# 基于层次聚类的网络新闻热点发现*

彭楠赟，王厚峰，凌晨添
北京大学 计算语言学研究所，北京 100871
E-mail: pengnanyun@pku.edu.cn; wanghf@pku.edu.cn; lct@pku.edu.cn

摘 要：网络新闻热点发现的主要目的是从海量互联网数据中发现人们感兴趣的热点话题。在已有研究中，主要采用基于单篇报道的增量聚类方法。本文则提出一套针对单日新闻进行层次聚类，发现每日热点，再对热点进行增量聚类的框架。在对每日新闻的层次聚类中，本文定义了类内凝聚度指标，并提出基于类内凝聚度的聚类阈值确定策略。实验证明，本文提出的方案在相关任务中都取得了令人满意的效果。

关键词：特征选取；层次聚类；阈值确定；热点发现

# Event Mining in On-line News Based on Hierarchical Clustering

Peng Nanyun, Wang Houfeng, Ling Chentian
Institute of Computational Linguistics, Peking University, Beijing 100871
E-mail: pengnanyun@pku.edu.cn; wanghf@pku.edu.cn; lct@pku.edu.cn

**Abstract:** The main goal of event mining in on-line news is to discover the hot topics which are frequently reported. Previous studies usually do a single pass incremental clustering algorithm to every newly arrived document to detect novelties. In this paper, we innovatively proposed a hybrid clustering algorithm which applied hierarchical clustering on daily news first, then do incremental clustering on daily hot topics. This hybrid algorithm reduces the error rate caused by limited information in the classical incremental clustering. To tackle the threshold determination problem in hierarchical clustering employed in daily hot topics detection, we defined a cohesion index within a topic. Consequently, we proposed a threshold determination policy based on this cohesion index and got satisfactory results on our data sets.

**Keywords:** feature selection; hierarchical clustering; threshold determination; event detection

## 1 引言

网络新闻中常出现人们普遍关注的热点话题。它们一般由一条普通的新闻报道开始，短时间内受到强烈关注和反复报道。由此开启了自然语言处理的一个新课题——网络新闻热点发现。这一课题有着广泛的应用前景：如网络舆情监控，网络信息安全，证券市场分析，行业调研等[1]。

网络新闻热点发现任务已受到学界广泛关注，已有方法主要建立在单路径增量聚类的基础上[1][2]。即对每篇新进文章，计算它与已有热点的相似度，若超过某一阈值，则将此文章划归与其相似度最高的一个已知热点；否则认为此文章代表一个新热点。在此框架下，很多研究将重点放在文本表示上[3][4][5][6]，试图通过抽取更合适的特征、改进 *tf-idf* 和相似度算法，来获得更好的聚类效果；也有研究改进传统的增量聚类，如吴永辉等使用了结合 LDA 模型和仿射传播（AP）的自适应聚类算法[7]，J. Zhang 等使用了概率模型[8][9]；此外，学界还关注适合网络新闻热点发现的新型聚类模型：如 Sayyadi 等使用了关键词图模型[10]，李建超使用了有向概率图模型[11]。

本文将网络新闻看作按时序切割的文章集合。在此视角下，完成热点发现任务可先对各时间单元内的文本集进行热点提取，再对所得热点进行增量聚类。受篇幅所限，本文只介绍每日新闻[1]热点提取。它建立在层次聚类基础之上，选用了更适合新闻热点发现的文本特征和 *tf-idf* 算法；采用

* 本文受国家自然科学基金资助，项目号：60973053，91024009。

[1] 本文以天为单位切割网络新闻流。

基于类内凝聚度的阈值确定方案；选取聚类结果中包含文章数超过一定阈值的类作为热点，并取其核心关键词作为热点表示。

本文第二章介绍本实验中采取的预处理策略和文本表示模型；第三章介绍本文提出的基于类内凝聚度的层次聚类阈值确定策略；第四章对实验结果进行了分析；第五章是总结与展望。

## 2 网络新闻表示模型

本文采用经典的向量空间模型表示网络新闻文本。鉴于新闻实时性强，特征及其权重的表示需要实现增量式；为提高运算效率，有必要筛选特征以减小向量空间的维度。为此，本文对传统的 *tf-idf* 算法进行了一定改进，使其能适应增量聚类的要求，提高聚类效率。

### 2.1 向量空间模型

向量空间模型将文本看作由一组正交特征向量组成的向量空间。假设所有文本的特征总数是 $M$，则每一个文本 $d_i$ 被表示为一个 $M$ 维的特征向量 $V(d_i)$。其中：

$$V(d_i)=(f_1,w_1(d_i);f_2,w_2(d_i);\ldots f_M,w_M(d_i)) \tag{2.1}$$

$f_j(j=1,2,..,M)$为第 $j$ 项特征；$w_j(d_i)(j=1,2,..,M)$为特征 $f_j$ 在文本 $d_i$ 中的权值。一般采用 *tf-idf* 算法得到：

$$W_j(d)=\frac{tf_j(d)\cdot\log(N/n_j)}{\sqrt{\sum_j(tf_j(d)\cdot\log(N/n_j))^2}} \tag{2.2}$$

公式(2.2)中的 $tf_j(d_i)$为特征 $f_j$ 在文章 $d_i$ 中出现的频率，$N$ 为文章总数，$n_j$ 为包含特征 $f_j$ 的文章数。为了快速处理海量网络文本信息，本文考虑了文本特征选择；同时，对传统的 *tf-idf* 算法进行了一定的改进，使其能适应增量聚类的要求。

### 2.2 特征选取策略

本文以词-词性对作为文本的特征单位。考虑到聚类效率和数据稀疏问题，只用了一元模型。只取名、动词性词语作为候选特征，并滤掉所有停用词和新闻文本中的高频词，如图 2.1：

图片/n　新华网/nt　新闻/n　新闻网/n　记者/n　新华网/nt　报道/v

新华社/nt　图片/n　图文/n　滚动/vn　日报/n　快讯/n

图 2.1　新闻文本中的高频词

上述高频词几乎在所有新闻文本中都出现，区分性不强。考虑到单字词携带的信息量较少，本实验也不予考虑。此外，还滤掉了文档频率低于 5（即出现在少于 5 篇文章中）的特征。因为能成为热点的话题，会在不同新闻中出现，自然会有较高的文档频率。此处 5 是一个经验性阈值。

### 2.3 改进的 *tf-idf* 算法

在新闻文本中，标题一般具有较强代表性；此外，新闻强调 5W，即时间、地点、人物、起因、事件五大要素。基于上述因素，本实验从两方面修改了传统的 *tf-idf* 算法：1）加大标题中出现的特征权重（使其变为原来的 2 倍）；2）对表示地点、人物的词加大权重，（使其变为原来的 1.5 倍）。修改后的 *tf* 计算公式为：

$$tf_j=(2Ct_j+1.5Cp,l_j+Cn_j)/w \tag{2.3}$$

其中 $Ct_j$ 表示特征 $j$ 出现在标题中的次数；$Cp,l_j$ 表示特征 $j$ 出现在人物、地点词中的次数；$Cn_j$ 表示特征 $j$ 出现在其他位置的次数。$W$ 表示一篇文章的特征总数。

网络新闻具有实时性，计算较早的文章特征的 *idf* 值时无法预料该特征在后来新闻中的文档频率，因此牵涉到 *idf* 值估算问题。本实验采用以天为单位进行聚类和热点发现的策略，较好的缓解

了这一问题：一天的新闻文本能达到数千甚至上万篇，直接采用当天特征的 *idf* 值可认为是对实际 *idf* 值的较好近似。每天的新闻处理完毕后，存储累计文档总数以及特征的累计文档频率。对之后的新闻计算 *idf* 值时，将之前的所有文章加入。这样就形成了如下的增量 *idf* 值算法：

$$idf_j = \frac{\sum_{j=1}^{i-1} N_j + N_i}{\sum_{j=1}^{i-1} n_j + n_i} \tag{2.4}$$

其中 $N_i$ 表示一天中的文章总数，$n_i$ 表示包含特征 $j$ 的文章数。

# 3 基于聚类的热点发现

网络新闻热点发现主要采用文本聚类算法。传统聚类算法包括划分聚类、层次聚类和密度聚类等。新闻热点发现需要较高效率和能以增量方式处理动态文本，为传统聚类算法带来一定挑战。

## 3.1 混合聚类算法设计

我们设计了一种结合层次聚类和增量聚类的算法，以期在效率和准确性之间取得平衡，这里不妨称为混合聚类算法。基本思想是：将网络新闻文本以天为单位进行划分，对每天的新闻采用层次聚类算法进行聚类，再对得到的话题簇进行传统的增量聚类。主要优点在于：1）每天的新闻报道数量大，提供了丰富的特征信息；2）在一天的范围内进行聚类，能获得比增量聚类更好的效果；3）层次聚类算法没有反复迭代过程，效率相对较高；4）将增量聚类运用在一个话题簇上而不是单篇文档，能获得更多信息，提高增量聚类的准确性。混合聚类算法具体过程如下。

1 以天为单位将网络新闻分为若干个增量 $N_i$

2 对一个增量进行层次聚类，以相似度为阈值确定聚类终止条件，将每天的新闻划分到 k 个类中

3 规定每个增量内部得到的聚类簇包含文章数大于某阈值时，该簇为热点话题。

4 调用话题表示算法获取其特征向量表示

5 如果 $N_i$ 为第一个增量，则所提取热点话题都为新话题，存入话题库，转到步骤 2；否则，将所得到的话题与话题库中已存在的话题进行比较和增量聚类，如果相似度大于某阈值，则将此话题归入已存在话题，并更新话题库特征向量；如果相似度均小于阈值，则认为此话题为新话题，存入话题库。重复处理所有增量，直至全部处理完为止。

图 3.1 混合聚类算法

受篇幅所限，本文只介绍单日新闻的层次聚类及热点发现。层次聚类面临聚类终止条件确定问题。传统方法一般采用给定类数或给定距离阈值来确定最终聚类结果。但对于网络新闻，每天的热点数量不该是固定的，因此采用给定类数的策略不合适。本实验采用阈值作为层次聚类终止条件。为了确定阈值，本文使用了“类内凝聚度”指标，根据此指标最终确定聚类终止阈值。

## 3.2 基于类内凝聚度的阈值确定策略

聚类将相似的文档聚合在一起，不相似的文档分散到不同的类中。因此，阈值的选择应保证聚类结果满足类内凝聚，类间距离大。为此，本实验层次聚类中采用组平均(GAC)作为距离度量方式；同时引入了“类内凝聚度”指标来度量类内文章的相似性，作为阈值确定的依据。

为了定义“类内凝聚度”，需要先引入两个概念：核心代表特征和核心文章。

**核心代表特征**是指在聚类结果中，某一类所有文章中累计 *tf-idf* 值最高的 20 个特征。

**核心文章**是指包含 m 个以上核心代表特征的文章。在本文中 m 取 3，是一个经验值。

于是，类内凝聚度指标定义如下：

**类内凝聚度**是衡量一个类主题显著性的指标，记为 $\gamma$ ($\gamma \in [0,1]$)，其中

$$\gamma = c_i / N \tag{3.1}$$

公式(3.1)中，$C_i$表示一个类的核心文章数，$N$表示该类文章总数。显然，$\gamma$值越大，表明该类核心代表特征的核心文章比例越大，也即主题越显著。

网络新闻的热点应该同时具备两大特征：1）报道量大；2）主题显著。本文使用了基于类内凝聚度的聚类阈值确定策略，以尽量保证聚类结果话题较显著。具体策略如下：

对一个给定的阈值 $\theta$[1]，取出聚类结果中文章数超过 60 的类[2]作为候选热点；分别计算这些候选热点簇的类内凝聚度 $\gamma_i$，并对其求平均，有：

$$\overline{\gamma} = \sum_{i=1}^{k} \gamma_i / N \tag{3.2}$$

若 $\overline{\gamma}$ <0.8，则减小阈值 $\theta$，以拆分内部距离较大的类，增加类内凝聚度；若 $\overline{\gamma}$ >0.8，则增大阈值 $\theta$，以合并一些相似度较高的类。总之，调整阈值 $\theta$ 以使 $\overline{\gamma} \approx 0.8$，作为聚类终止条件。

选择 $\overline{\gamma} \approx 0.8$ 是对主题显著性和热点规模的平衡：$\overline{\gamma}$ 越大，得到的聚类簇主题越显著，但相应的热点数也越少。实验结果表明，候选热点平均凝聚度 $\overline{\gamma} \approx 0.8$ 时，每天发现的新闻热点数将在 5-10 个之间，比较符合预期。在实际操作中，系统并没有为每一天的聚类都选定阈值 $\theta$，而是随机取一天的新闻，调整阈值使当天的候选热点平局类内凝聚度 $\overline{\gamma} \approx 0.8$，将这个阈值 $\theta^*$ 运用于之后的所有聚类，这将有利于提高算法效率。

需要说明的是，针对新闻热点话题发现，本实验并不关注整体的聚类结果，只关心“候选热点”，即被反复转载，相似文章数多于 60 篇的类。热点提取也只针对这些类内的文章。聚类结果中其余小类的文章被简单忽略。实验结果表明，在相同的阈值 $\theta^*$ 下，虽然每日聚类结果总类数差别较大，而热点数却稳定在 5-10 个之间。这从侧面反映出本实验简化的有效性。

## 4 实验与分析

本实验所用语料选自网络抓取的 2008 年腾讯新闻网新闻，分别取了 5 月 12 日至 5 月 21 日，8 月 12 日至 8 月 21 日及 9 月 12 日至 9 月 21 日三个时间段的文章。每个时间段跨度均为 10 天，包含新闻数分别为 27887 篇，22825 篇和 28786 篇。选这 3 个时间段，是因为其间有几个显著的新闻话题和热点：5 月的汶川大地震，8 月的奥运会和三鹿奶粉事件，用这段时间的新闻能帮助从直观上评价实验结果的好坏。9 月的新闻热点提取是用来与 5 月、8 月的新闻作对比的。

本文提出了“类内凝聚度”概念，并在单日热点发现的层次聚类算法中设计了基于类内凝聚度的阈值确定方案，取得了理想效果。为证明“类内凝聚度”这一指标有意义，本文先考察类内凝聚度和聚类距离阈值的关系，如图 4.1。由图可见，类内凝聚度随相似度阈值单调变化。阈值越大，类内凝聚度越低，符合直觉。因此，将类内凝聚度用作调整阈值的依据是有效的。选择 80% 为本实验目标，是因为虽然类内凝聚度越大话题一致性越高，但同时也限制了聚类簇的规模——对类内相似度要求过高会导致一些原本相关的话题被划分开——这将导致一些热点被遗漏。

此外，如 3.2 节所述，为提高算法效率，系统并没有依据类内凝聚度为每一天的聚类选定阈值 $\theta$，而是随机取一天的新闻[3]，调整阈值使当天的候选热点平局类内凝聚度 $\overline{\gamma} \approx 0.8$，再将这个阈值 $\theta^*$ 运用于之后的所有聚类。为了证明这种近似的有效性，实验统计了在选定阈值下，所有实验语料中的热点类内凝聚度宏平均，得到图 4.2。

如图 4.2 所示，在 5，8，9 三个月共 30 天的新闻文本中，相同阈值下聚类得出的每日热点类内凝聚度宏平均都较为接近，稳定在 75%到 85%之间。从这一结果看，本实验所作的简化可行。

---

[1] θ 表示两个类之间的距离，θ 越大，距离越大，相似度越小。本实验中采用的 cosine 距离。

[2] 60 为热点簇所包含新闻数的阈值，即单日超过 60 篇报道的新闻才是热点。这是一个经验值。

[3] 在实际应用中，为了取得更好的效果，可以随机抽取多天的新闻，调整阈值使它们的类内凝聚度的宏平均在 80% 左右。取这个阈值运用到所有数据上。

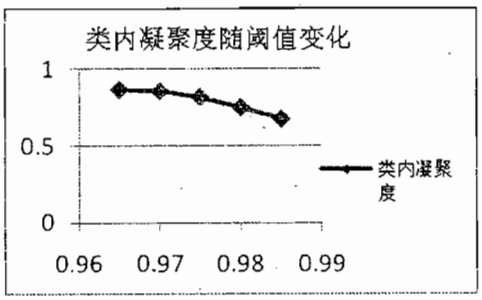


图 4.1 类内凝聚度随聚类阈值变化情况

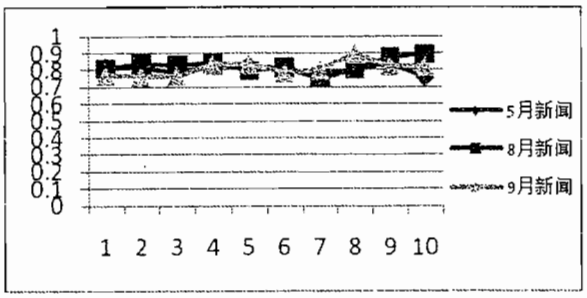


图 4.2 类内凝聚度宏平均

本实验基于类内凝聚度为单日新闻聚类选定了阈值，并规定最终聚类结果中，包含超过 60 篇新闻的团块被选为“每日热点”。下面分别从 5 月、8 月、9 月的新闻中各随机抽取 1 天，来展示系统挖掘出的当日热点。如图 4.3-图 4.5。

图 4.3 5 月 18 日热点话题展示

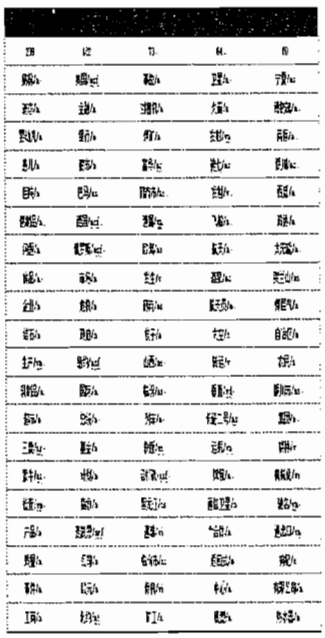

图 4.4 8 月 20 日热点话题展示

图 4.5 9 月 15 日热点话题展示

图 4.3，4.4，4.5 中每一列为一个热点话题，用它们的“核心特征”表示，第一行代表该热点的报道规模，从图中可以看到，5 月 18 日最热门新闻是地震灾后重建；8 月 20 日最热门新闻是三鹿奶粉事件；9 月 15 日最热门新闻是美国金融危机。经考证，8 月 20 日的热点没有奥运相关新闻，是因为当日没有热门赛事夺冠，加之奥运会接近尾声，相关报道减少。在 8 月其他几日的新闻中都对奥运有广泛报道。另外值得关注的还有 8 月 20 日热点四，关于神七发射的报道。神七是 2008 年 9 月 25 日发射的，但 8 月 20 日就有了成规模的报道，可见一些新闻热点的报道是跨越一定时间段的。相似的是 9 月 15 日热点六，关于奶粉质检的报道，它是之前三鹿奶粉三聚氰胺新闻的后续。综合看来，本文所提出的新闻热点发现算法，在每日热点的发掘中是有效的。

## 5 总结与展望

本文针对网络新闻热点发现任务，提取了更适合发现热点的文本特征，对 *tf-idf* 算法进行了修改，并支持增量计算。此外，设计了一种结合传统层次聚类和增量聚类的混合聚类方法；针对层

次聚类阈值确定问题，定义了核心代表特征，核心文章和类内凝聚度等概念，并提出了基于类内凝聚度的聚类阈值确定策略。实验结果表明，本文的算法在单日新闻热点提取中有较好效果。

由于篇幅限制，本文只介绍了混合聚类算法中针对单日新闻热点发现的层次聚类算法。针对单日热点的增量聚类将另文讨论。此外，一段时间的新闻热点发现还牵涉到1）“热点”定义问题：即如何确定某一话题是否属于新闻热点。2）“热点”排序问题，即如何确定一个话题的“热度”及其排名。这些都是今后工作需要研究的。

## 参 考 文 献


[1] 李保利，俞士汶. 话题识别与跟踪研究 [J]. 计算机工程与应用, 2003.

[2] 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.

[3] Y. Yang, T. Pierce, and J.G. Garbonell. A study on retrospective and on-line event detection [C]. In proc. of the SIGIR Conference on Research and Development in Information Retrieval, 1998.

[4] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking [C]. In proc. of SIGIR Conference on Research and Development in Information Retrieval, 1998.

[5] Y. Yang and J. Z. et al. Topic-conditioned novelty detection [C]. In Proc. of the SIGKDD international conference on Knowledge discovery and data mining, 2002.

[6] G. Kumaran and J. Allan. Text classification and named entities for new event detection [C]. In Proc of the SIGIR conference on Research and development in information retrieval, 2004.

[7] 吴永辉，王晓龙，丁宇新，徐军，郭鸿志. 基于主题的自适应、在线网络热点发现方法及新闻推荐系统[J]. 电子学报, 2010.

[8] J. Zhang, Z. Ghahramani, and Y. Yang. A probabilistic model for online document clustering with application to novelty detection [C]. Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference. MIT Press, Cambridge, MA, USA: Saul, Weiss, Y. and Bottou, L., (eds.), PP. 1617-1624. 2005.

[9] Z. Li, B. Wang, M. Li, and W. Ma. **A probabilistic model for retrospective news event detection** [C]. In Proc of the SIGIR conference on Research and development in information retrieval, 2005.

[10] H. Sayyadi, M. Hurst and A. Maykov. Event Detection and Tracking in Social Streams [C]. In Proc of the Third International ICWSM Conference, 2009.

[11] 李建超. 网页在线聚类的研究与实现[D]. 上海交通大学. 2007.