**文章编号：**153

# 基于话题检测的自适应增量 K-means 算法*

李胜东[1]，吕学强[2]，施水才[2]，孙军[3]

（1. 廊坊燕京职业技术学院 计算机工程系，河北 廊坊 065200；2. 北京信息科技大学 网络文化与数字传播北京市重点实验室，北京 北京 100101；3. 北华航天工业学院，河北 廊坊 065000）

**摘要：**根据话题检测任务的定义和特点，本文分析了传统的增量聚类算法和 K-means 算法的优缺点，提出了基于话题检测的自适应增量 K-means 算法，设计了话题检测实验，实验结果证明了该算法提高了话题检测性能，具有良好的应用前景。

**关键词：**话题检测；增量聚类；K-means 算法；话题检测与跟踪评测

**中图分类号：**TP391　　**文献标识码：**A

# Adaptive Incremental K-means Algorithm Based on Topic Detection

LI Sheng-dong[1], LV Xue-qiang[2] ,SHI Shui-cai[2], SUN Jun[3]

(1. Department of Computer Engineering, Langfang Yanjing Polytechnic College, Langfang Hebei 065200, China ; 2. Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing 100101, China; 3. North China Institute of Aerospace Engineering, Langfang Hebei 065000, China)

**Abstract:** According to the definition and characteristics of topic detection task, the paper analyzed the advantages and disadvantages of the traditional incremental clustering algorithm and K-means algorithm, proposed adaptive incremental K-means algorithm based on topic detection, and designed topic detection experiments. Experimental results prove that the new algorithm improves the performance of topic detection and has good prospects.

**Key words:** topic detection; incremental clustering; K-means algorithm; topic detection and tracking evaluation method

## 1 引言

互联网的出现，使信息急剧膨胀。这些信息包含有用的信息、无用的信息、感兴趣的信息、不感兴趣的信息等。在这种情况下，人们最关注的是如何快速而又准确地得到感兴趣的信息。目前，各种信息检索、信息抽取和信息过滤技术也都围绕这个目的展开[1]。但是，这些技术返回的信息冗余度过高，比如，仅仅因为信息中含有指定的关键词，许多不相关的信息就被作为结果返回了，其中，即使是相关的信息，也没有进行有效地组织。在这种背景下，研究者开始关注一种新的技术，它就是话题检测技术[2]。该技术就是研究如何检测新发生的事件，并帮助人们把分散的信息有效地组织起来。

在话题检测与跟踪研究中，话题检测[3,7]被定义为将输入的新闻报道归入不同的话题簇，并且在需要的时候建立新的话题簇。从定义可以看出，话题检测研究在本质上等价于一种无监督的聚类研究，即它的关键技术就是文本聚类算法。文本聚类算法[5]一般可以分为基于层次

* **收稿日期：2014-05-06　　定稿日期：2014-07-25**

**基金项目：**本项目得到网络文化与数字传播北京市重点实验室开放课题(No.ICDD201105、ICDD201205、ICDD201401)；国家自然科学基金项目（No.61271304 ）；北京市教委科技发展计划重点项目暨北京市自然科学基金 B 类重点项目（No.KZ201311232037）；2013 年河北省高等学校科学技术研究自筹资金项目（No.Z2013162）资助。

**作者简介：**李胜东（1984——），男，硕士，讲师，主要研究领域为文本数据挖掘、信息检索，智能优化。E-mail: lsd_6@126.com；吕学强（1970——），男，博士，教授，主要研究领域为中文信息处理、多媒体信息处理。E-mail: lxq@bistu.edu.cn；施水才（1966——），男，硕士，教授，主要研究领域为中文信息处理、信息检索与 WEB 应用。E-mail: shi.shuicai@trs.com.cn；孙军（1979——），男，硕士，讲师，主要研究领域为图像处理。E-mail: sunjun_79@163.com。

的聚类算法、基于平面划分的聚类算法、基于密度的聚类算法等，其中，最常用的是基于层次的聚类算法和基于平面划分的聚类算法。基于层次的聚类算法[15]可以达到很高的精确度，但是时间复杂度较高；以 K-means 算法为代表的基于划分的聚类算法，具有很高的效率，适合处理海量文本数据。

## 2 话题检测任务的特点

话题检测任务[7]的关键技术是文本聚类算法，这决定了它除了具有文本聚类的相似性，还有一些自己的特点。传统的文本聚类算法从全局的角度处理静态的对象，而话题检测任务从局部的角度以增量的方式处理动态的对象。这是话题检测任务的特点，也是话题检测与文本聚类算法的本质区别。

## 3 聚类算法

### 3．1 传统的增量聚类算法

传统的增量聚类算法处理话题检测问题时，其基本思想[16]是一次处理一篇报道。对于每一篇报道，先与每个已知话题进行比较，如果相似度大于阈值，则把该报道归入相似度最高的话题，如果对所有话题的相似度都低于阈值，则创建一个新话题，并更新话题数。

这种算法非常简单且易于实现，但缺点也很明显：一篇报道只能做一次决策，早期根据很少信息作出的错误判断，累计到最后的错误量可能很大。针对这个缺点，本文对比了国内外常用的聚类算法，分析了传统的 K-means 算法，发现传统的 K-means 算法和增量聚类算法具有优缺点互补的可能性，能够弥补传统的增量聚类算法的缺陷。

### 3．2 传统的 K-means 算法

K-means 聚类[6]的算法思想简单，易于实现，能够快速有效地处理大规模数据，已经成为数据挖掘等领域应用最广泛的聚类算法之一。它的核心思想[8]是通过迭代过程把数据划分到不同的聚类中，以使目标函数（1）最小化。

$$E = \sum_{i=1}^{K} \sum_{x \in C_i} \left| x - x_i \right|^2 \qquad (1)$$

在公式（1）中，$C_i$ 是语料中的第 i 类话题；x 是话题 $C_i$ 中的数据对象；$x_i$ 是话题 $C_i$ 中的均值；K 为初始化的聚类数，也是算法认定的话题数。

定义 1：根据两层阈值的话题/报道表示模型，报道 i 和报道 j 之间的余弦相似度函数 $Sim\left(d_i, d_j\right)$ 的定义为[9]：

$$Sim\left(d_i, d_j\right) = \frac{\sum_{k=1}^{M} w_{ik} \times w_{jk}}{\sqrt{\sum_{k=1}^{M} w_{ik}^2} \times \sqrt{\sum_{k=1}^{M} w_{jk}^2}} \qquad (2)$$

在公式（2）中，$d_i$ 表示报道 i 的特征向量；$d_j$ 表示报道 j 中的特征向量；参数 M 是基于两层阈值的话题/报道表示模型的特征空间维数。

定义 2：报道向量间的余弦距离 $Dis\left(d_i, d_j\right)$ 被定义为：

$$Dis\left(d_i, d_j\right) = 1 - Sim\left(d_i, d_j\right) \qquad (3)$$

从定义 1 和定义 2 可知，报道 i 与报道 j 之间的余弦相似度越大，它们越相似，因此，这两

个报道之间的余弦距离越小，传统的 K-means 算法越有可能把它们作为同一个话题。
传统的 K-means 算法通过迭代过程能够得到全局最优解，但初始化 K 值影响它的性能。

## 4 自适应的增量 K-means 算法

传统的 K-means 聚类算法可以通过迭代过程得到全局最优解，但初始化聚类数 K 制约着该算法的性能；而传统的增量聚类算法对一篇报道只做一次决策，能够得到局部最优解，很难得到全局最优解，但该算法不需要初始化聚类数 K。
通过分析传统增量聚类算法和 K-means 聚类算法的优缺点，发现它们的优缺点有互补的可能性；在此基础上，用传统的增量聚类得到初始聚类中心，产生自适应的 K 值，解决 K-means 算法对 K 值初始化敏感的问题；然后用传统 K-means 算法的迭代过程得到全局最优解，解决传统增量聚类中的局部最优问题，提出了自适应的增量 K-means 算法作为话题检测算法。
自适应的增量 K-means 算法把所有中文新闻报道语料划分为 r 个增量，每个增量报道的规模为 $N_i$（$i = 1,2,\cdots,r$）。对于每一个增量，先按传统增量聚类处理所有报道，得到 K 个聚类，接着对当前增量按照传统 K-means 算法进行迭代操作，每一次迭代都按要求进行适当的改变，直到聚类中心不变为止，然后处理下一个增量的报道。详细算法过程如下：

- Step1：对于每一个增量，设它的报道规模为 $N_i$（$i = 1,2,\cdots,r$），判定报道 S 是否是第一篇报道，如果是，使用报道 S 建立第一个话题，如果不是，计算报道 S 与其他话题中心的相似度。
- Step2：根据 S 与各个话题的相似度，找到与 S 相似度最高的话题 $T_1$。
- Step3：判定报道 S 与话题 $T_1$ 的相似度是否大于阈值 $\theta$。如果相似度大于阈值 $\theta$，就把报道 S 归入话题 $T_1$，否则，使用 S 建立一个新话题，并更新话题数 K。
- Step4：判定报道规模 $N_i$ 是否为 0。如果 $N_i$ 为 0，则输出话题数 K 和聚类结果，并转到 Step5；否则，转到 Step1，处理下一篇报道。
- Step5：根据传统增量聚类的结果，计算 K 个话题的均值，作为的传统 K-means 算法的初始聚类中心。
- Step6：根据公式（2）和公式（3），计算每个聚类中心与其余所有新闻报道之间的余弦距离。根据余弦距离的大小，把每个报道分配到余弦距离最小的聚类中心，也就是把每个报道分配到最近的聚类中心。
- Step7：重新计算每个聚类的均值，作为该话题类的新聚类中心。
- Step8：如果所有的聚类中心不发生变化，这说明目标函数收敛到最小值，转向 Step9；否则，修改聚类中心，按照 Step6 和 Step7 迭代。
- Step9：判断增量数 i 是否为 0。如果 i 为 0，算法终止，并输出话题数 K 和聚类结果；否则，转到 Step1，处理下一个增量的报道。

## 5 实验结果与分析

根据传统增量聚类算法、传统 K-means 算法和基于话题检测的自适应的增量 K-means 算法的算法思想，分别把它们作为话题检测算法设计话题检测实验，得到相应的话题检测与跟踪评测结果，对比评测结果评估基于话题检测的自适应的增量 K-means 算法作为话题检测关键技术的性能。

### 5．1 实验结果

在实验中，分词程序是中科院计算所软件室提供的 ICTCLAS[10]；语料是中科院计算所谭松波博士提供 14150 篇中文新闻报道[11,12]，第一层是 12 个主题，第二层是 60 个话题。对每个话题检测算法，在实验参数和实验环境相同的条件下，分别用 60 个话题进行测试，得到 60

个测试结果，对这 60 个结果进行归一化后，得到能够反映话题检测性能的话题检测与跟踪评测结果，即：归一化检测开销$(CDet)_{Norm}$[13,14]。最后，根据实验的话题检测与跟踪评测结果分别评估传统的增量聚类、传统的 K-means 算法和自适应的增量 K-means 算法的性能。实验的评测结果如表 1 所示。

表 1　话题检测与跟踪评测结果

| 关键技术 | 传统增量聚类 | 传统 K-means 算法(K=60) | 自适应的增量 K-means 算法 |
|---|---|---|---|
| $(CDet)_{Norm}$ | 0.4257 | 0.3972 | 0.3789 |

为了便于分析这个新算法对话题检测性能的影响，用 Excel 2003 中的图表向导工具把表 1 中的数据映射成图 1。

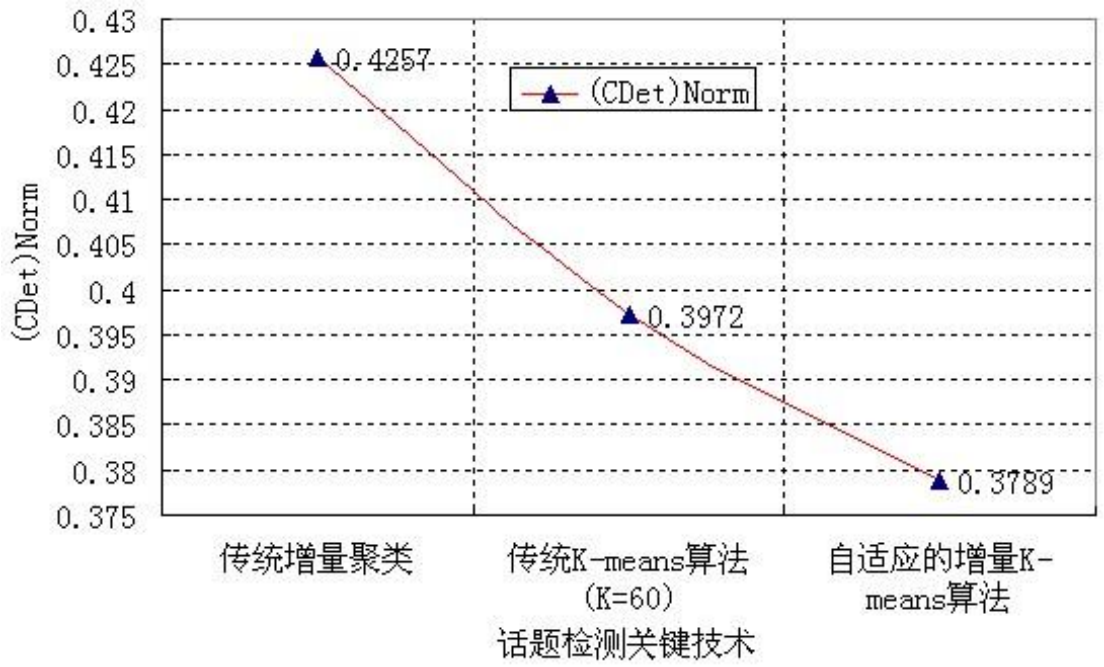


图 1　话题检测关键技术与话题检测性能之间的变化趋势图

## 5．2 实验分析

在实验中，根据评测结果评测话题检测性能，然后根据话题检测性能评估聚类算法作为话题检测关键技术的性能。在同等条件下，评测结果越小，说明话题检测性能越好，也表明该聚类算法作为话题检测关键技术的性能越好。

根据表 1 和图 1，传统的增量聚类作为话题检测关键技术时，TDT 评测结果为 0.4257；传统的 K-means 算法作为话题检测关键技术时，TDT 评测结果为 0.3972；自适应的增量 K-means 算法作为话题检测关键技术时，TDT 评测结果为 0.3789。因此，在同等条件下，基于自适应的增量 K-means 算法的评测结果比基于传统的增量聚类的评测结果减少了 10.994%，即自适应的增量K-means算法的性能比传统的增量聚类提高了10.994%;基于自适应的增量K-means算法的评测结果比基于传统的 K-means 算法的评测结果减少了 4.607%，即自适应的增量 K-means 算法的性能比传统的 K-means 算法提高了 4.607%。除此之外，传统的 K-means 算法需要对 K 值初始化，而且 K 的初始化值对该算法性能的影响很大，而自适应的增量 K-means 算法用传统增量聚类的思想自适应地调节 K 值，在很大程度上减少了 K 初始化值对该算法性能的影响；传统的增量聚类能够得到局部最优解，但很难得到全局最优解，而自适应的 K-means 算法通过迭代过程能够得到全局最优解，很好地解决了传统的增量聚类所面临的问题。

## 6 结论

本文分析了话题检测任务的定义和特点，对比了传统的增量聚类和 K-means 算法的优缺点，然后通过传统的 K-means 算法改进了传统的增量聚类，提出了基于话题跟踪的自适应增量 K-means 算法。在同等条件下，经过广泛而深入地研究和分析可知，新算法作为话题检测关键技术具有良好的性能。

## 参考文献


[1] 郑斐然 苗夺谦 张志飞等.一种中文微博新闻话题检测的方法[J].计算机科学，2012，39(1): 138~141

[2] 张阔，李涓子，吴刚等. 基于关键词元的话题内事件检测 [J]. 计算机研究与发展，2009，46（02）：245~251.

[3] 李忠俊.基于话题检测与聚类的内部舆情监测系统[J].计算机科学，2012，39(12): 241~244.

[4] 赵华，赵铁军，赵霞. 时间信息在话题检测中的应用研究[J].计算机科学，2008，35(01): 221~223.

[5] 马慧芳, 王博. 基于增量主题模型的微博在线事件分析[J]. 计算机工程, 2013, 39(3): 191-196.

[6] 吕明磊，刘冬梅，曾智勇.一种改进的 K-means 聚类算法的图像检索方法[J]. 计算机科学，2013，40(8): 285~288.

[7] Nist. The 2004 Topic Detection and Tracking (TDT2004) Task Definition and Evaluation Plan. http://www.itl.nist.gov/iad/mig/tests/tdt/2004/TDT04.Eval.Plan.v1.2.pdf.

[8] 毛嘉莉. 基于 K-means 的文本聚类算法[J]. 计算机系统应用，2009，（10）：85~87.

[9] Li Xinwu. Research on Text Clustering Algorithm Based on K_means and SOM [C]. ShangHai: International Symposium on Intelligent Information Technology Application Workshops, 2008. 341~344.

[10] 中科院计算所. 基于多层隐马模型的汉语词法分析系统 ICTCLAS. http://www.nlp.org.cn/project/project.php?proj_id=6.

[11] 谭松波，王月粉. 中文文本分类语料库 -TanCorp V1.0. http://www.searchforum.org.cn/tansongbo/corpus.htm.

[12] Tan, S.B., et al. A Novel Refinement Approach for Text Categorization[C].ACM CIKM 2005, 2005.

[13] Tim Leek, Richard Schwartzand Srinivasa Sista. Probabilistic Approaches to Topic Detection and Tracking [J]. Data Mining and Knowledge Discovery. 2003, 7 (3)：67~83.

[14] Yiming Yang, Jaime Carbonell, Ralf Brown, etal. Multi-Strategy Learning for Topic Detection and Tracking: a joint report of CMU approaches to multilingual TDT [C]. TDT 2002 Workshop. 2002. 85~114.

[15] 骆卫华，于满泉，许洪波等. 基于多策略优化的分治多层聚类算法的话题发现研究[J]. 中文信息学报，2006，20（01）：29~36.

[16] 洪宇，张宇，刘挺等. 话题检测与跟踪的评测及研究综述[J]. 中文信息学报，2007，（06）.



**作者简介：**李胜东（1984——），男，硕士，讲师，主要研究领域为文本数据挖掘、信息检索，智能优化。E-mail: lsd_6@126.com；吕学强（1970——），男，博士，教授，主要研究领域为中文信息处理、多媒体信息处理。E-mail: lxq@bistu.edu.cn；施水才（1966——），男，硕士，教授，主要研究领域为中文信息处理、信息检索与 WEB 应用。E-mail: shi.shuicai@trs.com.cn；孙军（1979——），男，硕士，讲师，主要研究领域为图像处理。E-mail: sunjun_79@163.com。

**通讯作者:** 李胜东，Tel: 18730622257； E-mail: lsd_6@126.com。


李胜东

吕学强

施水才