、"朝鲜半岛"、"第一夫人彭丽媛风采"、"国五条细则落地"、"撒切尔夫人去世"、"斯诺登逃往何处"、"四川雅安地震"和"中国航母追踪". 分别从每一个主题类别采集了 45 篇左右的帖子,共计 348 篇,每一篇帖子的发帖时间从 2013 年 3 月到 2013 年 11 月不等.

2.3 经典 K-means 算法实验

应用原始的 K-means 算法作为热点事件发现的聚类算法,采用 348 篇实验数据集,得到的结果如表 1 所示.

表 1 原始 K-means 算法的实验数据集结果

初始主题数	8							
初始主题	朝鲜半岛战火	第一夫人彭丽媛	国五条细则落地	红十字会与郭美美	英国撒切尔葬礼	斯诺登逃亡何处	四川雅安地震	中国航母追踪
初始文档数	40	40	45	47	47	40	47	42
聚类后主题数	8							
聚类后相关主题	朝鲜半岛战火	第一夫人彭丽媛	国五条细则落地	红十字会与郭美美	英国撒切尔葬礼	斯诺登逃亡何处	四川雅安地震	中国航母追踪
聚类后文档数	59	50	44	53	41	31	34	36
相关文档数	40	37	44	47	41	31	34	36
聚类文档差值	+19	+10	-1	+6	-6	-9	-13	-6

由表 1 可知,"朝鲜半岛战火"中多了 19 篇帖子,分别来自于"第一夫人彭丽媛"的 3 篇图片帖、"英国撒切尔葬礼"的 2 篇图片帖、"四川雅安地震"的 7 篇图片帖、"斯诺登逃往何处"的 2 篇图片帖、"中国航母追踪"的 5 篇图片帖. 所有的图片帖因为与所有聚类质心的相似度都为 0,所以被默认分配给第一聚类.

"红十字会和郭美美"版块多了 6 篇帖子,来自于"四川雅安地震". 这 6 篇帖子的主题都与地震和红十字会相关,相似度要高于"四川雅安地震"版块的聚类中心.

"第一夫人"版块中 3 篇图片帖划分在"朝鲜半岛战火",而来自"国五条细则落地"、"英国撒切尔葬礼"、"斯诺登逃往何处"和"中国航母追踪"版块中的 13 篇帖子被划分在"第一夫人"聚类. 原因在于"第一夫人"的聚类中心特征项中包含"英国、美国、北京"等主题词,这些被错分的帖子与此聚类中心的相似度较高. 表 2 是原始 K-means 聚类结果的评价指标.

表 2 原始 K-means 聚类结果的评价指标

实验数据集(348)	朝鲜半岛战火	第一夫人彭丽媛	国五条细则落地	红十字会与郭美美	英国撒切尔葬礼	斯诺登逃亡何处	四川雅安地震	中国航母追踪
召回率	1.00	0.93	0.98	1.00	0.87	0.77	0.72	0.86
准确率	0.65	0.95	0.89	0.88	0.89	0.83	0.87	0.90
F1 测试值	0.83	0.94	0.93	0.94	0.88	0.80	0.79	0.88

从表 2 可知,聚类结果的评价指标差别较大,平均召回率是 89%,平均准确率 86%,实验结果较差. 说明原始 K-means 算法聚类中心的选取随机性强,且代表性不足(特征项较少),这是本文对原始 K-means 算法改进的重点.

2.4 改进的 K-means 聚类算法

应用改进后的 K-means 算法作为热点事件发现的聚类算法,采用 348 篇实验数据集,得到的结果如表 3 所示.

本实验在 K-means 聚类之前,首先选择出了能够最大程度地代表各个聚类的聚类中心,通过表 3 可以看出,聚类结果基本没有任何差错. 而"红十字会与郭美美"版块少了 1 篇帖子,该帖子虽然出现了原版块的特征词,但与"四川雅安地震"版块的特征词更加匹配,其内容是"郭美美"事件的衍生帖;这篇帖子讨论的话题是关于"四川、雅安、重建",所以被划分在"四川雅安地震"版块中. 改进 K-means 聚类结果的评价指标如表 4 所示.

从表 4 可知,聚类间的评价指标差别不大,平均召回率是 99%,平均准确率 99%,实验结果较好,说明改进后的 K-means 算法聚类中心代表性较强. 表

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 3　改进 K-means 算法的实验数据集结果

初始主题数	8							
初始主题	朝鲜半岛战火	第一夫人彭丽媛	国五条细则落地	红十字会与郭美美	英国撒切尔葬礼	斯诺登逃亡何处	四川雅安地震	中国航母追踪
初始文档数	37	39	37	40	45	27	28	37
聚类后主题数	8							
聚类后相关主题	朝鲜半岛战火	第一夫人彭丽媛	国五条细则落地	红十字会与郭美美	英国撒切尔葬礼	斯诺登逃亡何处	四川雅安地震	中国航母追踪
相关文档数	37	39	37	39	45	27	29	37
不相关文档数	0	0	0	0	0	0	0	1
前后文档差值	0	0	0	-1	0	0	+1	0

表 4　改进 K-means 聚类结果的评价指标

实验数据集(348)	朝鲜半岛战火	第一夫人彭丽媛	国五条细则落地	红十字会与郭美美	英国撒切尔葬礼	斯诺登逃亡何处	四川雅安地震	中国航母追踪
召回率	1.00	1.00	1.00	0.97	1.00	1.00	1.00	0.97
准确率	1.00	1.00	1.00	0.99	1.00	1.00	0.99	0.97
F1 测试值	1.00	1.00	1.00	0.99	1.00	1.00	0.98	0.98

12 是改进 K-means 算法的验证数据集结果.

3　总结

本文详细介绍了经典划分式聚类算法 K-means 的特点,主要针对该算法中“如何制定聚类中心”的不足提出“选择最优聚类质心”的改进方案,改进后的算法得到了可靠的聚类质心. 最后,基于“中华网 BBS”文本数据对改进后的算法进行验证、分析,并与经典 K-means 算法聚类结果进行对比,得到了更为理想的实验结果,表明算法可行有效.

参 考 文 献

[1] 胡静. Web 文本挖掘中数据预处理技术研究[J]. 现代计算机, 2009(2):48-50.

[2] Yang Y, Pierce T, J Carbonell J. A Study on Retrospective and Online Event Detection[C]. The 21st Ann Int ACM SIGIR Conference.

[3] 话题发现 homepage. http://www.itl.nist.gov/iad/894.01/tests/话题发现/index.htm.

[4] 王旭东. 基于 Web 的信息抽取技术研究[D]. 西南交通大学硕士学位论文, 2007.

[5] 杨冠超. 热点话题发现策略研究[D]. 杭州: 浙江大学. 2011.1.20-22.

[6] 邱立坤, 程威, 龙志伟,等. 面向 BBS 的话题挖掘初探. 全国第八届计算语言学联合学术会议[C], 南京, 2005: 401-407.

Hot Topic Detecting in BBS with Developed K-Means

Ma Guodong　Li Hui

(Instructional Technology Academy of Capital Normal University, Beijing 100048)

Abstract

Introducing the feature of classic clustering algorithm K-means of that “cluster centers is difficult to determine”. I improved the program for the algorithm of “selecting the optimal clustering centroid”, the improved algorithm has been a reliable the cluster centroid. Finally I got a data results to prove the improved algorithm is reliable.

Key words: K-means, text clustering, hot topic detection.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net