A Comparative Study between Single-Pass Algorithm and K-means Algorithm in Web Topic Detection

Xi Ting [1,a] , Li Jufang [2,b]

[1]National University of Defense Technology， Changsha， Hunan， China

[2]National University of Defense Technology， Changsha， Hunan， China

[a]xtbeihang@sina.com, [b] lijufang@nudt.edu.cn,

*Corresponding author :Xi Ting

Abstract. As with the extensive application of the Internet, the explosive growth of information and unprecedented enthusiasm of users, the monitoring and management of Web content is becoming more and more imminent. Although traditional Single-Pass algorithm and K-means algorithm each has shortcomings, they are widely used in clustering analysis because of their relatively simple principles and fast computing speed. This paper firstly describes the overall flow of the entire topic of detection, then we make a comparison between Single-Pass algorithm and K-means algorithm. In order to verify the comparison, finally, an experiment is designed. The result shows that Single-Pass algorithm is better than K-means algorithm in Web topic detection.

Keywords: **Text data, Clustering algorithm, Topic detection.**

Single-Pass算法和K-means算法在Web话题分析中的对比性研究

习婷 [1, a] ， 李菊芳 [2,b,*]

[1]国防科学技术大学信息系统与管理学院，长沙，湖南，中国

[2]国防科学技术大学信息系统与管理学院，长沙，湖南，中国

[a]xtbeihang@sina.com, [b] lijufang@nudt.edu.cn,

[*]习婷

中文摘要. 互联网的广泛应用、信息的爆炸式增长以及网民参与热情的空前高涨，使得对 Web 内容的监控和管理迫在眉睫。传统的 Single-Pass 算法和 K-means 算法虽然各有缺点，但是由于其原理比较简单，计算速度快，而被广泛应用。本文首先介绍了整个话题发现的总流程，然后对 Single-Pass 算法和 K-means 算法进行了对比性研究，最后设计了实验进行验证，并对聚类结果进行了对比分析，为相关的信息使用者进行分析提供了一定的决策支持。

关键词: **文本数据；聚类算法；话题分析**

1. 引言

Internet已经成为继报纸、电视、广播之后的第四代媒体。据CNNIC发布的《第33次中国互联网络发展状况统计报告》显示，截至2013年12月，全国网名人数达到6.18亿，互联网普及率45.8%，手机网民规模达5亿。还有数据显示，全国60%的网民经常在网上发表言论，86.5%的网民把Internet当做是最重要的信息获取渠道，50%的网民对Internet的信任程度比电视、杂志高[1]。网络的广泛应用、信息的爆炸式增长以及网民参与热情的空前高涨，使得对Web内容的监控和管理迫在眉睫。但是，

面对如此海量的Internet信息，单纯地依赖人工监管是难以想象的，于是网络舆情监控系统也就应运而生。

网络舆情是通过Internet传播的公众对现实生活中某些现象所持有的较强影响力、倾向性的言论和观点，是网民关注的热点，是民众讨论的焦点，主要通过新闻评论、BBS论坛、博客、聚合新闻（RSS）、转帖等实现并加以强化，集中反映一个时期网络舆论的中心[2]。对网络舆情进行监控的主要流程如图1所示。其中，信息采集是基础，话题发现是核心，热点评估和跟踪预警是延续，分析处理是解决途径。

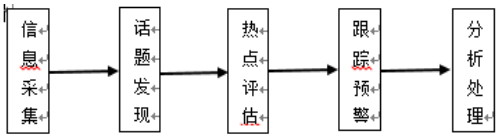

图1. 网络舆情监控流程

话题发现是依靠聚类的方法，将报道聚合成若干簇，簇间的报道之间相似度很低，簇内的报道之间相似度很高，以此来整合网络上大量的重复信息和同一话题的不同信息。使用TDT（topic detection and tracking）相关技术后，人们一话题为粒度发现新事件并了解事件的发展，在应对信息爆炸和新知识发现上有较重要的意义[3]。

聚类是一种常见的数据分析工具，其目的是把大量数据点的集合分成若干类，使得每个类中的数据之间最大程度地相似，而不同类中的数据最大程度的不同。目前现有的聚类算法有很多，依据其采用的基本思想将其大致分为五类，即层次聚类算法、分割聚类算法、基于约束的聚类算法、机器学习中的角力算法以及用于高维数据的聚类算法，如下图2所示。

其中，分割聚类 K-means 方法实现简单有效，适于球形聚类，但需要实现确定聚类的数目 K，并且对鼓励噪声点敏感，同时，聚类结果和效率跟初始类中心的选择有很大关系；层次聚类算法的有点在于：距离和规则的相似度容易定义，限制少；不需要预先制定聚类数；可以发现类间的层次关系。但缺点是时间复杂度较高，同时奇异值也会对算法产生较大影响[4]。CMU[5]使用经典的 Single-Pass 算法对新事件进行探测，该算法原理简单，计算速度快，但容易受新闻文本输入顺序的影响，聚类精度不是很高。雷震等人[6]提出一种改进 K-均值算法用于热点话题发现，该算法使用密度函数法进行聚类中心的初始化，以便客观的选择初始聚类中心。该算法既可以用用在线观测，也可以用于回溯探测，并且执行结果受新闻文档被处理顺序的影响较小，但是当新报道到来时，该算法需要针对所有样本重新计算，无法保证话题发现的实时性[7]。

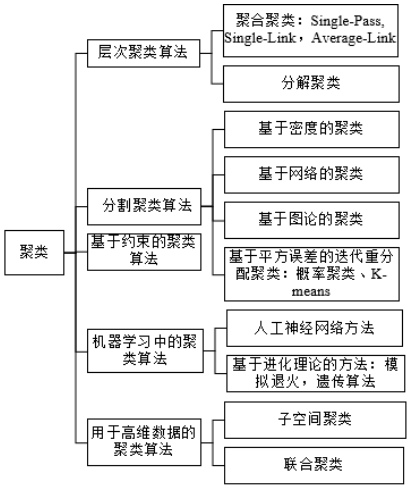

图2. 聚类算法分类

目前，从实用性的角度来看，应用最广的还是K-means算法和Single-Pass算法。虽然大部分学者对其进行了进一步的改进，但是缺乏对两者聚类效果之间的比较。鉴于此，本文就采用实验数据来对传统的K-means和Single-Pass两种聚类算法进行比较，并且在比较的基础上，希望能够进一步挖掘出一些有用的热点话题，为管理层进行决策、制定方案提供一些建议。

2. 文本聚类流程

2.1 网页信息采集与清洗

首先是数据的获取。一般情况下，web分析是采用网络爬虫来获取数据。通过Deep Web技术动态查找所需新闻网,通过网页清洗工具对网页的导航、广告、版权说明等噪声信息进行去噪后，使用节点智能识别技术抽取出相关文档字段，从而完成数据清理。因为本文是直接从搜狗词库中下载出来相关的新闻数据，系统已经对数据进行了去噪和清洗，可以直接应用。

2.2 网页信息预处理

（1）文本特征项的抽取

因为汉语中词语间无明确的分隔标记，所以要采用中文分词技术来实现对文本特征项的抽取。所谓中文分词，就是将一个汉字序列切分成一个个单独的词。它是文本挖掘的基础，对于输入的一段中文，成功的进行中文分词，可以达到电脑自动识别语句含义的效果。本文采用中国科学院技术研究所研制的汉语词法分析系统ICTCLAS对文本内容进行分词处理。

文本经过中文分词后词的向量空间维度非常大，使得后面的处理会花费大量的时间；此外，切分出来的特征词中包含了大量的无实际意义的词，这些词对话题的描述和识别无任何作用。所以，本文在分词的过程中根据词的长度和词性将介词、连词、助词等虚词以及词长较短的无实际含义词过滤掉。

（2）文本特征权重的计算

TF-IDF是一种用于信息搜索和信息挖掘的常用加权技术。TF词频指的是某一个特征词在该文档中出现的频率，IDF是特征词的反文档频率，指在整个文档集中包含特征词的文档的数目，是特征词普遍重要性的度量。TF越高，表明特征词在文档中越重要，可认为该特征词具有良好的类别区分能力，适合用来分类；IDF越大，即特征项在文档集中的多篇文档中出现，则表明该特征词并非本文档的独有特征，在聚类过程中对类划分所起的作用不大。

对于某一文档里的特征项t_i来说，它的重要性可以表示为:

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

上式中$n_{i,j}$表示该该词在文档d_j中出现的总次数，而分母则是文档d_j中所有词出现次数之和。

$$idf_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|} \tag{2}$$

其中|D|表示文档集中文档的总数，$|\{j:t_i\in d_j\}|$表示包含特征项t_i的文档数。如果该词在文档中未出现，就会导致分母为零，因此，一般情况下使用1+$|\{j:t_i\in d_j\}|$

$$tfidf_{i,j}=tf_{i,j}\times idf_i \tag{3}$$

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的TF-IDF。因此，TF-IDF倾向于过滤掉常见的词语，保留重要的词语。

（3）文本向量空间模型（VSM）的构建

VSM（vector space mode）用于表征文本，每个文本d_i均被映射成文本特征的权重向量。文本的每个特征项均被赋予一个权重w_i（即tfidf）,以表示该特征项在该文本中的重要程度。这样，就可以构建一个文本向量空间模型，方便后续进行数值运算。文本向量空间模型的表示如下：

$$D=D(t_1,w_q;t_2,w_{2;}\cdots t_n,w_n) \tag{4}$$

其中特征项t_i的权重为w_i，$1\le i\le n$。因此，根据(2)中计算出来的W，我们可以构建文本的向量空间模型。

2.3 Web话题聚类算法

2.3.1 Single-Pass 算法：

在TDT评测中，Single-Pass算法是使用最多的算法。其运算速度快，原理简单，但对文本输入的顺序比较敏感。一旦文本顺序发生了变化，聚类结果就会出现很大的不同。此外，在Web新闻报道的话题发现应用中，报道一般是根据时间来组织的，输入的顺序是确定的，所以该算法的缺点并不会对话题发现的结果造成太大影响，并且运行速度方面的优势非常适合新闻报道数量大对算法复杂度提出的要求。

本文采用传统的夹角余弦值公式计算各特征项之间的距离，并且比对相似度阈值将其判为已有类或新创建的类。

$$\cos\theta = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} = \frac{A \bullet B}{|A| \times |B|} \quad (5)$$

其中 A 和 B 为文本 A 和文本 B 的向量空间。

Single Pass 聚类算法顺序的处理输入的新闻文本，每次处理一篇，增量地更新聚类。预先设定一个相似度阈值 T_c，如果新闻文本与已有话题模型之间的相似度超过了 T_c，这篇新闻就归属于该话题模型文本类；否则根据该新闻文本创建一个新的话题及其对应的话题模型，同时把该新闻文本归属创建的新话题模型文本类。通过调整 T_c 可以控制聚类的精度[8]。具体算法如下：

Step1，接受一篇新闻文本向量 d；

Step2，对 d 与已有的话题模型向量分别求解余弦值，计算相似度；

Step3，如果余弦值大于 T_c，d 所对应的新闻文本分配给这个话题模型文本类，重新计算这个话题模型的中心向量，跳转至 Step5；

Step4，如果余弦值小于 T_c，d 所对应的新闻文本不属于已有的话题模型，因此，创建新的话题，同时把这篇文本归属创建的新话题模型，并且该新闻文本向量就是新话题模型的中心向量；

Step5，结束。

2.3.2 K-means 算法

K-means 算法是一种经典的分割聚类算法，其基本思想是在聚类开始时，用户预先设定一个类簇数为 K，然后据此从所有文本库中随机选择 K 个文本，将这些文本作为初始类簇的中心，对于文本库中剩余的每个文本，计算其到每个类簇中心的欧几里得距离，并将其划分到最近的类簇中；全部分配完之后，重新计算每个类簇的中心，以及每篇文本到新类簇中心的距离，把文本重新划分到最近的类簇中；不断循环，直到所有的样本都不能再重新分配为止。

算法有如下优点：第一，对待处理文本的输入顺序不太敏感；第二，对凸型聚类有较好结果；第三，可在任意范围内进行聚类。 然而 K-means 算法也有难以消除的缺陷：第一，对初始聚类中心的选取比较敏感，往往得不到全局最优解，得到的多是次优解；第二，关于算法需要预先设定的 K 值，限定了聚类结果中话题的个数，这在非给定语料的应用中并不可行；第三，该算法容易受到异常点的干扰而造成结果的严重偏差；第四，算法缺少可伸缩性[9]。

3 Web话题分析

本文采集了从 2014 年 1 月 1 日起至 2014 年 3 月 31 日之间的新浪、腾讯、网易、搜狐上发表的报道(贴文)，从中选取文档内容长度适合聚类分析、话题中报道数适中的话题 80 个，共 3000 篇报道，经过格式清洗、内容提取、分词和向量化后得到的文本向量，进行仿真。

3.1 算法聚类结果对比分析

系统运行时间单位为秒(S)，图 3 为仿真结果。

其中 K-means 方法按照算法运行开始时的设定把聚类数目设为 80 个，算法运行时随机指定每个类的初始中心，所以该算法得出 80 个话题类的时间特别快，之后进入缓慢耗时的重定位过程； single-pass 算法将话题聚为 91 个，按照文档的相似程度逐一对文本向量进行计算得出话题的，时间复杂度也基本上和话题数目成正比。可以看出，采用 single-pass 算法得出最终结果的响应速度要比 K-means 算法快，这是因为 K-means 算法指定初始类中心时是随机的，而后需要大量的重定位，这严重影响了其运算速度，而 single-pass 算法在聚类时首次遇到一篇文档时就依据与已有类

的相似性为其指定最可能的类属，从而节约了大量的运算时间。

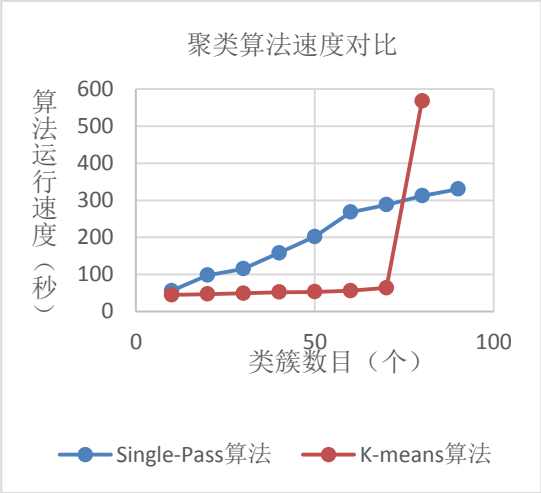

图 3. 聚类算法速度对比图

3.2 算法精度比较

通常，人们使用漏检率和错检率以及耗费函数来评价话题发现的质量，即聚类的精度。假设与某话题相关的 Web 报道文本数是 A，不相关 Web 文本数是 B；其中 A 里面有 A1 篇被聚到该话题类中，B 中有 B1 篇被聚到该话题类。则

错检率：$$F=\frac{B_1}{B}, \quad (6)$$

漏检率：$$E=\frac{A-A_1}{A} \quad (7)$$

而耗费函数综合考虑了漏检和错检的代价，其计算公式为：

$$C=C_E*P_E*P+C_F*P_F*(1-P) \quad (8)$$

其中 C_E 和 C_F 是漏检和错检的代价，P_E 和 P_F 是漏检和错检的概率，P 文本属于某话题类的先验概率。根据 TDT 评测标准，本文设定 $C_E=1, C_F=0.1, P=0.125$

算法精度统计结果如下表所示：

表1. 算法精度对比

	Single-Pass 算法	K-means 算法
话题数	92	80（指定）
错检率	0. 00368	0. 00124
漏检率	0. 207	0. 144
耗费函数	0. 0262925	0. 018632

观察可知，因为 K-means 算法是预先制定要聚类的数目，所以该算法的聚类结果有 80 个话题，而 single-pass 算法是根据相似阈值来自动生成类簇的，所以它生成话题的数目不可以控制，而且当用户觉得生成的类簇数目过多而将相似阈值调小企图减少话题数目时，算法的错检率将升高，本不属于某话题的报道可能被错误归并到这个话题中；当用户觉得类簇过少而提高阈值时算法的漏检率将升高，话题分裂的可能性加大，也会影响算法的整体性能。但这种话题数目的不可控制性，也可以说成不用控制性，正是 single-pass 算法适用于网络话题发现的原因——动态数据源情况下无法提前获知话题的数目或话题数目不是固定值。

K-means 算法初始中心是随机指定的，在一定程度上影响了其精度，当然也带来了运算速度上的问题，并且该算法往往使聚类结果陷入局部最优解，而全局来看并不最优，所以该算法的整体耗费函数要大于改进算法，而 single-pass 算法多出来的 11 个话题都被当做了错检漏检，所以耗费函数大。

4. 总结

在对数据挖掘和聚类方法等相关理论进行一定的回顾和总结的基础上，对 Single-pass 算法和 K-means 算法从聚类结果和算法精度上进行了对比性分析，总结了各自的适应条件。为管理者和决策者进行决策提供了一定的技术支撑。

本文在研究中发现了以下几个问题：（1）在构建文本向量之前，需要建立一个包含所有特征项的词表，然后依据每个特征项的 TF-IDF 构建文本向量，其中有些特征项在该篇文本中的 TF-IDF 可能为零，但也被包含在向量空间里，这样导致整个向

量空间的维度特别大，增加后续余弦值的计算量；（2）相似度的阈值大部分都是依赖于经验值，但是对于某些特殊问题，其经验值可能会影响到聚类的结果。这些都是算法改进中需要解决的问题。

References

[1] CNNIC. The 33rd statistical report of the China Internet network development [EB/OL].[2014-01-14].http//www.cnnic cn/upload files/pdf/2014/1/14/170516 pdf

[2] CHEN H,CUI D W,LI X，et al.The harmonious evolution of ethnic group algorithm [C]//Proc of the 3rd International Conference on Natural Computation：IEEE Computer Society,2007:380-384

[3] Xie Qian long，Xu Wei ran. An Algorithm for Online Sudden Topic Detection Based on Microwave of Automatic Machines [J].Microsoft，2012,33（12）：109-113

[4] Seo Y W，Sycara K. Text clustering for topic detection[Z].USA：Carnegie Mellon University,2004

[5] Zhang Xiao yan，Wang Ting. A study of topic detection and tracking technology [J]. Journal of Frontiers of Computer Science and Technology，2009,3（4）：347-357

[6] Lei Zhen，Wu Ling da，Lei Lei, et al. A Incremental K-means Method Based on Initialize Class Center and Its Application in News Events Detection [J]. Journal of the China Society Scientific and Technical Information，2006，25（3）：289-295

[7] Wang Wei, Xu Xin.The Cluster-based Network Public Opinion Hotspots Detection and Analysis[J]. New Technology of Library and Information Service,2009（3）:74-79

[8] Shui Yidong, Zhai You li, Huang Kuan hou. Topic Detection and Tracking Method Combined with Periodic Classification and Single-Pass Clustering [J]. Journal of Beijing Jiaotong University，2009（5）：85-89

[9] Yin Feng jing, Xiao Wei dong, Ge Bin, Li Fang fang. A Incremental Text Clustering Algorithm for Network Topic Detection [J].Application Research of Computer, 2011,28（1）:54-57