AR-Grams：一种应用于网络舆情热点发现的文本聚类方法

王贤明[1]，潘佳玲[1]，胡智文[2]
（1. 温州理工学院 数据科学与人工智能学院, 温州 325035； 2. 浙江工商大学 计算机与信息工程学院, 杭州 310018）

摘要：网络舆情热点发现是一种常用且处理速度要求较高的应用。针对网络舆情热点发现这一特殊应用场合，本文提出了一种基于随机 N-Gram 的文本聚类方法 AR-Grams。该方法通过随机 N-Gram 的文本相似度计算方法，确立待聚类文档集中各个初始聚类的标志文档并完成初步的聚类操作，继而通过聚类元素数阈值来确定初始聚类，并可根据实际情况确定是否执行聚类合并。该方法生成的聚类内聚性好，准确率高。另外，为了便于评估整体的聚类效果，提出了聚类的整体覆盖率和正确覆盖率。实验结果表明：与对比方法 DR-Grams 相比，在低阈值时，AR-Grams 的准确率、召回率、F-score、正确覆盖率分别提高了 11.9%、9.1%、10.2%和 9.2%，提升效果尤为明显；在高阈值时，效果基本相当；在整体上，前述 4 项指标则分别提高了 4.5%、2.9%、3.5%和 3.0%，优于对比方法 DR-Grams。
关键词：文本聚类; N-Gram; 网络舆情

AR-Grams: A novel text clustering approach to determining online public opinion of hot events

WANG Xian-ming[1], PAN Jia-Ling[1], HU Zhi-Wen[2]
(1. School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou 325035, China;
2. School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou, 310018, China)

Abstract: Determining online public opinion of hot events is a commonly used application with high processing speed requirements. To address the challenges of such scenarios, we propose a novel text clustering method AR-Grams based on N-Gram. AR-Grams approach first employs a text similarity algorithm based on random N-Grams to determine the symbolic documents of each initial cluster in the document sets to be clustered and complete the preliminary clustering. Then, the initial clustering is determined by the threshold of cluster element number. Moreover, whether to merge the initial clustering depends on actual scenarios. The clusters generated by the method have good cohesion and high accuracy. In addition, this paper also proposes the overall coverage rate and the correct coverage rate to evaluate the overall performance of text clustering. Arguably, AR-Grams approach shows better performance than that of the baseline approach DR-Grams. Concretely, the accuracy rate, recall rate, F-score, and correct coverage rate of AR-Grams respectively improve by 11.9%, 9.1%, 10.2%, and 9.2% when the threshold is small, and they are basically the same when the threshold is large. Overall, AR-Grams can respectively achieve 4.5%, 2.9%, 3.5%, and 3.0% improvements over the previous state-of-the-art performances of DR-Grams.
Keywords: text clustering; N-Gram; online public opinion

1 引言

在文本挖掘领域，文本聚类是一类常见而又重要的数据挖掘手段，同时也是很多其他挖掘操作的前置工作。顾名思义，聚类即按照某些特征和规则将整个数据集分成若干组的过程，各个组内元素在某些特征方面具有较高的相似性，而组间元素则在这些特征方面具有较大的差异性，所得到的各个组即为一个聚类，也常称之为“簇”。聚类作为一种无监督的机器学习方法，无需人工对数据进行标注和训练，自动化程度高。目前已被广泛应用于计算机科学、情报学、社会学、生物学等多个领域。随着互联网的高速发展，文本聚类在 Web 数据处理相关方面应用尤其广泛，例如推荐系统、网络舆情[1-2]、各类文本挖掘及相关应用[3-5]。

在诸多 Web 相关研究领域中，网络舆情研究近年来发展很快，是一个兼具实用价值和学术价值的综合性研究领域，被学术界和政府管理部门重视，吸引了计算机科学、情报学、社会学、新闻学、统计学等多个学科研究人员投入到相关研究中。在网络舆情研究中，其中一个重要的研究方向即网络舆情热点的发现。由于网络数据的海量性，导致网络热点的发现对聚类算法的实时性要求较高，计算资源消耗也大。不过正

项目基金：教育部人文社会科学研究青年基金项目“基于社交网络大数据的网络舆情涨落机制研究”(17YJCZH178)；国家社会科学基金项目“基于网络舆情大数据的主流媒体公信力和影响力测度及其建设研究”（19BTJ031）；媒体融合与传播国家重点实验室（中国传媒大学）开放课题“基于情感驱动的网络信息传播动力学建模与分析”
作者简介：王贤明(1979-)，男（汉族），湖北黄冈人，温州理工学院教授。E-mail:xmwung@ustc.edu；潘佳玲（2001-），女（汉族），浙江绍兴人，温州理工学院本科生；胡智文（1975-），男（汉族），湖北黄冈人，浙江工商大学教授。E-mail: huzhiwen@zjgsu.edu.cn。

是由于网络数据的海量性，一旦某个热点产生后，围绕该热点的大量媒体数据将迅速发布并传播开来，也就是说，由于相关热点数据非常多，无需获取其全部相关数据，而只需要获取其中一部分数据，且保证这部分数据足够“纯”，即足够分析出相关热点，这是一种典型的准确率重要性远大于召回率的情况。然而目前鲜有专门针对这种情况的聚类算法。

2 相关研究及问题

2.1 聚类方法

文本聚类的研究历史悠久，取得了丰硕的成果，相关聚类方法层出不穷。目前，较为知名的文本聚类方法如划分聚类、层次聚类、基于密度的聚类[6-7]等。近年来，基于语义的聚类[8-11]和深度学习的聚类[12-13]逐渐受到关注，尤其以后者更为明显。同时也有不少混合型方法或集成聚类方法[14-17]。此外，也有一些适用于特殊场合的聚类方法，例如目前针对短文本的聚类[18-20]也获得了不少关注。

在上述方法中，基本都需要特征项或词支撑，并且不同的特征或特征组合效果是不同的[21-22]，因而决定了特征选择[23-24]或降维[25-26]对聚类是一项重要的前置研究内容。对中文而言，往往离不开分词的支持[27]，相应的分词准确性问题也随之而来，最终也将影响聚类的速度、准确率和召回率。

N-Gram 是一种经典的统计语言模型，目前已被广泛使用于各种各样的文本应用场合[28-30]及非文本应用场合[31]。由于 N-Gram 的特点，因此可以应用于文本相似度的计算[32-33]。文献[34]提出了一种基于 N-Gram 相似度算法的文本聚类方法，该方法无需分词支持，对语言也无要求，速度和准确率可以方便地调控。其适用场景是：对准确率和速度要求较高，但对召回率要求次之。典型的应用如网络舆情实时热点发现。在舆情热点发现过程中，对准确率和速度要求是必然的；热点分析要求有一定量的相关主题文档即可进行，并不要求识别得足够全面，也就意味着对该聚类的召回率并无太高要求。不过该方法在聚类阈值相对较小时，初始聚类结果较为“粗糙”，且准确性也相对稍低，可能存在着将毫不相干的内容聚到同一个类中的弊端。

2.2 聚类的评估

文本聚类的评估较为困难，方法多样。例如采用专家人工评估、熵（Entropy）评估、准确率、召回率、F-score 等，其中尤以准确率、召回率最为普遍，它们评估的是每个单独的聚类，且一般都尽量在这两者间取得平衡。

由于在舆情热点分析类似的应用过程中，往往会同时得到多个聚类，并且在该评估过程中，准确率的重要性远高于召回率。本文基于实际需要和便利性，拟从当前的聚类评价指标构建综合性的评估指标。

本文方法优势及创新点如下：

1） 相较于常规聚类方法，本文方法由于是基于 N-Gram，避免了很多聚类方法中的分词、特征提取等操作，同时具备语言无关性，而且也可以轻松地通过调整阈值实现对聚类速度、聚类精细程度等的调控。
2） 本文方法所得初始聚类的“内聚性”强，相应的，各个初始聚类的准确率高。因而最终聚类往往准确率也较高。
3） 定义了适合本文聚类方法的综合评估指标。

3 方法及原理

3.1 应用于网络舆情热点发现的文本聚类方法

设原始文档集个数为 k，每个文档集对应一个主题，文档集分别记为 $D_1 = \{d_1^1, d_2^1, \cdots, d_{n_1}^1\}$，$D_2 = \{d_1^2, d_2^2, \cdots, d_{n_2}^2\}$，…，$D_{\mathrm{k}} = \{d_1^k, d_2^k, \cdots, d_{n_k}^k\}$。实验文档集为上述文档集的并集，在不必区分或者无法区分文档的归属时，可将文档集记为：$D' = \{d_1', d_2', \cdots, d_n'\}$，其中 $n = \sum_{i=1}^{k} n_i$，为文档集中的文档数。聚类过程中，文档相似度采用文献[32]中方法计算，相似度阈值为 T，即若文档相似度值不低于该值，则将这些文档归属到一个类中。聚类中文档数阈值为 C，即若某个初始聚类中的文档数不低于该值，则认定该初始聚类为一

个有效聚类，否则舍弃。

聚类的主要流程如下图 1 所示。

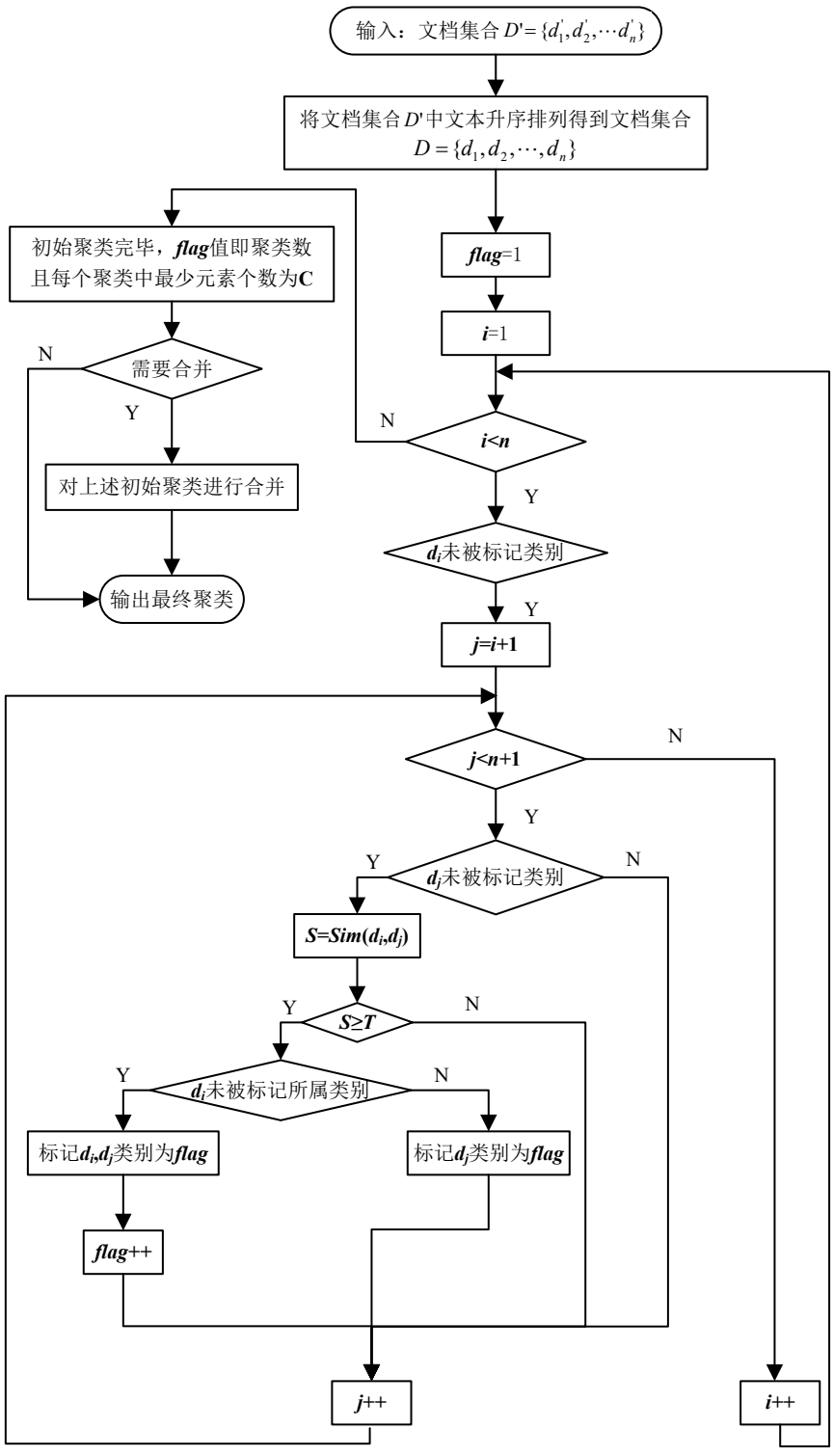

图 1　聚类过程

其中，上述 flag 变量既可以用于记录初始聚类完毕时所得的聚类数，也可以用作各聚类的序号。$S = Sim(d_i, d_j)$是文档d_i和d_j的相似度值，范围为[0,100]。聚类完毕，根据 flag 值即可知所获得的初始聚类个数，且每个初始聚类中最少元素个数为C。此处的初始聚类是指经过上述方法聚类后的直接聚类结果，以便和最终的聚类区分开。

经由上述方法聚类后，所得初始聚类结果可以直接用于类似网络热点识别之类的应用场景。倘若需将其聚类应用到其他更为广泛的聚类场合，则需要对上述初始聚类结果执行合并处理。所谓合并处理，即对各个初始聚类进行二次聚类。二次聚类可以通过两种方式进行。第一种是准确性更好的方式，即将各个初始聚类视为一个整体来对待，例如计算各个初始聚类中文档集的频繁项集，该频繁项集对应于该初始聚类，然后利用频繁项集的方法[35]即可完成初始聚类的合并，亦即完成最终聚类。第二种是一种快捷的方式，即以各个初始聚类中的最长文档作为该聚类的代表文档，并对各个代表文档进行聚类计算。若代表文档聚为一类，则意味着其对应的初始聚类可以合并为一个大的聚类。一般情况下，采用第二种方式也可以取得较为满意的结果。由于聚类的合并可以采用多种常规的聚类方法，因此不再赘述。

3.2 聚类覆盖率

设与原始文档集对应的各个合并聚类中元素数为 $n_1^p,n_2^p,\cdots,n_k^p$，正确的元素数分别为 $n_1^q,n_2^q,\cdots,n_k^q$，则聚类的整体覆盖率定义为：$C_a=\frac{\sum_{i=1}^{k}n_i^p}{n}$，即所有聚类中文档数之和与总文档数的比值；正确覆盖率定义为：$C_r=\frac{\sum_{i=1}^{k}n_i^q}{n}$，即所有聚类中正确的文档数之和与总文档数的比值。显然，聚类覆盖率和正确覆盖率的取值范围均为[0,1]，且满足$C_r \le C_a$，其中正确覆盖率表征了聚类的整体性能，其值越大，表明聚类整体效果越好。本文除了使用传统的准确率、召回率和 F-score 来讨论聚类结果外，还将使用聚类覆盖率指标对聚类结果进行讨论。

4 实验及结果分析

4.1 实验方案

为了便于比较，本文采用与文献[34]相同的实验数据、相似度计算参数和实验方案，其中文献[34]中方法记为 DR-Grams，本文方法记为 AR-Grams。

4.2 实验结果与分析

1）聚类阈值与初始聚类数的关系

初始聚类数是利用 AR-Grams 进行聚类后的直接聚类结果，亦即未进行聚类合并之前的聚类情况。相关实验结果如下图 2 所示。

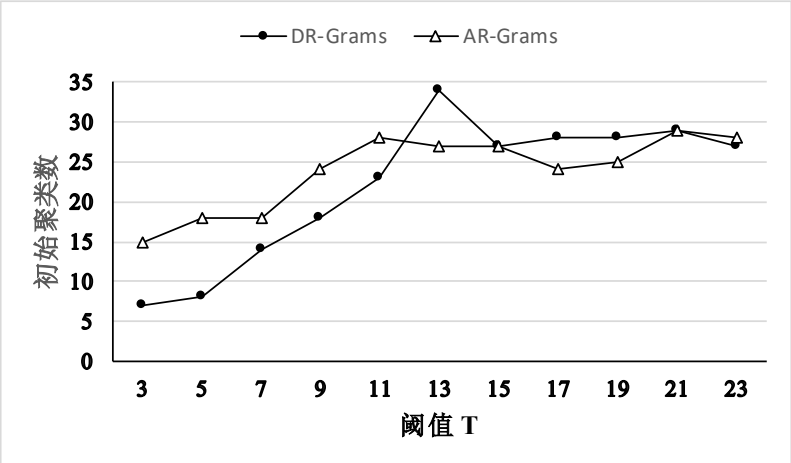

图 2 初始聚类数与阈值的关系

从该图可见，当阈值较小时，获得的初始聚类较少，随着阈值的增大，所得初始聚类逐渐增多，当阈值增大到一定范围时（对本例是[11,15]），聚类数呈现基本稳定的状态，但当阈值增到足够大时，聚类数开始逐渐下降。

呈现上述现象的原因在于：当阈值较小时，阈值对不同聚类元素的辨识度有限，且相对较容易受到因采用随机 n-Grams 相似度计算中的随机性影响，因而更容易将本不该隶属于一个类中的文档聚到一起，从而最终获得的聚类较少，容易推断，此时的准确率也应该相对较低。当阈值逐渐增大时，阈值的辨识度逐渐增大，各文档更容易被归属到其应该的聚类中，因而聚类相对更为准确，聚类数也就更多，这正是聚类数增多的原因。当阈值增大到一定范围时，此时可以较为准确地划分各个文档的类别归属，并且由于此时聚到同类中的文档确实是存在相当程度的重复，因而在阈值不是足够大时，一定程度的阈值变化是不会有太大影响的。这正是聚类数存在一段相对稳定区间的原因。并且，此阶段各个聚类的大小相对更大，同时各个聚类的准确率基本维持在 100%，该阶段正是适合于用作类似于网络热点分析相关研究或应用的时机。随着阈值的继续增大，只有几乎完全相同的文档才会被聚到一个类当中，不过完全相同的文档数毕竟有限，因而此时获得的聚类数将开始逐渐降低。需要交代的是，此时得到的初始聚类其实较多，不过只是有些聚类太小，即元素数在阈值C之下，因而被过滤掉了，留下的有效初始聚类数在减少。

另外，对比 AR-Grams 和 DR-Grams，可以发现两者随着阈值的变化趋势相同，但在不同的阈值阶段上，具体聚类数有所差异。在低阈值时，AR-Grams 获得的聚类相对更多，最为重要原因就在于 AR-Grams 在低阈值下聚类更为精细，不像 DR-Grams 聚类结果那么粗糙，因而获得的聚类数更多，相应的，整体上各聚类更小。但随着阈值的增大，阈值已能够准确地进行聚类而不至出错，因而两种聚类方法在高阈值时的表现基本相同。

此外，当初始聚类数趋于稳定时，意味着此时所对应的阈值T为较好的选择。根据这一特征，可以实现聚类过程中聚类阈值T的自动化确定。

2）聚类阈值与准确率、召回率及 F-score 的关系

准确率是经典的聚类评估指标，AR-Grams 聚类准确率结果如下图 3 所示。

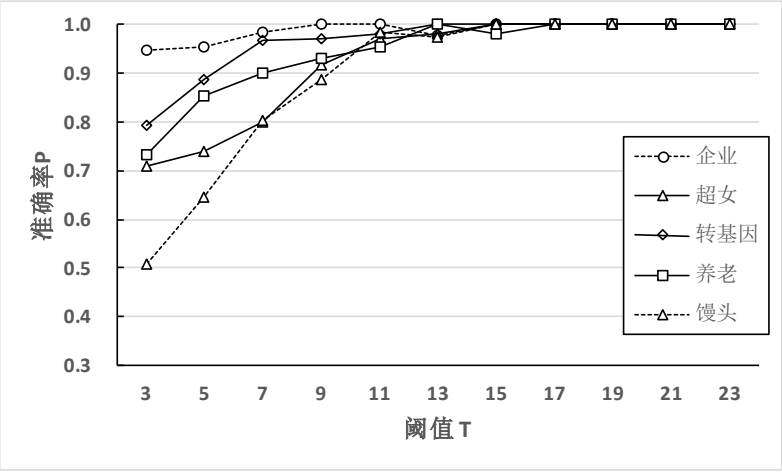

图 3　聚类阈值与准确率的关系

与 DR-Grams 聚类一样，阈值越小，各文档归属出错的可能性越大；阈值越大，各文档归属出错的可能性越小，聚类阈值对聚类结果起着决定性作用。对比 AR-Grams 和 DR-Grams 结果可见，两种方法所得结果的变化趋势相同，即准确率随着聚类阈值的增大而增大，直至为 100%。并且在阈值$T=11$时，准确率已经接近 100%，当阈值$T\geq 15$ 时，准确率几乎已为 100%。故从聚类准确率来看，聚类阈值在 AR-Grams 和 DR-Grams 下具备同样的作用。在 AR-Grams 聚类下，相似度阈值范围可初步确定在区间[11,17]。

针对各个单一数据集而言，在图中，“馒头”的准确率明显低于其他数据集，原因在于文档集 D 中的最小可聚类文档来自于“馒头”数据集，在 AR-Grams 聚类下，该文档将首先成为聚类标志文档，并将获得最多的与其他文档进行相似度计算的机会，因而也将纳入更多的文档到该类中，这就是“馒头”的最低准确率的根本性原因。这一点，是 AR-Grams 和 DR-Grams 的共同特性，即较先的可聚类标志文档所在的聚类往往具备较低的准确率。不过随着阈值的增大，该情况逐渐被改善。

与 DR-Grams 相比， AR-Grams 聚类在低阈值（$T\leq 9$）时准确率提高了 11.9%，在整体上则提高了

4.5%。由此可见，AR-Grams 在低阈值下的改进效果明显。原因正如前文所述，低阈值下的 DR-Grams 聚类结果较为粗糙，而 AR-Grams 结果则较为精细，精细的聚类结果其准确率必然高得多。实验结果显示 AR-Grams 下的聚类准确性整体高于 DR-Grams，高阈值下的结果相当，因而可以认为 AR-Grams 优于 DR-Grams。

聚类评价的另外一个重要指标为召回率，AR-Grams 聚类阈值与召回率关系的结果如下图 4 所示。

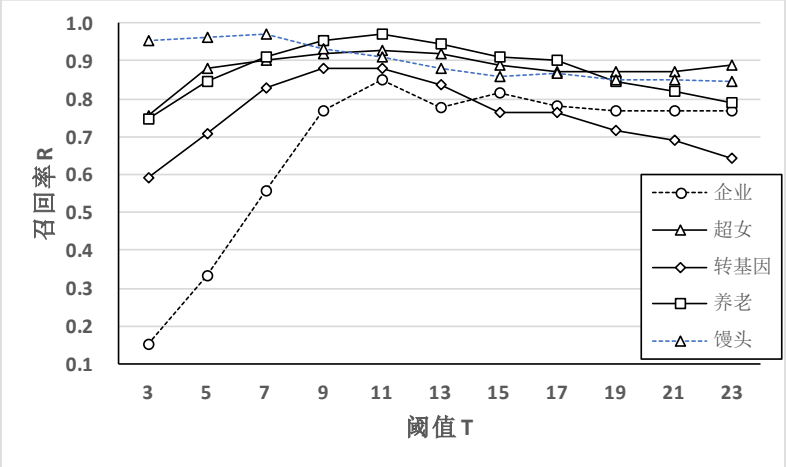

图 4　聚类阈值与召回率的关系

对比两种方法的召回率曲线可知，两种方法下聚类阈值和召回率存在相同的关系，即随着聚类阈值的增大，召回率呈现先增后降的态势，并且最佳聚类阈值范围为[9,11]。

和 DR-Grams 一样，在聚类阈值较小时（$T \le 9$），阈值作用归结为“类间纠错”，即阈值的增大，将逐渐减少文档被归属错误的可能性。但当T逐渐增大时（$T > 11$），文档的归属已基本完全正确，正如图 3 中所示，$T = 11$ 时的准确率已基本为 100%，因而此后的阈值作用将主要体现为把各个聚类划分为更为精细的、且准确率依然保持为 100%的更多小聚类，亦即“类内细分”的作用。类内的细分一方面将会获得更多稍小的聚类，同时又将使得较多过小的聚类被阈值C过滤或者一些单一的文件不被归属到任意聚类中，这正是召回率曲线下降的原因。

与 DR-Grams 相比，AR-Grams 在低阈值下的召回率提高了 9.1%，在整体上则提高了 2.9%。可见本文方法在低阈值时的改进作用明显。

综合性的评价指标 F-score 曲线如下图所示。

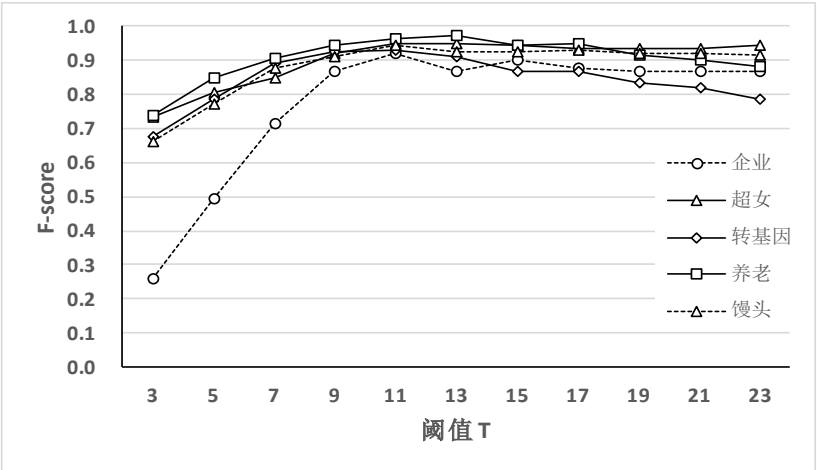

图 5　聚类阈值与 F-score 的关系

由上图可见，在整体上，F-score 曲线的升降趋势与召回率一致，即先升后降。这一点与 DR-Grams 聚类一样。但具体值方面，本文方法的 F-scoce 在低阈值时提高了 10.2%，在整体上则提高了 3.5%。这主要归功于 AR-Grams 的精细特性，在造就更多精细聚类的同时，提高了聚类效果。

2）聚类阈值和覆盖率的关系

整体覆盖率 C_a 和正确覆盖率 C_r 实验结果如图 6 所示。

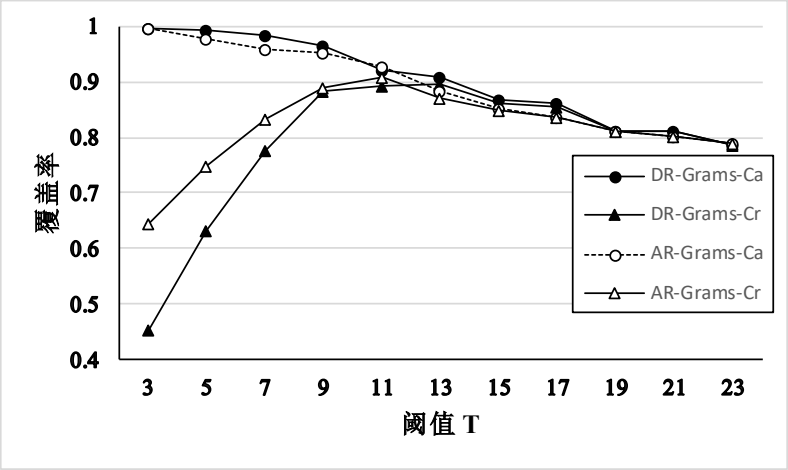

图 6　聚类阈值与覆盖率的关系

由上图可见：整体文档覆盖率随着聚类阈值的增加呈现单调递减趋势，正确文档覆盖率则呈现先升后降的趋势。显然，随着聚类阈值的增大，文档将更难以聚到一起，或者难以聚成较大的类。由于各个聚类对纳入该类文档的限制更为严格，这将导致越来越多的文档成为独立于任何聚类的个体文档，或者由于所含文档过少而无法被认定为有效聚类，在宏观上即呈现为整体文档覆盖率的持续下降。对正确文档覆盖率而言，则与上述情形有所不同。在阈值较小时，虽然绝大多数的文档都被归属到相关聚类中，但是正如前文所述，低阈值时的归属错误率极高，这一问题随着阈值的增大将逐渐缓解（即低阈值时阈值呈现为“类间纠错”功能），这正是正确文档覆盖率在开始阶段呈现增长趋势的原因。在阈值较大时，由于阈值的“类内细分”作用，诸多大类被分割为多个细小的聚类甚至一些独立的文档，在该过程中，越来越多的独立文档和极其细小的聚类被排除在有效聚类之外，宏观上即呈现为正确文档覆盖率的缓慢下降。这在另一个侧面再次印证了前文所论述的阈值的两种典型作用。当阈值增大到一定程度时，阈值已具备充分的辨识能力，可确保被归属到同一个类中的文档在实际上也的确是同类文档，此即当阈值较大时，两条曲线基本重合的原因。

和 DR-Grams 相比，本文方法的正确覆盖率提高了 9.2%，在整体上则提高了 3.0%，可见本文方法的主要效果表现在低阈值时对正确覆盖率的提升上，主要原因与前文的聚类准确率相同，不再赘述。

3）AR-Grams 聚类特性及应用场景解释

AR-Grams 聚类方法的特性可总结为：高准确率、低召回率、聚类精度和速度易于调控。该方法可通过调整相似度计算中 N-Gram 的数目及各项阈值来实现聚类精度和速度的调控，故决定了其可用于实时性较高的场合也可用于精度要求较高的场合，但并不能用于召回率较高的场合。另外由于该方法可以获取多个准确率高的聚类，通过其中的较大聚类即可完成类似网络热点发现之类的应用需求。这主要是由于在实际情况下，网络热点一旦产生，虽然围绕着一个热点话题的数据往往涉及多个方面，但其中往往存在着大量由于转载或其他原因而导致有较大重复率的文档。只要能把这些重复率较高的文档识别出来，就足以分析出相关热点，而并不需要识别出该热点所有相关数据，这正是本文聚类方法具有实用价值的客观支撑条件。本文聚类方法并不适用于类似于文献[36]中的艺术类数据聚类（包含音乐、舞蹈、书画等数据）。从本实验的初步聚类结果来看(即在不进行聚类合并条件下的聚类结果)，虽然聚类数较多，但其中较大的聚类却并不多，在实际进行网络热点分析时，只需利用其中的几个较大聚类即可实现。另外，由于实现海量网络数据中热点的识别只需要能够取得其中一个较大的且准确率高的聚类即可，至于该类中元素是多一些还是少一些，都不会影响热点分析结果，这就决定了虽然本文方法仍然是基于阈值进行聚类的，但是却对阈值要求却很低，只需要阈值较大，例如在 0.5 以上，但不要高于 0.9 即可。

另外，虽然采用本文方法时，取较小的阈值能够获得较少的聚类，不过由于此时各聚类中包含了一定

数量的归属错误的文档，这些对热点分析不利，故低阈值并不适合于进行热点分析。

5 结束语

本文提出的 AR-Grams 的文本聚类方法，具有语言无关性、高准确率、低召回率、聚类精度和速度易于调控等特点，相较于常规聚类方法，省却了繁琐的特征提取等操作，同时也避免了 DR-Grams 聚类可能导致的将毫不相干的文档聚到一个类中的缺陷，从而提高了低阈值下的准确率，因而也提高了 F-score，相应的也提高了聚类的正确覆盖率。这使得 AR-Grams 能在更广的阈值范围内应用于网络话题检测或者网络热点识别等场合。不过，在极低阈值时的聚类效果仍有待进行更为深入的研究。

参 考 文 献

[1] 马永军,杜禹阳,蔡润身.可调节聚类系数的 BBV 网络舆情传播模型研究[J].情报科学,2019,37(11):34-37,93.

[2] 夏火松,李保国,杨培.基于改进 K-means 聚类的在线新闻评论主题抽取[J].情报学报,2016,35(1):55-65.

[3] 徐小龙,杨春春.一种基于主题聚类的多文本自动摘要算法[J].南京邮电大学学报（自然科学版）,2018,38(5):70-78.

[4] 谭章禄,彭胜男,王兆刚.基于聚类分析的国内文本挖掘热点与趋势研究[J].情报学报, 2019,38(06):578-585.

[5] 谭章禄,王兆刚,胡翰,等.基于文本聚类的煤矿安全隐患类型挖掘研究[J].中国安全科学学报. 2019,29(03):145-148.

[6] 袁逸铭,刘宏志,李海生.基于密度峰值的改进 K-Means 文本聚类算法及其并行化[J].武汉大学学报(理学版). 2019,65(05):457-464.

[7] 刘颖莹,刘培玉,王智昊,等.一种基于密度峰值发现的文本聚类算法[J].山东大学学报(理学版),2016,51(1):65-70.

[8] Wei Song, Jiu Zhen Liang, Soon Cheol Park. Fuzzy control GA with a novel hybrid semantic similarity strategy for text clustering[J]. Information Sciences, 2014,273:156–170.

[9] Vivek Mehta,Seema Bawa, Jasmeet Singh. Stamantic clustering: Combining statistical and semantic features for clustering of large text datasets[J]. Expert Systems with Applications,2021,174(15):114710.

[10] 吴锦池,余维杰.融合知识库语义的文本聚类研究[J].情报杂志,2021,40(05):156-164.

[11] 钱志森,黄瑞章,魏琴,等.半监督语义动态文本聚类算法[J].电子科技大学学报,2019,48(06):803-808.

[12] 饶毓和,凌志浩.一种结合主题模型与段落向量的短文本聚类方法[J].华东理工大学学报(自然科学版). 2020,46(03):417-429.

[13] Gianni Costa, Riccardo Ortale. Jointly modeling and simultaneously discovering topics and clusters in text corpora using word vectors[J]. Information Sciences,2021,563:226-240.

[14] 杨玉娟,冯霞,王永利.QH-K:面向新闻文本主题抽取的改进 H-K 聚类算法[J].南京邮电大学学报（自然科学版),2020,40(1):82-88.

[15] Bharti Kusum Kumari, Singh P K. A three-stage unsupervised dimension reduction method for text clustering[J]. Journal of Computational Science,2014, 5:156-169.

[16] 张颖怡,章成志,陈果.基于关键词的学术文本聚类集成研究[J].情报学报,2019,38(08): 860-871.

[17] Karpagalingam Thirumoorthy, Karuppaiah Muneeswaran. A hybrid approach for text document clustering using Jaya optimization algorithm[J]. Expert Systems with Applications,2021, 178(15):115040.

[18] 贺超波,汤庸,张琼,等.基于增量式鲁棒非负矩阵分解的短文本在线聚类[J].电子学报,2019,47(05): 1086-1093.

[19] Wen Aihong,Yan Nan,Xu Caocao. An efficient Particle Swarm Optimization approach to cluster short texts[J]. Cluster Computing ,2019,22:S4119–S412.

[20] 贾瑞玉,陈胜发.结合新概念分解和频繁词集的短文本聚类[J].小型微型计算机系统,2020,41(06): 1321-1326.

[21] 赵华茗,余丽,周强.基于均值漂移算法的文本聚类数目优化研究[J].数据分析与知识发现,2019,(9):27-35.

[22] 张旭,孙玉伟,成颖.不同特征对文本聚类效果的比较研究——以新闻文本为例[J].情报理论与实践,2020,43(1): 169-176.

[23]田夏利,熊堂.融入新的特征选择机制的文本数据聚类算法[J].计算机工程与设计,2021,42(3):734-741.

[24] Kusum Kumari Bharti, Pramod Kumar Singh. Opposition chaotic fitness mutation based adaptive inertia weight BPSO for feature selection in text clustering[J]. Applied Soft Computing,2016,43: 20-34.

[25] Mohamed A.A. An effective dimension reduction algorithm for clustering Arabic text[J]. Egyptian Informatics Journal, 2020,21(1):1-5.

[26] Kusum Kumari Bharti, Pramod Kumar Singh. Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering[J]. Expert Systems with Applications, 2015,42:3105-3114.

[27] 祖坤琳,赵铭伟,林鸿飞.基于有序聚类的专利知识演化研究[J].计算机工程与科学,2016,38 (4):786-791.

[28] Alessandro Cucchiarelli, Christian Morbidoni, Luca Spalazzi, et al. Algorithmically generated malicious domain names detection based on n-grams features[J]. Expert Systems with Applications,2021,170(15):114551.

[29] Nidal Nasser, Lutful Karim, Ahmed El Ouadrhiri, et al. n-Gram based language processing using Twitter dataset to identify

COVID-19 patients[J]. Sustainable Cities and Society,2021,72:103048.
[30] 李超,刘辉.一种基于关联分析与 N-Gram 的错误参数检测方法[J].软件学报,2018,29(8):2243-2257.
[31] 任卓君,陈光,卢文科.基于 N-gram 特征的恶意代码可视化方法[J].电子学报,2019,47(10):2108-2115.
[32] 王贤明,胡智文,谷琼.一种基于随机 n-Grams 的文本相似度计算方法[J].情报学报,2013,32(7): 716-723.
[33] 黄贤英,谢晋,龙姝言.基于公共词块及 N-gram 模型的问句相似度算法[J].重庆理工大学学报(自然科学),2017,31(10):175-179,197.
[34] 王贤明,谷琼,胡智文.基于 R-Grams 的文本聚类方法[J].计算机应用, 2015,35(11):3130-3134.
[35] ZHANG W,YOSHIDA T,TANG X, Wang Q．Text clustering using frequent itemsets[J]．Knowledge-Based Systems,2010, 23(5):379-388.
[36] 陈笑蓉,刘作国.文本聚类的重构策略研究[J].中文信息学报,2016,30(2):189-195.