基于 K-Means 改进算法在微博话题发现中的应用研究①

张云伟, 宋安军

(上海海事大学 信息工程学院, 上海 201306)

摘 要: 在传统的 K-means 算法中, 聚类结果很大程度依赖于随机选择的初始聚类中心点以及人工指定的 k 值. 为了提高聚类精度, 本文提出了利用最小距离与平均聚集度来对初始聚类中心点进行选取, 将层次聚类 CURE 算法得到的聚簇数作为 k 值, 从而使聚类精度得到提高. 最后, 将改进后的 K-means 算法应用到微博话题发现中, 通过对实验结果分析, 证明该算法提高了聚类结果精度.

关键词: K-means; 微博; 话题; 聚类

Application of Improved Algorithm Based on K-Means in Microblog Topic Discovery

ZHANG Yun-Wei, SONG An-Jun

(College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China)

Abstract: In the traditional K-means algorithm, the clustering results greatly depend on the random selection of initial cluster centers and the artificial K values. In order to improve the clustering accuracy, this paper proposes to select the initial cluster centers by using the minimum distance and the average clustering degree. The number of clusters is obtained by the hierarchical clustering CURE algorithm as K value, so that the clustering accuracy can be improved. Finally, the improved K-means algorithm is applied to the micro-blog topic discovery. Through the analysis of the experimental results, it is proved that the algorithm can improve the accuracy of clustering results.

Key words: K-means; microblog; topic; clustering

1 引言

随着互联网的快速发展, 微博[1]作为一个信息获取与共享的重要平台, 在广大用户中逐渐的流行开来. 由于每天都有大量的用户在微博上发表和评论信息, 微博的信息量呈现出爆炸性的增长. 随着数据挖掘技术的广泛应用, 可以利用文本挖掘的方法去获取微博中有用的信息, 挖掘出用户讨论的热点话题, 使人们及时了解最新的社会动态, 具有十分重要的现实意义.

针对微博话题发现, 传统的 K-means[2]算法对于初始聚类中心点依赖程度比较高, 聚类结果的精度受初始聚类中心点选取的影响. 为了提高聚类精度, 本文提出了一种初始化中心点选取的改进算法, 利用最小距离与平均聚集度来对初始聚类中点进行选取, 将层次聚类 CURE 算法得到的聚簇数作为 k 值, 将改进后的 K-means 挖掘算法运用到微博话题发现中, 并通过聚类[3]评价方法对改进的 K-means 算法进行评价, 结果表明改进算法准确度有了一定的提高.

2 K-means算法

2.1 K-means 算法思想

K-means 的算法思想是: 随机选取 k 个初始聚类中心点[4], 计算聚类中心点与每个数据对象之间的距离, 找出离数据对象最近的簇, 并把该数据对象加入到最近簇中. 然后调整新簇的聚类中心点, 直到连续两次聚类中心点没发生变化, 则数据对象调整结束, 得出最优的聚类结果.

2.2 K-means 算法流程

对于数据集 $X=\{x_1,x_2,\ldots,x_n\}$, 聚类成 k 个聚簇

① 基金项目:国家自然科学基金(61502298)

收稿时间:2016-02-19;收到修改稿时间:2016-04-11 [doi: 10.15888/j.cnki.csa.005461]

$L_1,L_2,\ldots,L_k$, 每个聚簇 L_k 存在一个质心 p_i, 其中 $p_i=\frac{1}{n_i}\sum_{x\in L_i}x$, n_i 是簇 L_i 中数据点的个数. 聚类效果好坏由函数 S 表示:

$$s=\sum_{i=1}^{k}\sum_{j=1}^{n_i}dist(x_j,p_i) \quad (1)$$

在公式(1)中, $dist(x_j,p_i)$是 x_j 与 p_i 之间的欧式距离. 如果 S 值越小, 表明聚簇中的数据对象越紧凑.

K-means 算法步骤如下:

输入: 数据集 $X=\{x_1,x_2,\ldots,x_n\}$, 聚类数目 k;

输出: k 个聚类 L_j, $j=1,2,\ldots,k$;

步骤 1: 令 $t=1$, 随机选择 k 个数据点作为 k 个类簇的初始聚类中心 $m_j(t)$, $j=1,2,\ldots,k$;

步骤 2: 计算数据对象 x_i 与这 k 个聚簇中心的距离 $dist(x_i,m_j(t))$, $i=1,2,\ldots,n$, $j=1,2,\ldots,k$, 当满足公式(2)时, 则把数据对象 x_i 加入到聚簇 L_j 中;

$$dist(x_i,m_j(t))=\min\{dist(x_i,m_j(t))\} \quad (2)$$

步骤 3: 重新选择 k 个聚簇中心点, 根据公式(3)计算 $m_j(t+1)$, 其中 N_j 是聚簇 L_j 中数据对象的个数;

$$m_j(t+1)=\frac{1}{N_j}\sum_{\substack{i=1\\x_i=p_i}}^{N_j}x_i,\ j=1,2,...,k \quad (3)$$

步骤 4: 如果 $m_j(t+1)\neq m_j(t),j=1,2,...,k$, 令 $t=t+1$, 执行步骤 2, 否则执行步骤 5;

步骤 5: 输出聚类结果, 算法结束.

2.3 K-means 算法的优缺点分析

K-means 算法是一种基于划分的聚类算法, 该算法其优点表现在算法思想简单, 又容易实现, 具有高效性, 对于大多数数据集来说具有一定的伸缩性.

存在这些优点的同时, 该算法也有不足之处需要改进, 传统的 K-means 初始中心点[5]的选取一般都是随机的, 孤立点存在使得聚类结果精度不高.

3 K-means算法的改进

3.1 改进算法基本思想

传统 K-means 算法选取 k 个点作为初始聚类中心点, 然后进行迭代, 聚类结果的精度受初始聚类中心点选取的影响. 为了提高聚类结果的精度, 本文提出根据数据的稀疏性特点, 通过计算各点的聚集度, 结合最小距离和平均聚集度从聚集度比较高的数据中选取 k 个数据对象作为初始化中心点, 来提高聚类结果的准确性.

基本思想: 对于给定的数据集 X, 首先计算出两两数据之间的欧式距离 $dist(x_i,x_j)$, 存放在 N*N 的矩阵中, 每一行存储着一个数据对象和其它数据对象欧式距离, 然后根据聚集度计算公式(4)计算出每一个数据对象的聚集度, 按照聚集度进行升序排列保存在集合 DegData, 然后把小于平均聚集度的数据对象删除, 保留下来的数据对象作为初始化中心点候选集合 D, 选取 k 个数据对象作为 K-means 的初始聚类中心点.

3.2 相关参数定义

定义 1(聚集度). 对于数据集 X 中的样本点 x_i, 离 x_i 距离小于等于 R 的所有样本点的个数, 称之为聚集度, 记作 $Deg(x_i)$:

$$Deg(x_i)=|\{x_i\,|\,dist(x_i,x_j)\le R,x_j\in X\}| \quad (4)$$

在公式(4)中, $dist(x_i,x_j)$代表数据对象之间的欧式距离.

定义 2(邻域半径 R). 对于数据集中样本点聚集程度的划分, 需要考虑到划分半径的大小, 领域半径公式如下:

$$R=\frac{Avg(X)}{n^{releR}} \quad (5)$$

在公式(5)中, $Avg(X)$表示所有数据对象平均距离, n 代表数据集中数据对象的个数, $releR$ 是邻域半径调节系数, $0<releR<1$. 一般的 $releR$ 取值为 0.13, 聚类的效果比较好.

定义 3(欧式距离). 设两个 n 维向量 $X_i=(x_1,x_2,\ldots,x_n)$ 和 $Y_j=(y_1,y_2,\ldots,y_n)$分别代表两个数据对象, 它们之间的欧式距离记为 $dist(X_i,Y_j)$:

$$dist(x_i,x_j)=\sqrt{(x_1,y_1)^2+...+(x_n,y_n)^2} \quad (6)$$

定义 4(平均聚集度). 数据集中有 n 个数据对象, 平均聚集度记作 $Aed(n)$:

$$Aed(n)=\frac{\sum_{i=1}^{n}Deg(x_i)}{n} \quad (7)$$

3.3 算法的一般步骤

输入: 聚类数 k 和数据对象个数为 n 的数据集;

输出: 满足条件的 k 个聚类结果;

步骤 1: 根据欧式距离公式(6)计算出两两数据对象之间的距离, 将结果存入 N*N 矩阵中;

步骤 2: 根据公式(4)算出每一个数据对象的聚集

度, 加入到集合 DegData 中;

步骤 3: 对 DegData 中的数据对象进行筛选, 把小于平均聚集度的元素进行删除, 剩余的数据对象加入到集合 BigDeg 中;

步骤 4: 将集合 BigDeg 中的数据对象从大到小排列;

步骤 5: 从 BigDeg 集合中选取聚集度最大的一个数据对象 z, 加入到集合 C 中, 并删除集合 BigDeg 中的 z;

步骤 6: 根据 N*N 距离矩阵找到距离 z 最近的一个数据对象 o, 如果 $Deg(o)>Aed(n)$则把 o 加入到集合 C 中, 并将 o 从 BigDeg 中删除, 否则不做处理;

步骤 7: 重复步骤 5, 6, 直到集合 C 中有 k 个数据对象, 则选取结束;

步骤 8: 把集合 C 中的 k 个数据对象作为 K-means 聚类算法步骤 1 中的初始聚类中心, 输出聚类结果.

3.4 聚类数 k 值的确定

传统 K-means 算法中, 都是根据事先的经验来选择 k 值, 对于微博话题发现这样的数据集来说, 由于话题数目的不确定性, 不能事先判断 k 的取值大小, 本文利用凝聚型层次聚类[6]CURE 算法, 先对数据集进行初步的聚类, 得到聚类结果类别的个数, 作为改进 K-means 算法的输入值, 来确保在微博话题发现中不至于有太大的偏差.

4 改进的K-means算法在微博中的应用与实验分析

4.1 实验环境

实验采用 PC 机(酷睿 i7CPU, 8G 内存), win7 系统, JDK1.7 运行环境, 以 java 语言实现.

4.2 实验数据的获取以及预处理

本实验所采用的数据是通过爬虫工具, 获取到新浪微博 2015 年 8 月 11 日到 8 月 26 日的部分微博数据 17735 条数据, 对于少于 20 个字的微博进行剔除, 剔除之后为 15493 条数据, 将这些微博数据存入到 Mysql 数据库中作为实验数据集, 获取到的数据格式如表 1 所示.

表 1 实验数据集

编号	内容	url	来源	日期
1	#突发#【天津滨海新区开发区发生爆炸…	http://weibo.c...	头条新闻	8.13
2	【警方证实中国传媒大学失踪女研究生…	http://weibo.c...	天府早报	8.11
3	虽然说我不能反对你们阅兵. 但是他们…	http://weibo.c...	寒夜星空	8.26

对微博数据进行预处理, 对于微博中的特殊符号, 如#, 【】, @等符号通过程序进行剔除, 清洗过数据集之后, 需要对每一条微博进行特征提取[7]. 在特征提取的过程中, 需要先对微博数据进行分词, 本文采用的是中科院研究的 NLPIR 分词系统, 通过分词系统对微博文本进行分词、词性的标注、未登录词识别以及使用哈工大停用词表对停用词进行剔除, 构造出微博文本的特征向量, 接下来对这些特征向量进行聚类分析.

4.3 实验结果

对上面处理过的数据集进行改进的 K-means 算法聚类, 此次聚类结果中得到部分话题如表 2 所示.

表 2 部分话题列表

序号	关键字	话题	微博数量
1	天津|爆炸|事故|消防|求援|受伤	天津爆炸事件	1969
2	中国|抗战|阅兵|反法西斯	抗战阅兵即将举行	1036
3	人民币|贬值|经济|交易|影响	中国人民币贬值	786
4	海航|航班|颠簸|受伤	HU7148 航班颠簸	508
5	阿里巴巴|苏宁|电商|购物	阿里巴巴投资苏宁	335

通过对微博进行挖掘, 可以从一段时间的微博中发现大家所讨论的话题, 把改进的 K-means 算法运用在微博中, 为了验证算法的有效性以及准确性, 我们从数据集中人工筛选出天津爆炸事件、中国人民币贬值两个话题 250 条微博, 再挑选其他 3 个话题的微博 500 条数据加入到测试数据集中, 并对测试数据集中的天津爆炸事件和中国人民币贬值微博进行人工话题标记.

4.4 评价指标

文本聚类领域常用的查准率(Precision)、查全率(Recall)和 F 值(F-Measure)[8]三个评价指标对结果进行评价. 设 A 代表微博话题, A_i 代表实验数据集中所标记话题为 A 的微博数目, A_j 表示实验结果聚簇中包含微博数目, A_{ij} 表示实验结果聚簇中标记的微博数目, 则

P、R、F 值的计算公式如下:

$$P = precision(i, j) = \frac{A_{ij}}{A_j} \tag{8}$$

$$R = recall(i, j) = \frac{A_{ij}}{A_i} \tag{9}$$

$$F = \frac{2PR}{P+R} \tag{10}$$

取 k=5,利用改进后的 K-means 算对数据进行聚类,得到天津爆炸和人民币贬值话题实验结果如图 1 所示.

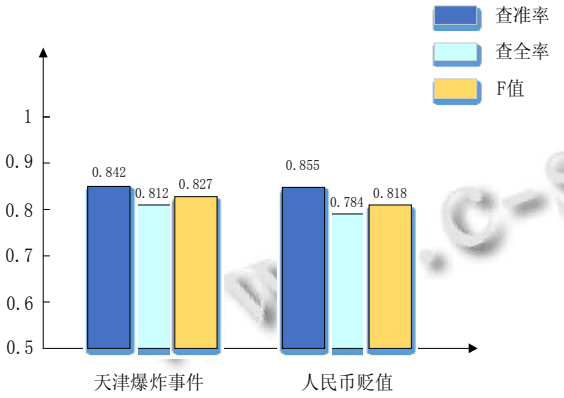

图 1 话题评价

如图 1 中可见, 经过本实验得到的热点话题准确率基本上能够达到 80%以上, 结果让人比较满意, 对于两个事件的综合评价标准 R 的值也比较高, 进一步对改进算法进行验证, 得到实验结果如图 2 所示.

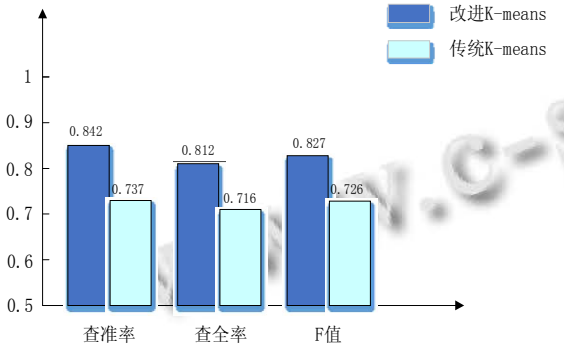

图 2 算法结果对比

通过从查准率、查全率和 F 值评价指标对改进的 K-means 算法进行评价, 从图 2 中可以看到改进后的 K-means 算法比传统的 K-means 算法在查准率、查全率和 F 值有较大的提高, 使得聚类结果更加准确.

5 结语

微博的信息量大, 为了从中获取到微博中的话题, 本文采用 K-means 聚类算法对微博数据进行挖掘, 通过对传统的 K-means 算法不足之处的分析, 提出了优化的初始中心点选取的改进算法, 并把改进后的算法运用到微博话题发现中, 通过聚类算法评价标准, 证明改进算法在聚类结果精度上得到了明显提高, 实验得到了预期的结果, 改进的算法有可能产生局部最优解, 有待以后对算法做进一步的改进.

参考文献

1 蒋盛益,麦智凯,庞观松,吴美玲,王连喜.微博信息挖掘技术研究综述.图书情报工作,2012,56(17):136–142.

2 Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. on Knowledge and Data Engineering, 2005, 17(6): 734–749.

3 樊宁.K 均值聚类算法在银行客户细分中的研究.计算机仿真,2011,28(3):369–372.

4 孙吉贵,刘杰,赵连宇.聚类算法研究.软件学报,2008,19(1):48–61.

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

6 段明秀.层次聚类算法的研究及应用[硕士学位论文].长沙:中南大学,2009.

7 彭时名.中文文本分类中特征提取算法研究[硕士学位论文].重庆:重庆大学,2006.

8 Wang CH, Nan LL, Ren YP. Research on the text clustering algorithm based on latent semantic analysis and optimization. IEEE International Conference on Computer Science and Automation Engineering, 2011, 10(12): 470–473.