**热点话题研究综述**

**韩东成 杨晓雅 谷睿 杨可**

(北京理工大学计算机学院 北京 100081)

**摘要** 热点话题是一定时期间人民群众最关心的事件，热点话题的发现与研究有利于我们了解当下社会的舆论焦点，同时也便于政府对舆情进行监督和正方向引导。本文首先论述了热点话题的发现与演化过程，包含媒体关注度、用户关注度、复合关注度等话题发现模型。其次介绍了国内外对话题发现的研究现状，对相关理论和技术进行简要概述。之后，着重分析PLSA和LDA两个主题模型，并对K-means算法的传统方法在话题文本聚类和自生成K值两方面加以改进，最后进行对比实验，验证改进方法的有效性。

**关键词** 热点话题；话题演化；PLSA模型；LDA模型；K-means算法

**A Survey on Hot Topics**

**HAN Dongcheng YANG Xiaoya GU Rui YANG Ke**

(School of Computer Science and Engineering, Beijing Institute of Technology, Beijing 100081)

**Abstract** Hot topics are the most concerned events of the people during a certain period of time. The discovery and research of hot topics will help us understand the focus of public at the current time, and also facilitate the government to supervise and guide the public opinion. This paper firstly discusses the discovery and evolution process of hot topics, including topic discovery models such as media attention, user attention, and compound attention. Secondly, it introduces the research status of topic discovery at home and abroad, and gives a brief overview of relevant theories and technologies. After that, the paper focuses on analyzing the two topic models of PLSA and LDA, and improves the traditional methods of K-means algorithm in topic texts clustering and self-generated K value. Finally, the comparison experiment is carried out to verify the effectiveness of our improved methods.

**Keywords** Hot topics; topic evolution; PLSA model; LDA model; K-means algorithm

组员分工：

韩东成（K-means的传统和改进方法及实验,综述整理）

杨晓雅（综述热点以及文本预处理）

谷睿（PLSA 、 LDA主题模型）

杨可（热点话题的发现与演化）

1. **引言**

近年来，随着互联网技术相关技术的不断更新和发展，互联网已经成为下一代信息媒体，其发展速度和影响力远远超过传统的广播、电视、报纸等以往的主流媒体。在其蓬勃发展的同时，相应的问题也出现了：如今新媒体入行门槛低，网络流量大，网民数量多，也容易让网络上出现一些三俗、暴力、恐怖甚至反动的信息和言论。通过有效的网络热点舆论管理，相关部门可以准确地应对网络环境下的舆情特点所带来的挑战，热点话题发现技术就是实现自动化网络舆情管理的核心技术之一，如何快速准确的发现热点话题是当下网络舆情研究的重点内容。

1. **热点话题的发现与演化**

网络热点话题主要指从大量的网络新闻中提取出被大众广泛讨论的话题，那么热点话题必然也是当前新闻报道的最重要的事件，网络群众关注的焦点。那么事件的被关注程度就是从海量新闻中筛选出热点话题的依据，那么也可以将它作为话题热度的重要指标。

以下内容分别为热点话题的定义以及判断其热度的重要指标、什么是热点话题的演化过程、话题发现模型、热点话题演化发展过程和热点话题演化发展过程。

* 1. **热点话题的定义以及判断其热度的重要指标**

一个话题能够被称之为热点话题，首先必须需要多个新闻媒介的报道，与此同时，该话题能进一步发展成社会热点关注的话题，还必须与网民的关注和社会大众的关注有密切关系。

网络新闻媒体对话题的关注称之为是媒体关注度，网民及社会群众对新闻话题的关注称之为是用户关注度。一个话题如果能够被称之为热点话题，那么该话题必须同时具有较高的媒体关注度及用户关注度。

媒体关注度从新闻发送方的角度考虑新闻话题受到各个新闻媒体的关注程度，用户关注度从新闻接收和二次传播方面决定了一个话题能否成为热点话题，如果一个新闻话题被用户广泛关注，那么必然会对新闻媒体造成正向反馈效果，各个媒体将会迎合大众需求增加对该话题的报道。

用户关注度主要体现该话题的传播范围和社会对该话题的反响程度。用户关注度基于用户的网络行为体现，如是否主动点击该新闻话题，是否进行二次搜索该话题，是否评论该话题新闻等。

* 1. **热点话题的演化过程**

热点话题往往具有以下几个特点：1、传播速度快、范围广。2、一般情况下网络上新闻话题热度达到一定程度后，将会由线上转移到线下，网络舆情一定程度上能够引导社会舆情。3、热点话题在受到关注后，会在极短的时间内蔓延并持续一段时间，然后关注程度迅速下降直到消退热点话题的关注度总会有一个明显的爬升过程，即使某些持续热点话题在每个时间点的关注度都不低，但仍然会在被某个事件刺激后被关注程度突然骤升，达到最高点，也有的可能是新出现的话题，因此前期被关注程度很低，甚至没有用户关注度，只有为数不多的媒体报道。但当受到网络群体的关注后热度也会达到峰值。热点话题的被关注程度在骤升过程后进入一个相对平稳的阶段，我们即称之为“热点话题的流行阶段”，在这个阶段，该话题始终保持很高的关注度。热点话题的流行阶段持续时间越长表明它的话题性越大。并且在相对平稳的流行阶段中，热点话题的被关注程度到达个最高峰，流行过后，该话题被关注程度将会呈现下降趋势，直到消亡。

每个热点话题从开始形成到结束消失都有生命周期，一般包括产生、发展成熟、消亡四个阶段。其中在话题的成熟期，往往还会衍生出其他子话题，这些子话题有时作为该话题的补充，有时会完全取代该话题，成为一个新的热点话题。即新产生的话题是已有话题的延续，已有的老话题在成熟期或者发展期直接消退，演化产生的相关话题直接成为一个成熟的热点话题，或进一步成为更热门的话题。

* 1. **话题发现模型**

**2.3.1 基于改进TF-PDF的媒体关注度模型：**

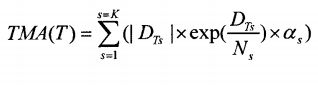
TF-PDF算法认为一个新闻话题能够被称之为热点话题必须满足以下条件：

（1）必须有足够多新闻网站报道该话题。

（2）各新闻网站报道该话题的新闻数量必须明显多余其他新闻数量

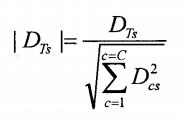
目前中国各大门户网站上的新闻质量是不一样的，人们对每个网站的认可程度也不相同。即同样一条新闻，如果由新华网或新浪、网易等知名网站发布与由其他地方性、非知名网站发布对网民的影响力明显不同。因此我们在传统的TF-PDF的基础上引入了“网站认可度”这一参数表明网民对网站的认可程度。

话题T的媒体关注度的公式如下：

（1）

其中表示网站总数，表示网站上的话题总数，表示网民对网站的认可度。

话题在网站上的新闻报道数目的公式如下：

（2）

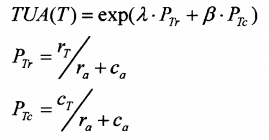
为了避免新闻报道频数对整体关注度造成影响，因此需要对话题报道频度进行标准化，即上式。

（3）

是指网站上对某个话题报道的数目占总的报道次数的百分比指数，话题相关报道越多，则该话题的值越大，则该话题成为热点话题的可能性也越大。

**2.3.2 基于用户浏览行为的用户关注度模型：**

用户关注度从新闻接受者和二次传播的角度考量网络新闻热点话题的热度，通过网民关注新闻的浏览行为可知，一篇新闻的被阅读次数、被评论次数基本可以确定该新闻是否是被网民用户所关注。通常情况下，网民会优先选择对具有话题性的新闻进行阅读和评论，而具有话题性，能够引起社会讨论的新闻往往也就是热点新闻、热点话题。基于用户行为的用户关注度模型如下：

（4）

表示话题丁的用户关注度，、分别表示话题被阅读次数和被评论次数占总次数的比例。、是调节因子，表示话题被阅读的总次数，表示话题被评论的次数，、是所有话题被阅读和评论的次数。

**2.3.3 新闻热点话题发现复合关注度模型**

1545228922(1)（5）

、动态调节媒体关注度与用户关注度的权重，显然，在门户网站媒体关注度对话题的热度影响要大于用户关注度，但在某些论坛上，用户关注度的影响力又要大于媒体关注度。

**2.3.4 在线热点话题发现算法**

在对网络新闻热点话题的提取和分析时，以下为一种基于时间窗口的在线话题分段聚类方法，将每一天看成一个时间窗口进行聚类分析，并计算每个时间窗口下各个话题的复合关注度及每个话题内部各核心事件占比，提取出每日热点话题，然后再将每日新产生的热点话题与当前已有热点话题合并。为后续该话题的发展走势研究和演化分析奠定基础。

该算法如下：

输入：当前时间

已有热点话题集合*={t1、t2.....tn}*

今日新闻文本数据集*D={d1，d2，d3，．．．，dn}*

类簇相似度阈值

输出：今日热点话题集合*={t1、t2.....tn}*

更新后的热点话题集合*T*

算法步骤：

a)对集合*D*采用算法进行聚类分析，得到多核心话题模型集合。并计算各个话题下各核心事件关注度占比；

b)对集合中所有话题根据复合关注度公式计算其话题关注度，并按热度从大到小排序，给出今日前十热点话题，保存到今日热点话题集合；

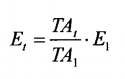
c)将集合与已有热点话题集合T进行合并，设表示集合的第*i*个话题，表示集合丁的第*j*个话题，如果*sim(，)>β*，合并与，并在对应的话题走势图上标记。如果对于，不存在与之对应的，则表明为新话题，加入到集合*T*中，并建立对应的话题走势图；

d)删除掉集合*T*中没有被更新的话题(即已经消亡的话题)；

e)当前时间*f=t+1*，进行下一时间窗口新闻文本分析，转到步骤a)。

* 1. **热点话题演化发展过程**

由以下式子可以表示话题当天的话题指数：

（6）

表示第*t*天该话题的话题指数，表示该话题第一天的话题指数。表示该话题第*t*天的关注度。

从以上可以得到一条该话题的发展走势曲线图，根据话题的发展走势曲线图，我们可以直观的观察到当前热点话题处于话题生命周期的哪个时期，若话题发展走势曲线呈上升趋势，则说明该话题目前处于发展阶段，话题热度正在持续上升，若发展走势曲线呈现平稳阶段，且相当于第一天具有较高关注度，说明该话题正处于成熟阶段，若当前处于下降阶段，则表明该热点话题正在走向消亡。

* 1. **话题演化偏移分析**

在基于多核心话题描述模型下，热点话题的演化可以看成话题内核心事件偏移的过程，某些话题从话题产生到消亡这一整个生命周期都只有一个核心关注事件点。而对于某些热点话题，在话题进入成熟期或者消亡期时，由于受到广泛的关注从而产生了更深层次的调查讨论，导致该热点话题的关注核心发生偏移，媒体的报道与用户的关注点都偏向另一个事件。

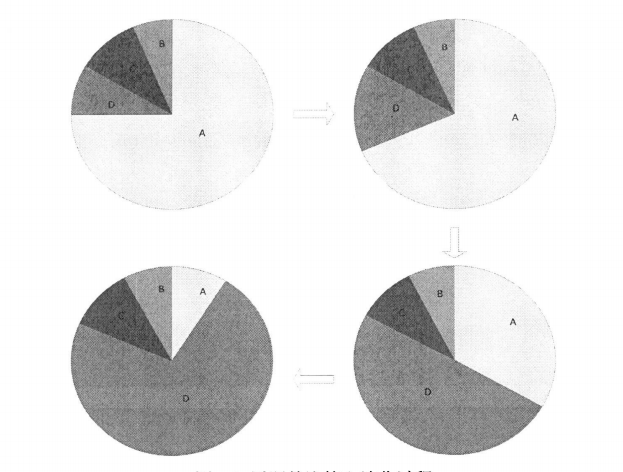


图1 热点话题关注核心的转变过程

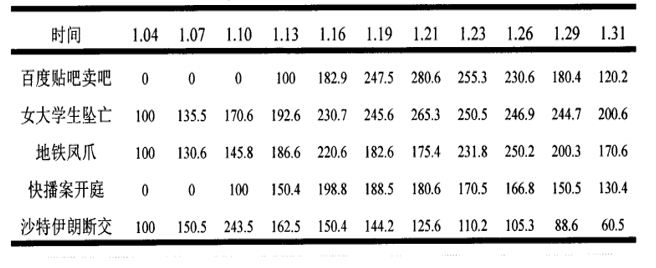
图1展示了一个热点话题内关注核心的转变过程，首先该热点话题在产生时包含四个核心关注点，即该话题的多核心模型拥有4个质心，分别称之为A、B、C、D，扇形区域大小表示当前阶段各关注核心受关注程度占该热点话题总受关注度的比例。

由图1可知，在该话题刚产生时，明显关注核心是事件A，其占有该话题绝大部分关注度，是大众关注的焦点，而事件B、C、D是属于该话题的其他三个周边核心事件。随着时间的推移，我们发现该话题的核心关注焦点开始转向事件D，事件D开始占据该话题主要关注点，这一过程我们称之为该话题的演化过程。如果采用单核心话题描述模型，前后两个时间的核心关注事件A与核心关注事件D可能会被当成两个热点话题，事实上，事件D被广泛关注是由于人们对事件A广泛关注讨论后所引起的连锁反应(后续报道、深度报道、引发思考等)，理应将事件A与事件D归为同一话题。

* 1. **具体实例分析**

**2.6.1关于百度卖吧事件走势图的分析：**

表1 走势话题指数表



走势话题指数表中展示了五个热点话题的话题指数，可以看出，几乎每个热点话题都有经历“产生、发展、顶峰、消退”这四个阶段。

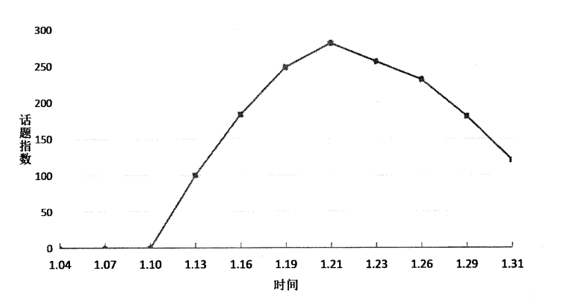


图2 话题指数走势图

走势图中可以看出，百度爆出卖吧事件后，短短几天时间就达到了热度顶峰，且居高不下，到了1月23号才开始呈现下降趋势。

**2.6.2关于地铁凤爪事件的分析**

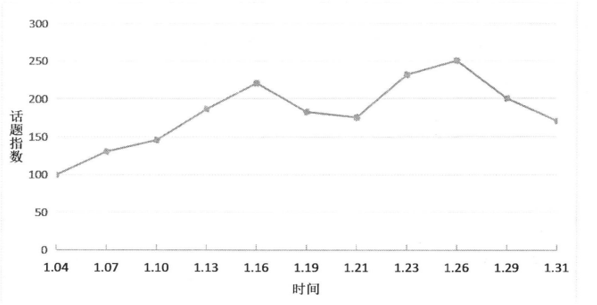


图3 话题指数走势图

从发展走势曲线图中“地铁凤爪事件”的话题指数不是一个单峰的话题，“地铁凤爪事件”在1月16日和1月26日处于高热度状态。即该事件在1月16日到达事件“顶峰”后，正处于消退状态时，经过新一轮的报道，将该话题又再次推上热度“顶峰”。



图4 关注占比表

从图中可以看出，最开始的该话题核心是针对“地铁吃凤爪乱吐骨头”这一不文明现象的报道与谴责。这一话题核心从产生到16号达到讨论顶峰。开始进入消退阶段，然而这时，针对这一事件的深入报道得知，该“凤爪女”故意导演了这一不文明现象，借助网络媒体的报道与传播达到“炒作”的目的，这时，网民对该话题的关注核心已经变成了对“炒作”事件的谴责与自己被消费的不满。

1. **研究现状**

话题检测与追踪(Topic Detection and Tracking，即TDT)，主要用来发现和跟踪新闻报道中的而心生新闻的时间，自动识别新闻报道，自动发现突发新闻时间，能够自动跟踪新闻时间的完整发展过程等。近年来，互联网正逐渐取代报纸、电视成为新闻报道的主体，网络新闻的爆炸式发展，以及对网络热点新闻话题的发现与跟踪的研究都取得了很大的进步，对网络新闻话题的发现与追踪采用的技术基本上是基于早期的TDT技术的。

TDT评测会议把话题检测和跟踪分为以下五个子任务：

1）报道切分：找出所有新闻报道的边界，将输入的新闻文本数据分割成各个独立的新闻报道。

2）话题跟踪：根据源新闻报道给出某话题的几则报道，然后将接下来的相关报 道归类到该话题集群。

3）话题检测：检测所有新闻报道集，发现新闻报道集中的新话题集群。

4）首次报道检测：在流式的新闻报道集中检测首次讨论某个话题的新闻报道。

5）关联检测：通过新闻报道的相似性判断两则新闻报道讨论的是否是同一个话题。

目前，有关网络舆情研究、热点话题发现的论文在不断增加，越来越多的学者加入到这个领域来进行学术研究，很多学校和科研机构也己形成了专门的研究课题小组。通过研究网络舆情的基本理论和改进已有的技术方法，已经有很多的科研成果在各大舆情监测系统中成功应用。

国外对于话题发现的研究比较成熟，基于语言模型的方法由Spitters于2002年首次提出，他将新闻文本的表示及文本之间相似度的计算都基于语言模型；Smith等人采用命名实体标注的方法来判断同一话题下的子话题集合；Pouliquen等人选择使用对数似然检验计算词的权值而不是采用传统的TF—IDF值。IBM采用了一种基于两层聚类策略的方案，使用Okapi公式来比较两篇新闻文本的相似度。

国内对于TDT领域的研究都是基于对NLP(Natural Language Processing，自然语言处理)的研究基础上展开的，这主要由于汉语与英文句子的组成存在差异，英文词只需要按照空格区分即可，而中文则根据字意可拆分情况具有多变性，虽然国内对TDT的研究起步比较晚，但目前仍取得了不错的研究成果。

黄晓斌等人在网络舆情的描述、网络舆情的相关性计算、网络舆情真实性鉴定等方面做了深入的研究与实验，在网络舆情信息分析中运用了文本挖掘技术，取得了不错的效果。

许鑫等人对网络信息采集技术、预处理技术、关键词抽取等方面都进行了分析与研究，构建了一个集信息采集、处理、发布为一体的网络舆情监测研究平台。

贾自艳等人首次使用单通道聚类方法Single—Pass用于处理流式文本数据，他们在Single-Pass文本聚类方法基础上还提出了一种动态进化模型的事件探测和追踪算法。

骆卫华提出了基于多策略优化的分治多层聚类的话题发现方法，通过多层聚类的MLCS算法对话题进行组织，他们在信息检索时充分利用了目前完善的NLP技术。

刘星星等人提出了一种使用两层聚类算法的热点话题自动发现方法，并且给出了话题的热度计算公式，实现了完整的热点话题自动发现系统。

张小明等人提出了基于增量聚类的自动话题检测研究模型，通过提炼出具有较强的主题辨别能力的文本特征来提高文本聚类的准确率，同时在聚类过程中通过BIC来判断话题类簇数。

李凤岭等人在基于微博的话题发现 领域引入了LDA模型，增加了对文本语义相关方面的研究。截至到目前，基于新闻话题的研究一直是文本数据挖掘、自然语言处理等研究领域的热点研究目标。

1. **相关理论和技术**
   1. **数据获取和采集**

网络爬虫是搜索引擎的基础技术，是采集技术和采集策略的结合，通过将网络数据采集到本地，为之后的数据提供前提条件。社交网络平台提供的开放编程接口（API）和软件设计组件（SDK），可以更加简洁地获取相应的Web数据，为高效获取原始数据提供了保障。出于对数据安全性和服务负载能力考虑，目前社交网络所提供的API和SDK都存在着时间和空间上的访问限制。

国内外研究自然语言处理和文本挖掘的机构，提供种类丰富的文本数据集，如被普遍使用的LDC语料库和国内使用较多的Sogou语料库。

* 1. **中文分词**

中文分词是指将一个汉语序列划分为彼此独立的词语。分词就是将连续的文字序列按照特定的规则重新划分成词语的过程。在英文的行文规范中，词与词地分隔符是空格，而中文语境中，只是字、句和段能够通过明显的分隔符来明显划界，词和词之间没有一个明确的分隔符，中文处理地复杂难度要远远高于英文。

中文分词地研究成果已经趋于完善，且分词效果也趋于稳定，当下研究中常用的主流中文分词工具有：汉语词法分析系统ICTCLAS、IKAnalyzer和Paoding。

* 1. **文本表示**

文本数据通过格式预处理、中文分词和去除停用词等操作后，成为由一串字或者单词组成的文本串。由于构成文本的短语不同、词语维度各异，挖掘算法无法对文本数据进行直接计算，为了方便文本挖掘处理，通常将文本数据用一种文本模型统一标识，向量空间模型和布尔模型是两类常见的文本表示模型。

* 1. **特征提取**

文本特征提取的含义是利用特定的特征提取算法，从特征集合中提取中那些对文本挖掘意义大得特征词并予以保留，同时将那些对文本挖掘完全无意义或者意义及其小的特征词汇进行删除。

特征项是文本数据中出现的单字、词语和词组等，由于词语具有独立的含义，并且相对容易切分、在文本信息中具有一定的统计学规律，所以传统意义上会选择词语作为特征项，就是用特征词语的统计值作为文本向量的项值。然而直接使用文本中的词语作为特征项词语会由于特征项的高维性，导致“维数灾难”。在利用特征词建立文本向量之前，必须对文本的特征值进行降维处理，包括特征选择和特征抽取。

在文本挖掘方面，一般采用TF-IDF（Term frequency-inverse document frequency，即词频-逆向文档频率）算法作为特征值的权重计算方法。TF-IDF权值方法基于一下原理：文本的特征词必然会在该文本中多次出现，并且在其余不同类文本中出现的次数较少，那么该词就可以被认为是该文本的特征词，即该词能够很好的区分该文本。TF-IDF的权值标准化计算公式如下：

（7）

其中表示特征词i在文档中出现的次数，考虑到文章有长短之分，为了便于不同文章的比较，进行“词频”标准化：将词频数除以该文章总词数得到标准化“词 频”。N为新闻报道总数，为包含特征词i的新闻报道数，a是调节因子，一般取 1。

区别于新闻文本不同于其他普通文本，一般来说，对于网络新闻报道中的标题往 往会涵盖该新闻中所报道的大量信息甚至全部关键信息，新闻的正文只是对新闻标题 的一个补充和细化，因此，对于新闻文本，我们计算其特征词权重时需要赋予标题中出现的特征词一个更高的权值。本文在词频上需要作特殊处理：

（8）

其中与分别表示标题与正文中特征词出现的次数。

在计算出每个文本每个词的TF-IDF值后，按值的大小排序，选择排行靠前的特定数目的词作为该文本的特征词项即可，经过研究发现，一般选择新闻字数5％个数的特征词项即可很好的表示该新闻报道的内容。

* 1. **相似度分析**

文本聚类就是将相似的文本项目归一成统一的一个类目。要判断文本之间的相似性，就涉及到文本之间的相似度分析。相似度分析主要有以下两种方法：①相似度，相似度就是比较两个对象之间相像程度数值度量。两个对象越像，它们之间的相似度就越高。

通常，相似度是正数，并常常在0（不相像）和1（完全相像）之间取值。②相异值，相异度时比较两个对象区分程度的数值度量。对象之间越像，它们之间的相异度就越低。通常，相异度是正数，并常常在0（不相像）和1（完全相像）之间取值。

本文对两种常用的相似度计算方法进行介绍：

1. 夹角余弦相似度

在文本向量空间建模后，通过计算两个空间向量的夹角余弦距离来定义两个文本之间的相似程度。计算公式如下：

（9）

1. JS(Jenen-Shannon)距离

JS距离最早由香农提出来，利用信息熵的概念解决对信息的量化度量。信息论中， 熵是指对事件不确定性的描述，熵越高，即表示能传输的信息越多，反之，则表示能够传输的信息量越少。信息熵的定义如下：

（10）

KL距离是一种基于信息熵的距离的表现形式，主要衡量某一相同事件两个不同概率分布之间的差异情况。KL距离全称为Kullback．Leibler Divergence，由于其对熵的计算，KL距离也常常被称作为信息相对熵：

（11）

KL距离可以理解为相同事件空间中P(x)和Q(x)概率分布之间的差异性。其物理意义为：在相同事件的空间分布下，概率分布为P(x)的事件空间，若采用Q(x)进行 编码时，平均每个基本编码长度增加的比特数。

显然对于任意事件空间，按照自身概率分布进行编码是最好的方式，因此始终大于0，当且仅当时，等于0。同时，KL距离不具有对称性。

JS距离是对KL距离对称性的改进，公式如下所示：

（12）

P、Q为某个相同事件的两个空间概率分布，显然，JS距离越大，表示两个分布之间的差异越大。反之，则表明两者之间的分布差异越小。夹角余弦距离利用空间两个向量的夹角余弦值表示两个文本向量的相似度，适用于所有的向量空间模型。因此在基于TF-IDF的向量空间模型中的话题发现算法中经常使用。

而不同于夹角余弦距离，JS距离是对空间概率分布差异情况的描述，从相对熵的角度度量两个概率空间分布的差异性，比较两个概率分布的差异性时，JS距离的精度优于夹角余弦相似度，因此本文在基于LDA主题模型建模后计算的LDA 模型下文本向量之间的相似度时，因为使用文本-潜在主题这一概率分布情况对文本进行向量建模，所以选择JS距离作为文本之间的相似性度量效果更优。

* 1. **文本聚类**

热点话题发现算法从本质上来讲是属于数据挖掘中的文本聚类算法，文本热点被准确发现和文本聚类中的算法本身特征以及各种阈值的选取是密不可分的。

关于聚类算法的研究在20世纪90年代开始引起人们的广泛关注，也由此却得了不错的成绩。而对于文本聚类而言，并不是所有的聚类算法都可以应用到文本聚类中。目前的聚类算法包括基于划分的方法和基于平面层次方法两大类，除此之外，还包括基于密度的方法、基于网格的方法和基于模型的方法。

在热点话题发现上，被广泛应用的聚类算法一般只有基于划分的方法。

1. **PLSA、LDA主题模型**

传统的话题发现算法在话题聚类时多利用向量空间模型，即采用特征词-权值这个组合来进行文本向量空间建模，将文本抽象成特征向量，词组抽象成特征项，用特征向量之间的距离来度量文本之间的相似程度。这种方法的弊端就是只关注了文本词组的重复程度和出现次数，而忽略了文本背后的语义信息。

主题模型可以避免这一问题。构思文章时首先要做的是确定主题，然后围绕主题遣词造句，表达成文。比如构思一篇自然语言处理相关的文章，我们会涉及语言学、概率统计等多个方面，而提到语言学，人们会联想到语法、句子等词语；说到概论统计，人们会联想到概率、均值、马尔科夫链等词语。

主题模型是对文本中潜在主题进行建模。在文本和词之间增加了一层主题层，先让文本和主题产生某种关联，然后再从主题中寻找词的概率空间分布。利用主题模型模拟文本生成以后，进行产生文档的逆过程，利用看到的文档推断其隐藏的主题分布，从而进行话题聚类，这才是主题建模的目的。本文将介绍两种主题模型。

* 1. **PLSA——概率潜在语义分析模型**

假设我们要生成M篇文档，有K个可选主题，这些主题各自在文档中出现的概率大小不一样，同时每个主题有V个可选的词。一篇文档由不同的词组成，确定了文档里每个位置上的词，一篇文档就生成了。

把文档生成看作掷骰子的过程。假设有两种骰子，一类是文档-主题骰子，每个骰子有K个面，每一面对应一个主题。另一类是主题-词骰子，每个骰子有V个面，每一面对应一个词。

在生成文档前，我们准备一个文档-主题骰子和K个主题-词骰子。先掷“文档-主题”骰子选择主题，得到主题结果后，使用和主题结果相对应的那颗“主题-词”骰子掷出词结果。重复N次以上两步操作得到N个词从而完成一篇文档。重复M次以上操作，则完成M篇文档。

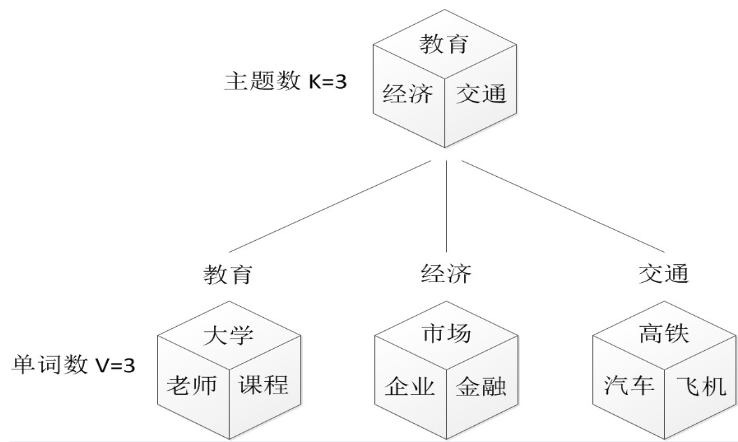


图5 文档生成举例

比如介绍一个国家的文档中，往往会分别从教育、经济、交通等多个主题进行介绍，如图5所示。

令K=3，即制作1个含有3个主题的“文档-主题”骰子，这3个主题可以是：教育、经济、交通。

然后令V = 3，制作3个有着3面的“主题-词项”骰子，其中，教育主题骰子的3个面上的词可以是：大学、老师、课程，经济主题骰子的3个面上的词可以是：市场、企业、金融，交通。主题骰子的3个面上的词可以是：高铁、汽车、飞机。

先掷“文档-主题”的骰子，假设（以一定的概率）得到的主题是教育，那么下一步便是掷“教育”主题骰子，（以一定的概率）得到“教育”主题筛子对应的某个词：大学。

在实际中，“掷骰子”的过程是随机的，遵循一定的概率分布。比如可能选取“教育”主题的概率是0.5，选取“经济”主题的概率是0.3，选取“交通”主题的概率是0.2，那么这三个主题的概率分布便是{教育：0.5，经济：0.3，交通：0.2}。

将各个主题Z在文档D中出现的概率分布为主题分布，同理各个词语W在主题Z下出现的概率分布为词分布。在PLSA模型中需要做的就是如何求出这两个概率。

利用看到的文档推断其隐藏的主题分布的过程（即产生文档的逆过程），便是主题建模的目的：自动地发现文档集中的主题分布。概率关系图如图6所示。节点D和W表示可观察到的文档和单词，节点Z表示观察不到的隐藏变量，表示隐含的主题。

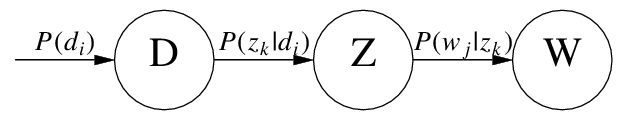


图6 PLSA概率关系图

𝑃(𝑑𝑖)表示词在文档𝑑𝑖中出现的概率，𝑃(𝑧𝑘|𝑑𝑖)表示某个主题𝑧𝑘在给定文档𝑑𝑖下出现的概率，𝑃(𝑤𝑗|𝑧𝑘)表示某个词𝑤𝑗在给定主题𝑧𝑘下出现的概率，𝑃(𝑤𝑗|𝑑𝑖)表示某个词𝑤𝑗在给定文档𝑑𝑖下出现的概率。

文档D和词W是可以观测到的样本，对于任意一篇文档，其𝑃(𝑤𝑗|𝑑𝑖)是已知的。根据大量已知的文档-词项𝑃(𝑤𝑗|𝑑𝑖)，训练出文档-主题𝑃(𝑧𝑘|𝑑𝑖)和主题-词项𝑃(𝑤𝑗|𝑧𝑘)，公式如下：

（13）

故得到文档中每个词的生成概率为：

（14）

𝑃(𝑑𝑖)可以事先计算得出，而𝑃(𝑤𝑗|𝑧𝑘)和𝑃(𝑧𝑘|𝑑𝑖)未知，所以=(𝑃(𝑤𝑗|𝑧𝑘),𝑃(𝑧𝑘|𝑑𝑖))是要估计的参数值，利用EM最大期望算法最大化求解出𝑃(𝑤𝑗|𝑧𝑘)和𝑃(𝑧𝑘|𝑑𝑖)。

EM算法的基本思想是：首先随机选取一个值去初始化待估计的值，然后不断迭代寻找更优的使得其似然函数L()比原来的L()要大。

EM的关键便是要找到L(θ)的一个下界Q(;)（注：L(θ)=log p(X|θ)，其中X表示已经观察到的随机变量），然后不断最大化这个下界，通过不断求解下界Q(;)的极大化，从而逼近要求解的似然函数L(θ)。

EM算法的一般步骤为：

1、随机选取或者根据先验知识初始化；

2、不断迭代下述两步

①给出当前的参数估计，计算似然函数L(θ)的下界Q(;)

②重新估计参数θ，即求，使得

3、上述第二步后，如果L(θ)收敛（即Q(;)收敛）则退出算法，否则继续回到第二步。

上述过程好比在二维平面上有两条不相交的曲线，一条曲线在上（简称上曲线L），一条曲线在下（简称下曲线Q），下曲线为上曲线的下界。

在已知下曲线，上曲线未知的情况下，为了求解上曲线的最高点，我们试着不断增大下曲线，使得下曲线不断逼近上曲线。一旦下曲线在某一个点达到局部最大值并与上曲线在这点的值相等，记录下这个值，然后继续增大下曲线，寻找下曲线上与上曲线上相等的值，迭代到L(θ)收敛（即Q(;)收敛）停止，从而利用当前下曲线上的局部最大值当作上曲线的全局最大值（换言之，EM算法不保证一定能找到全局最优值）。

假定用表示词在主题上的多项分布，则可以表示成一个向量，每个元素表示词项出现在主题中的概率，即P(|)=，；用表示所有主题Z在文档上的多项分布，则可以表示成一个向量，每个元素表示主题出现文档中的概率，即P(|)=，。最终要求解的是两个矩阵和。

由于词与词之间是相互独立的，所以整篇文档N个词的分布为:

P(W|)=（15）

由于文档与文档之间也是相互独立的，所以整个语料库中词的分布为：P(W|D)=（16）

其中，表示词项在文档中的词频，n()表示文档中词的总数，显然n()=。从而得到整个语料库的词分布的对数似然函数为：

L(=（17）

我们需要最大化L(来求解参数和。对于这种含有隐变量的最大似然估计，使用EM算法分为两步：先E-step，后M-step。首先假定参数已知，计算隐变量的后验概率。利用贝叶斯法则可以得到：

P(|)=P(,)/

= P(|)P(|) P()/

= P(|)P(|) /

=/（18）

然后带入隐变量的后验概率，最大化样本分布的对数似然函数，求解相应的参数。根据E-step的计算结果，把= P(|)代入 L(，于是只要最大化下面这个函数L即可。

L=（19）

利用多元函数求极值的方法对求偏导，可估计出参数。

* 1. **LDA——隐含狄利克雷分布模型**

PLSA模型中主题分布和词分布是唯一确定的，LDA主题模型对潜在主题分布和每个潜在主题下词的分布也进行了随机化，他们出现的概率由Dirichlet先验随机确定。

图1的例子在PLSA模型下能明确的指出主题分布可能就是{教育：0.5，经济：0.3，交通：0.2}，词分布可能就是{大学：0.5，老师：0.3，课程：0.2}。但在LDA模型下，主题分布和词分布不再唯一确定不变，即无法确切给出。例如主题分布可能是{教育：0.5，经济：0.3，交通：0.2}，也可能是{教育：0.6，经济：0.2，交通：0.2}，它是随机的可变化的。但无论如何变化都服从一定的分布，即主题分布跟词分布由Dirichlet先验随机确定。



图7 LDA概率关系图

概率关系图如图7所示。除W是观测变量外，其余均是隐变量。其中，表示词分布，表示主题分布，是主题分布的先验分布，是词分布的先验分布。LDA把要估计的主题分布和词分布看作是其先验分布是Dirichlet分布的随机变量。在估计主题分布、词分布的过程中，它们的先验分布（即Dirichlet分布）事先由人为给定，LDA模型下要做的就是求它们的后验分布。利用Gibbs Sampling（吉布斯采样）求解出后验分布从而解决主题分布和词分布两参数未知的问题。

Gibbs抽样是马尔可夫链蒙特卡尔理论（MCMC）中用来获取一系列近似等于指定多维概率分布（比如两个或者多个随机变量的联合概率分布）观察样本的算法。

Gibbs Sampling的主要步骤为：

（1）初始化，对语料库中的每篇文档的每个词，随机赋予一个主题topic编号。

（2）重新扫描语料库，对每个词w，按照Gibbs抽样公式重新采样语料库的主题，在语料中进行更新。

（3）重复以上语料库的重新采样过程直至Gibbs抽样收敛。

（4）统计语料库中的topic-word的共现频率矩阵。

得到主题下词对的概率分布后，通过统计每篇文档中的主题，就可以得到文档下主题的概率分布。

1. **K-means的传统方法及改进**

作为数据挖掘的经典算法之一，K-means是聚类算法中，应用最为广泛的一种。该节主要介绍传统的K-means聚类算法、改进的K-means算法在话题聚类上的应用以及相关试验结论。

* 1. **划分聚类算法**

新闻话题聚类在舆情监督、热点话题发现、突发事件实时追踪等领域有着重要的应用。

新闻话题聚类通常依据新闻所报道的文本内 容相似度差异来进行聚类分析，隶属于同类话题的 新闻有着较高的文本相似性，而不同类的新闻话题 往往具有较低的文本相似性。

新闻话题聚类在信息检索，热点话题发现，舆情监督等领域都有着广泛的应用，被越来越多的研究人员所重视。新闻话 题的聚类常常本质上是一个文本聚类问题，新闻话题的相似性通常通过计算新闻文本之间的相似性来度量。

K-means算法是最经典的基于划分的聚类方法之一，在文本聚类领域有着广泛的应用。其基本思想是以空间中K个点为初始簇中心进行聚类，根据最邻近原则进行对象划分，通过迭代方法更新所有的对象和簇中心向量，将每一个对象都划分到离自己最近的簇中，并且实时地更新簇中心向量，使得所有的对象都能够划分到最合适的簇中，直到迭代停止。

基于 K-means 的文本聚类算法由于算法简单易实现，时空复杂度低，聚类效果优异等特点被广泛用作新闻话题聚类算法。但传统的 K-means 算法又具有其局限性，如对初始中心点的选择敏感和用户必须自定义分组K等，导致算法收敛于局部最优而无法得到全局最优解。针对传统的 K-means 算法中初始聚类中心点随机选择导致聚类结果不稳定的问题，提出了一种改进的 K-means 算法 用于新闻话题检测，该算法基于新闻报道相似性选择初始聚类中心点，保证各新闻话题集群具有很好的区分度。并在此基 础上，根据新闻话题覆盖率自动确定话题集群个数 K。实验结果表明，改进后的算法能够生成稳定的，高质量的话题集群。

目前 K-means 文本聚类算法的变种主要有:

基于渐变中心的K-means文本聚类算法。传统的 K-means 在每次迭代完成后，等所有的对象都已经重新划分到合适的簇后，再更新各个簇的中心向量，而渐变中心方法则会根据当前处理的对象 实时更新簇中心向量。实验结果表明该优化方法是有效的，本文算法也基于此种迭代方式。

基于二分的K-means聚类算法[3]。其思想是 先将所有对象作为一个簇，而后选择能最大程度降 低聚类代价函数的簇划分为两个簇，迭代进行，直 到簇的数目等于用户给定的 K 为止。

通用的划分聚类算法，如图8所示，即在给定的N个元素的数据集里面，通过划分的方法法将这N 个元素分别划分到K个簇。其算法的主要步骤为，首先选择初始聚类中心，其次，将元素按照相似度进行划分，从而修改聚类中心。当满足划分阈值时，划分结束，完成聚类操作；当未满足划分阈值时，继续划分元素并不断修改聚类中心，直至阈值被满足则完成聚类。

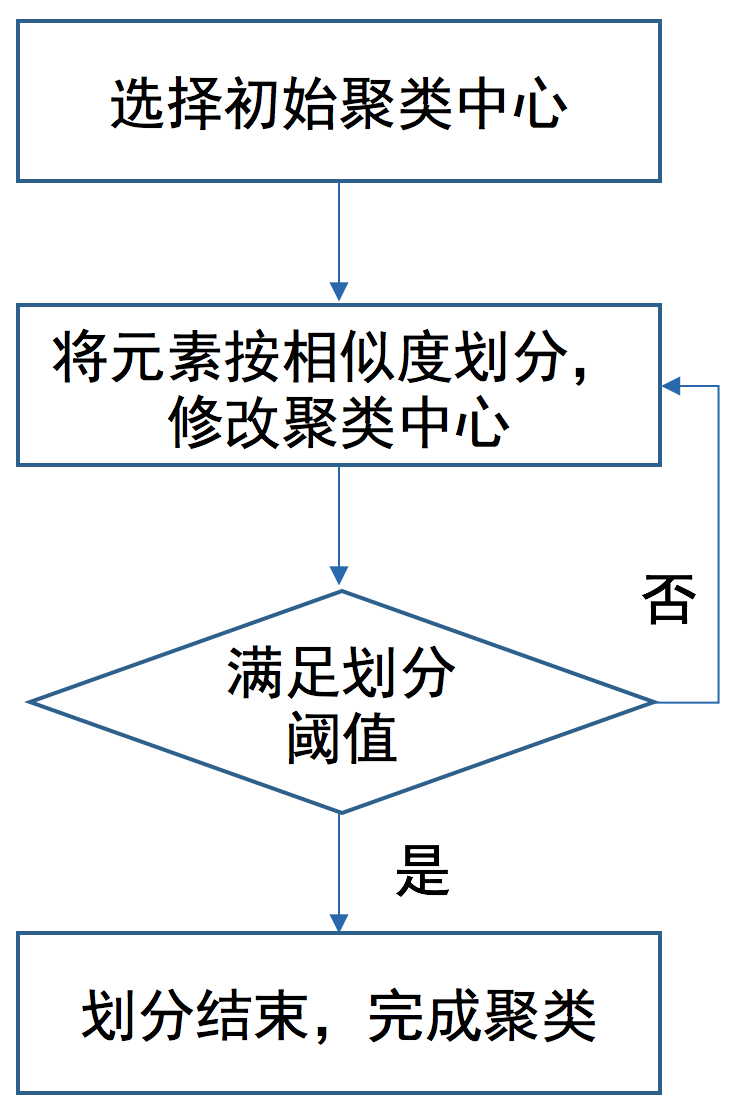


图8 划分聚类算法流程图

K-means算法是划分聚类中常见的算法之一，传统的K-means聚类算法初始化时选定K个初始文本作为聚类中心点，然后将剩余文本依次归类到上述K个簇，并实时更新每个簇的中心点向量，迭代执行直到分组不再改变或者执行了足够多次的迭代。它的优点是计算简单，适合大规模数据，但由于传统的K-means聚类算法对初始聚类中心（选择的随机性）和新闻语料输入顺序敏感，故无法将两个相似话题集群合成一个话题集群。

传统K-means 聚类算法在应用于新闻话题聚 类时因其对初始聚类中心和新闻语料输入顺序敏感，迭代过程可能会终止于一个局部最优值，将某一热点新闻话题划分到多个子话题中，无法得到预期的聚类结果，进一步导致该热点话题被忽视。 即传统K-means文本聚类算法不具备将两个相似话题集群合并成一个话题集群的功能。

该算法需要用户预先设定簇分组 K，不会去掉噪声 数据，完全按照最邻近原则进行对象划分，会将噪声数据划分到其最近的簇，在初始聚类中心选择了 噪声数据时，甚至有可能会导致噪声数据成为一个包含很少对象的簇。

* 1. **改进K-means算法在话题聚类上的应用**

针对以上缺点，改进的K-means聚类算法在初始聚类中心选择和话题数K上进行优化。为了保持各新闻话题集群间具有较好的区分度，且保证聚类结果不受语料输入顺序影响，改进的初始聚类中心选择算法为，首先初始化新闻文本，设置话题数目，循环遍历新闻文本中的各条新闻，对与初始聚类中心集（初始化为0）相似度低于阈值的新闻，加入到初始聚类中心集和话题数目集合中；对与初始聚类中心集相似度高于阈值的新闻，说明其相似性较高，则将其加入到与其报道内容最相似的新闻集里。最终返回初始聚类中心集和各话题集合所包含的新闻集。

具体算法如下：

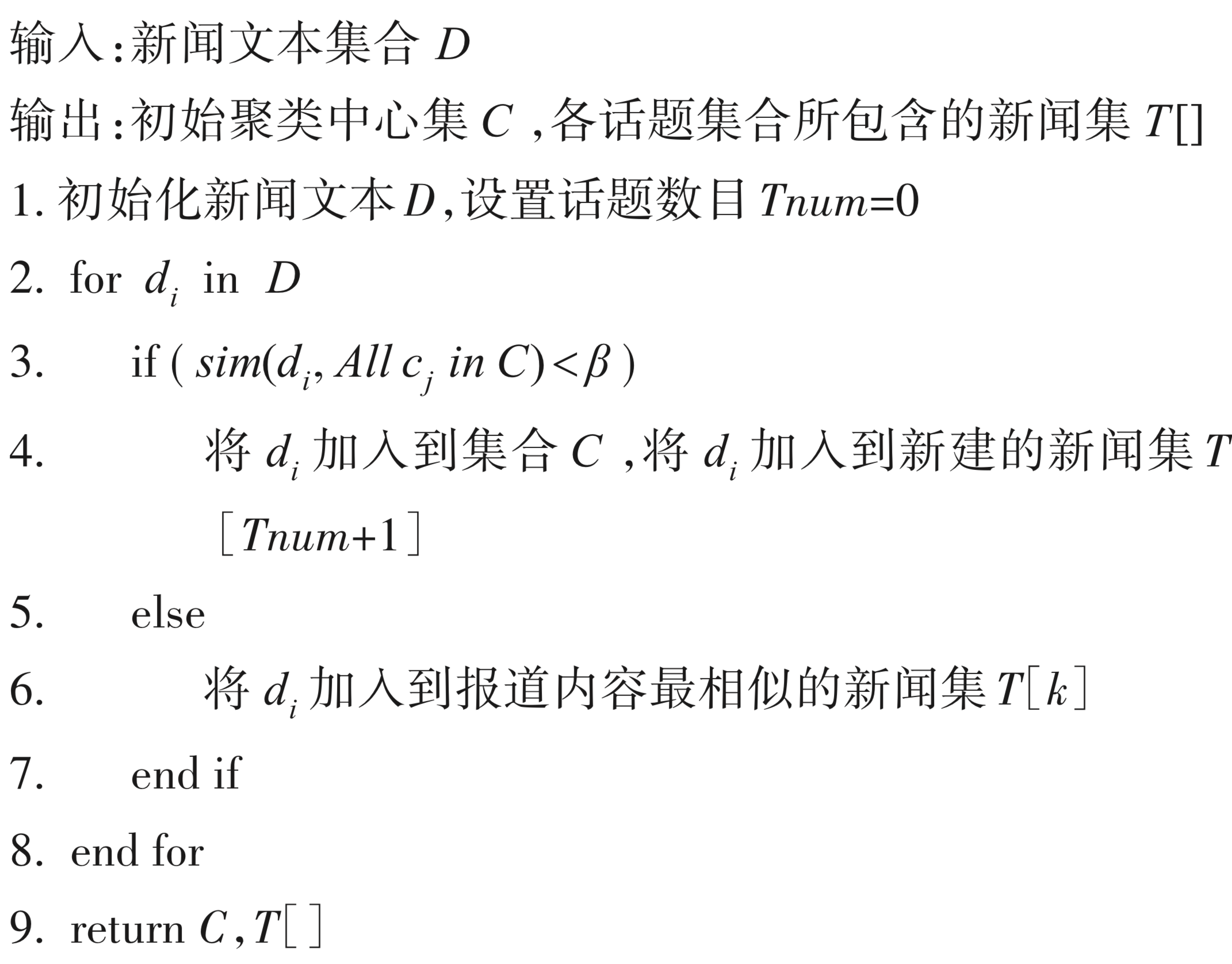


图9 初始聚类中心选择算法图

由于存在噪声数据，所以话题数目虽然会远远小于文本集合大小，但仍比K大，所以我们通过改进传统K-means算法来实现自生成话题数K。首先，利用上文初始聚类中心选择方法得到初始聚类话题中心集、新闻集和话题个数。其次，按照新闻集中各个新闻的数量，对其做降序排列。之后，将前x个话题个数中的新闻占总数的覆盖率，与整个新闻文本集群中所有能够形成话题的新闻占总数的覆盖率进行比较，当前x-1个话题覆盖率小于整体覆盖率且前x个话题覆盖率大于等于整体覆盖率时，我们即确定K为x。最后，我们通过迭代直至标准测度函数收敛为止。

其中，定义话题的新闻文本覆盖率:

（20）

具体算法如下：

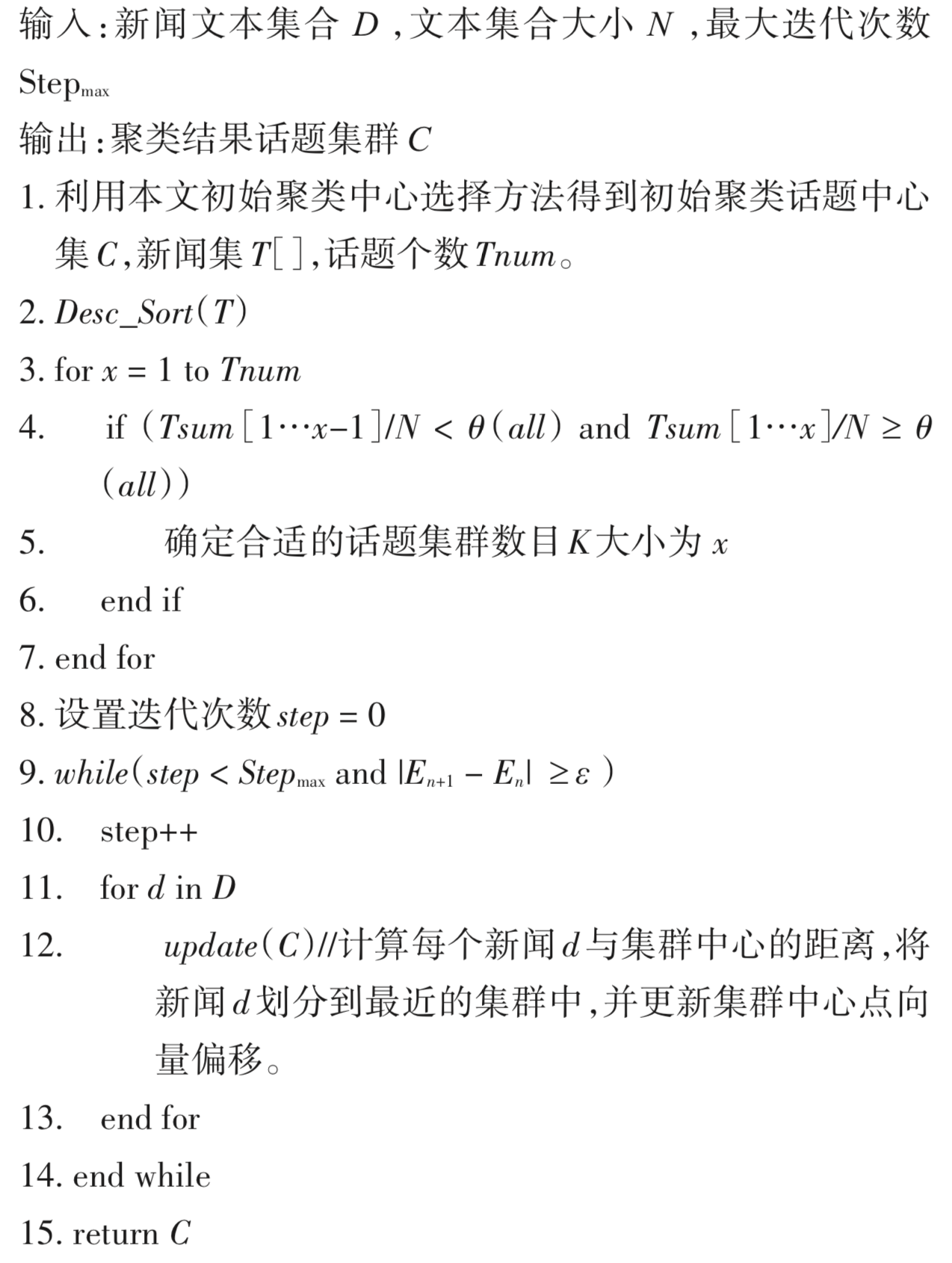


图10 基于改进 K-means的新闻聚类算法图

其中表示整个新闻文本集群中所有能够形成话题的新闻占总数的覆盖率。若输入数据含噪声数据较多，则应适当减少。迭代步骤中的E为K-means算法的标准测度函数，主要用来判断是否需要进行下一次迭代，一直到收敛停止。

* 1. **实验**
     1. **前期准备**

数据集选取从新浪新闻、网易新闻等门户网站爬取的当日产生的新闻文本，判断标准主要为评价话题集群内新闻报道内容相似度是否高，两话题之间新闻 报道内容区分度是否高。

平均相似度表明的是话题集群内的各新闻内容相似度的平均值，设新闻话题集群的大小为，则新闻话题集群的平均相似度定义如下:

（21）

(标准化互信息)表明的是话题集群间的新闻内容的相似程度,的定义如下：

（22）

其中，，分别表示向量和向量的信息熵，的信息熵的定义如下：

（23）

表示向量，向量的互信息值(Mutual Information),其定义如下：

（24）

表示和的联合分布概率，和分别表示和的概率函数。

测量值是根据准确率()和召回率()广泛用于信息检索和统计学分类领域的度量值。其定义如下:

（25）

* + 1. **结果分析**

为了验证本文算法在新闻话题聚类时的聚类效果优于传统 K-means 算法及其变种，我们将本文算法聚类结果与传统 K-means 算法，基于渐进中心 的 K-means 算法和基于二分的 K-means 算法的聚 类结果作比较。

我们对每个算法实验10次，每次对语料的输入顺序不同，在图11中的横轴上进行体现，纵轴表示新闻话题集群内部相似值。可以发现，这四种算法在集群内部的相似度上没有明显差异，说明本文算法在针对初始中心特定选择后并不会对每个话题集群内部相似性造成影响。

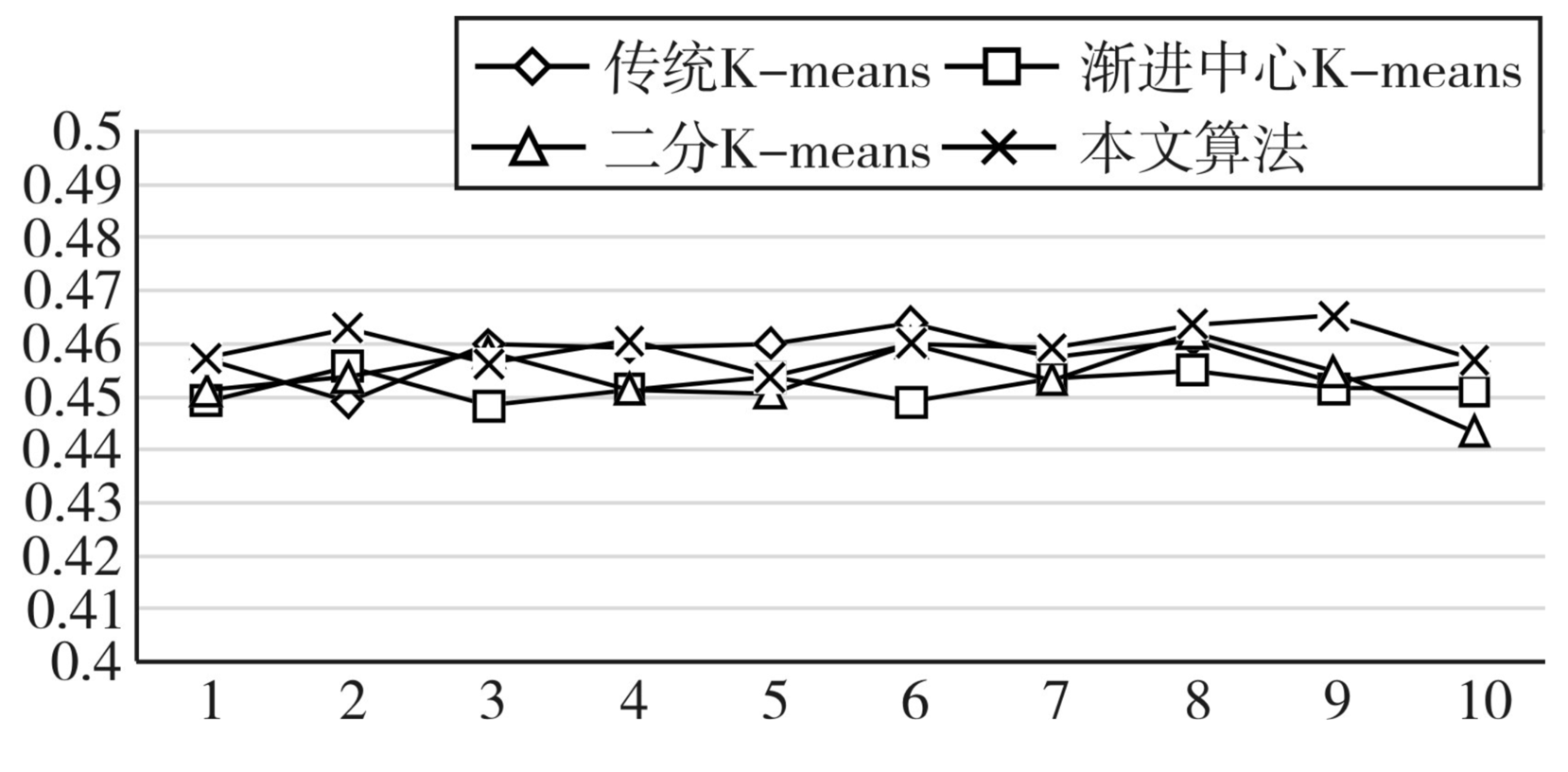


图11 10次试验相似度值

图12表示的是聚类结果中话题集群间的相似性度量，纵轴为值。可以看到，本文改进的算法在不同话题集群的相似性上表现出更加优秀的效果，不同的话题集群间区分度更明显，聚类效果更好 。

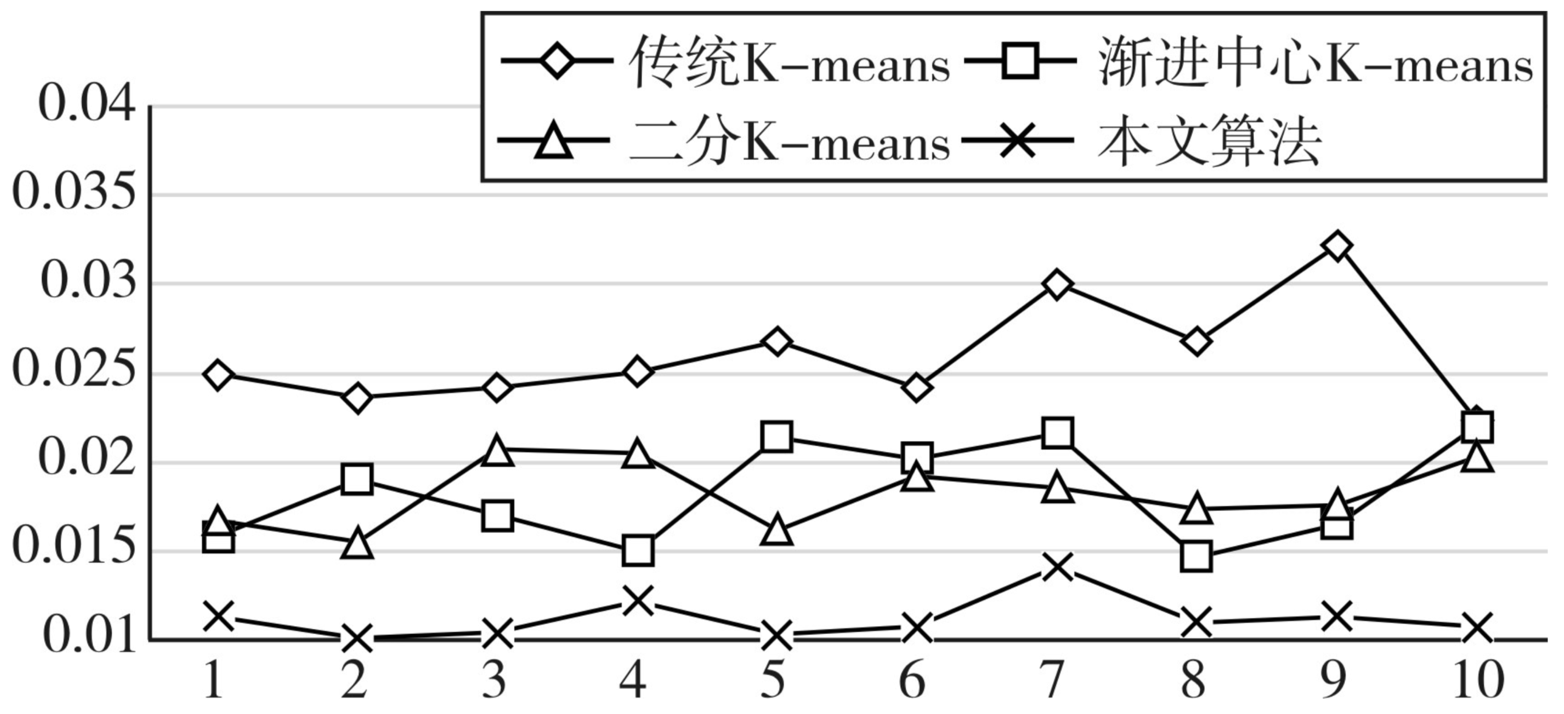


图12 10次试验值

如图13所示，本文算法表现效果在综合评价指标F值上也优于渐进中心K-means和二分K-means算法，且本文算法波动性更小，说明其受语料输入顺序的影响较低。

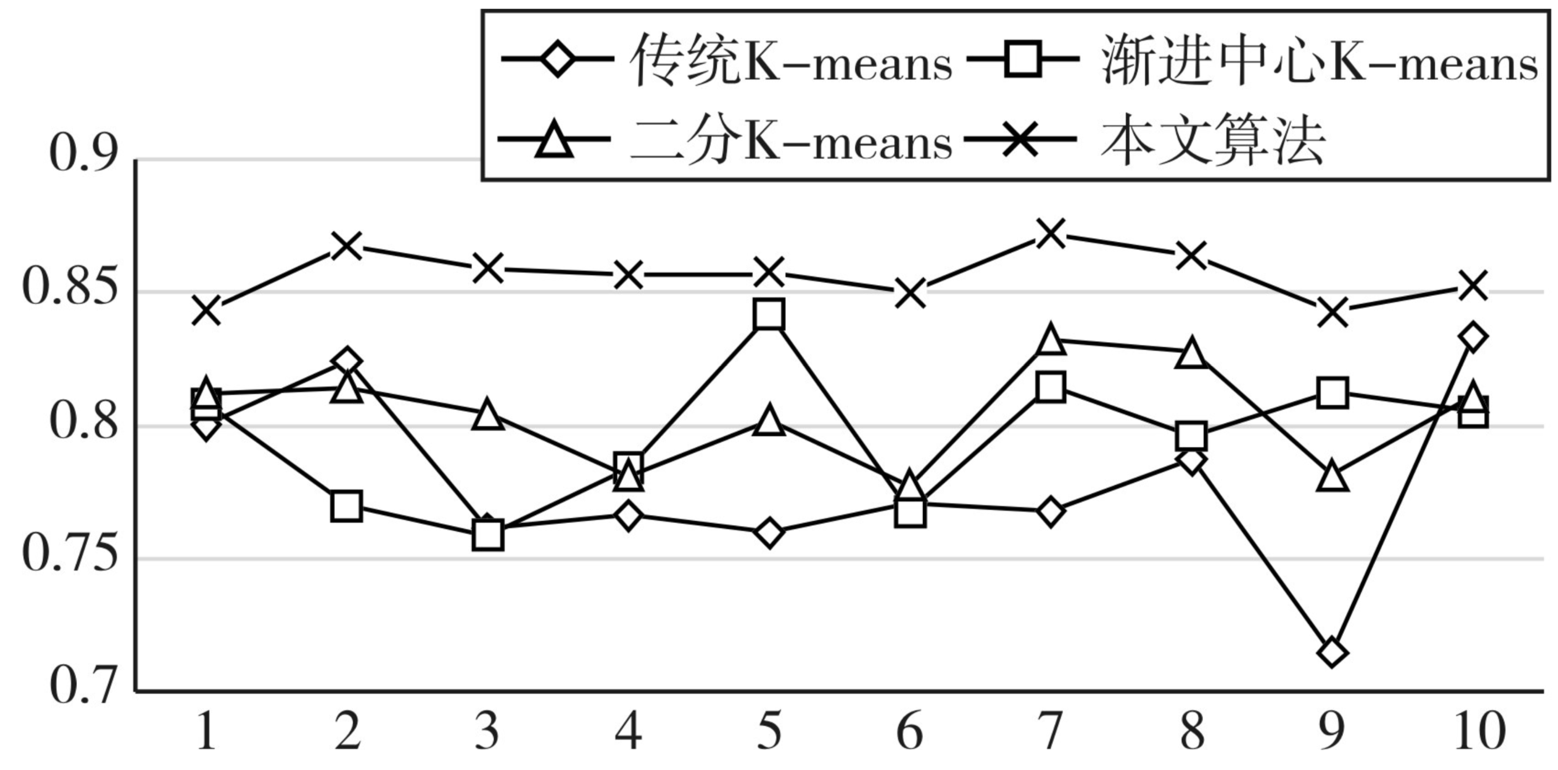


图13 10次试验测量值

如表2所示，在 NMI值与F-值测量上，该优化算法无论在平均值还是在标准差上都明显优于其它算法，其优越性如下：

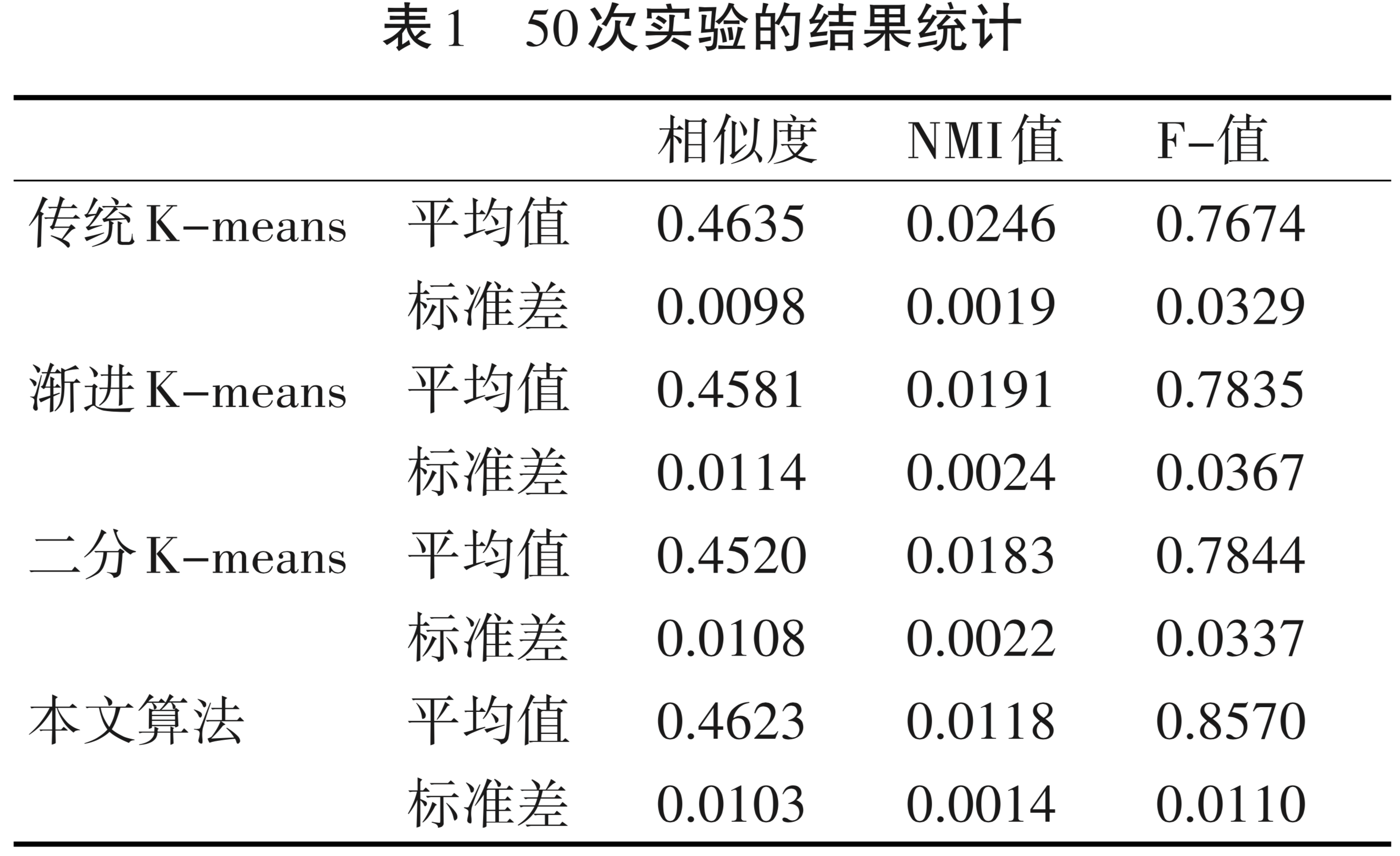
（1）在新闻话题的划分上要更加精确；

（2）聚类结果基本不受语料顺序干扰；

（3）其稳定性要优于传统、渐进和二分K-means算法；

（4）F-值综合测量显示出该优化算法能在话题集群内的相似度和话题集群间的区分度上更好地取得平衡。

表2 50次试验结果统计



1. **总结与展望**

本文对新闻热点话题的发现和演化分析进行了研究，并且将研究成果应用到了热点话题发现及演化分析系统上。

在TDT领域和网络话题自动发现领域，文本聚类技术是基础，而文本聚类质量的好坏关键在于文本相似性的描述，传统的基于词频统计的向量空间建模忽视了中文语义之间的联系，聚类结果存在明显语义缺陷，本文引入LDA主题模型对新闻报道进行主题分布层面上的文本建模，同时采用传统的TF-IDF特征词-权值模型和LDA模型对文本之间的相似度进行度量，并将新闻内容之间的相似性用于优化K-means聚类算法在新闻话题聚类时的初始聚类中心选择上，保证了聚类结果中各话题集群的区分度，使得改进后的新闻话题聚类算法不会陷入局部最优，保证了话题集群的完整性。

NMI测试结果表明了本文优化方法的有效性，测试结果也表明本文算法更加稳定，基本不受语料输入顺序的影响。同时，本文算法根据新闻话题集群的覆盖率提出了自生成话题集群分组数K的方法，测试结果表明这一优化是可行的，能够得到准确的K值。

但该算法具有一定的局限性，具体表现为如下几点：

1）对于大规模数据无法进行快速分析， 无论在建模方法还是数据的处理方面都有可以改进的地方；

2）对于话题的演化分析采用的方法比较单一，对于多核心话题描述模型可以使用多种话题演化分析方法进行融合，提高话题演化的描述能力；

3）本文对于热点话题的发展走势未做完整预测研究，未来可进一步对热点话题的发展走势进行行为预测分析，提前预测话题的走势。

对于K-means算法，新闻垃圾数据对聚类集群自生成分组K有一定影响，如何降低垃圾数据对分组K的影响将是进一步研究的重点。

**参考文献**

[1]Zahra S，Ghazanfar M A，Khalid A，et al. Novel centroid selection approaches for KMeans-clustering based recom⁃ mender systems[J]. Information Sciences，2015，320(C): 156-189.

[2]谭晋秀，何跃. 基于 K-means 文本聚类的新浪微博个 性化博文推荐研究[J]. 情报科学，2016(4):74-79.

[3]Wang J，Wang J，Ke Q，et al. Fast Approximate K-Means via Cluster Closures[J]. Multimedia Data Mining & Ana⁃ lytics，2015，157(10):3037-3044.

[4]Pavlas M. Fast Regularized Least Squares and k-means Clustering Method For Intrusion Detection Systems[J]. Ceskoslovenská Epidemiologie Mikrobiologie Imunologie， 2015，16(4):228-36.

[5]García M L L，García-Ródenas R，Gómez A G. K-means algorithms for functional data[J]. Neurocomputing，2015，151:231-245.

[6]J. Han，J. Tan，H. Huang，et al. Power-supplying block partition based on improved K-means clustering algorithm [J]. Dianli Zidonghua Shebei / Electric Power Automation Equipment，2015，35(6):123-129.

[7]Bharti K K，Singh P K. Chaotic Artificial Bee Colony for Text Clustering[J]. Soft Computing，2015:337-343.

[8]Forsati R，Keikha A，Shamsfard M. An improved bee colony optimization algorithm with an application to docu⁃ ment clustering[J]. Neurocomputing，2015，159(1): 9-26.

[9]Chabchoub Y，Fricker C. Classification of the vélib stations using Kmeans，Dynamic Time Wraping and DBA averaging method[C]// International Workshop on Computational Intelligence for Multimedia Understanding. IEEE，2015.

[10]中国互联网络信息中心.第38次中国互联网络发展状况统计报告[R].北京:2016 年8月.http://tc.people.com.cn/nl/2016/0803/c183008-28606650.html999.

[11]江华丽.中文分词算法研究与分析[J].物联网技术，2016(1):87-89.

[12]谢邦昌，朱建平，李毅.文本挖掘技术及其应用[M].厦门大学出版社，2016.