增量采样聚类驱动的新闻事件发现

陈晓琪， 谢振平， 刘渊

引用本文:
陈晓琪, 谢振平, 刘渊. 增量采样聚类驱动的新闻事件发现[J]. 智能系统学报, 2020, 15(6): 1175–1184.
CHEN Xiaoqi, XIE Zhenping, LIU Yuan. News event detection driven by incremental sampling clustering[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(6): 1175–1184.

在线阅读 View online: https://dx.doi.org/10.11992/tis.201912037

DOI: 10.11992/tis.201912037

增量采样聚类驱动的新闻事件发现

陈晓琪[1,2], 谢振平[1,2], 刘渊[1,2]

(1. 江南大学 人工智能与计算机学院, 江苏 无锡 214122; 2. 江南大学 江苏省媒体设计与软件技术重点实验室, 江苏 无锡 214122)

摘 要: 为获得更好的事件发现和代表性新闻抽取性能, 引入数据集代表点采样聚类的视角, 研究实现了一种事件发现及表示的集成分析方法。对于给定的新闻流数据, 首先引入信息支撑度定义新闻间关系权重和事件关系权重, 并通过引入双层近邻传播算法的迭代构建整体时间流上的单向事件内容支撑度网络, 实现代表性新闻的分层增量采样, 进一步考虑以最大相似度划分策略实现代表性新闻上的整体新闻流数据聚类。实验结果表明, 相比于现有相关方法, 新方法在大规模新闻流数据上具有显著的计算效率, 可提取出新闻流中极有代表性的新闻, 以及获得更好的新闻文档聚类质量, 其热点事件发现结果与权威机构评选的重大新闻有极高吻合度。

关键词: 新闻流数据; 事件发现; 代表性新闻; 增量采样; 信息支撑度; 近邻传播; 事件网络; 分层聚类

中图分类号: TP391 **文献标志码:** A **文章编号:** 1673−4785(2020)06−1175−10

中文引用格式: 陈晓琪, 谢振平, 刘渊. 增量采样聚类驱动的新闻事件发现 [J]. 智能系统学报, 2020, 15(6): 1175–1184.

英文引用格式: CHEN Xiaoqi, XIE Zhenping, LIU Yuan. News event detection driven by incremental sampling clustering[J]. CAAI transactions on intelligent systems, 2020, 15(6): 1175–1184.

News event detection driven by incremental sampling clustering

CHEN Xiaoqi[1,2], XIE Zhenping[1,2], LIU Yuan[1,2]

(1. School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China; 2. Jiangsu Key Laboratory of Media Design and Software Technology, Jiangnan University, Wuxi 214122, China)

Abstract: For obtaining better performance of event detection and representative news extraction, an integrated analysis method of event detection and representation is proposed by introducing the sampling clustering strategy on news documents. For a given news flow data, first, we present two-weight definitions on the relationships between news and events by introducing an information supporting degree concept and then construct a one-way event content support network on the whole time flow using the iterative algorithm of double-layer nearest affinity propagation to realize layer-by-layer incremental sampling of representative news. Furthermore, overall news clustering was performed by using the maximum similarity division strategy. According to our experimental results, compared with existing related methods, the new method has significant computational efficiency for processing large-scale news flow data. It can extract the most representative news from the news flow and obtain better clustering quality of news documents. Its hot event detection results are highly consistent with the major news selected by the authority.

Keywords: news flow data; event detection; representative news; incremental sampling; information supporting degree; affinity propagation; event network; hierarchical clustering

随着互联网的不断发展, 网络媒体蓬勃兴起并成为新闻传播的重要途径之一。近几年, 互联网新闻用户规模日益庞大, 用户需求的多样性和网络信息传播的快捷性使得网络新闻数量呈井喷式增长。而另一方面, 大量新闻数据并没有能提高用户获取有效信息的效率, 且会因为相似新闻过多而增加用户不必要的阅读时间。因此, 抽取

收稿日期: 2019−12−31.

基金项目: 国家自然科学基金项目 (61872166); 江苏省“六大人才高峰”项目 (2019XYDXX-161).

通信作者: 谢振平. E-mail: xiezp@jiangnan.edu.cn.

大量新闻文档中的重要新闻，以及组织新闻流为新闻事件正成为信息推送服务的关键技术需求。

事件发现技术是信息推送服务的关键技术之一。目前，有关事件发现的大多数方法基于聚类思想实现，包括 single-pass 算法[1-2]、k-means 算法[3-4]、AP 算法[5-6] 等。其中主流的一种方法是 Allan 等[7] 提出的在线事件发现系统，每当有新的文档到来时，需与已知的事件计算相关度，通过预先设定的事件相似度阈值判断将该文档嵌入已知事件或生成新的事件。以该方法为基础，研究人员做出了许多改进工作，主要涉及文本表示形式的改进、更多语料特征的利用和文本聚类方法的建立。针对现有话题演化挖掘缺乏对话题特征随时间发展而动态演变的深入分析所导致的挖掘结果偏斜问题，赵旭剑等[8] 引入话题特征演变特性提出一种新的特征计算模型，利用已有话题文档和最新文档进行话题信息动态扩增，有效修正话题演化挖掘结果的偏斜。Yin 等[9] 针对短文本聚类的狄利克雷多项式混合模型，提出了一种折叠吉布斯抽样算法 GSDMM，可自动推断聚类数量，结果完备性、同质性平衡良好，能够解决短文本的稀疏高维度问题。周楠等[10] 基于带背景、语言的概率潜在语义分析模型 (PLSA with Background Language, PLSA-BLM)，结合关键词聚类发现事件内部子话题，在维基百科等知识库基础上生成事件子话题标签的模型 ET-TAG，与已有子话题发现算法相比有更好的性能。Xu 等[11] 使用唐森−香农散度 (Jensen-Shannon Divergence) 来度量话题相似度，引入时间衰减函数来提高相似时间的话题之间的相似度，改进 single-pass 算法并结合潜在狄利克雷分布 (latent dirichlet allocation, LDA) 达到有效监测和跟踪话题的目的。

事件发现的另一个趋势是增量或分层地处理文档。黄九鸣等[12] 为解决在线社交网络文本流所含热点短语指向的突发事件和热点话题问题，提出结合 AC-Trie 前缀树构建的无需分词且适用多种热度度量函数的热点短语挖掘技术。Chen 等[13] 考虑解决在线主题模型固定话题数、话题重叠问题，因为层次主题模型处理话题重叠的适配性，提出了基于知识的半监督层次在线话题检测框架。此外，一些其他方法也被应用于事件发现。Sayyadi 等[14] 基于复杂网络思想，提出了基于关键词共现性的 KeyGraph 话题检测方法。Chen 等[15] 则将层次隐树分析 (hierarchical latent tree analysis, HLTA) 引入话题检测，改进期望最大化 (expectation-maximization) 算法，可以得到更好的主题层次结构。柏文言等[16] 则开发一种融合用户关系的话题追踪方法，实现微博等短文本的事件跟踪。

当前有关信息推送服务的研究大多集中在文本聚类的事件发现上，关于事件表示的代表性新闻的提取研究还较少，文献 [17] 的做法是在事件发现的基础上通过计算得到与聚类中心最相近的文档作为代表性新闻。此外，许多现有方法要求保留所有的已处理文档[7,14] 作为历史信息与新到文档进行比较，随着数据规模的扩大以及数据流的不断到来，计算量和所需存储空间也会逐渐增大。文本考虑事件发现和事件表示的集成分析，兼顾大规模新闻流数据事件发现的可行性，提出一种以代表点选取为核心的增量采样聚类驱动的事件发现方法，该方法一方面在数据的增量处理中不断约简非代表性新闻以保证事件发现的效率，另一方面以采样获得的代表性新闻为基础实现聚类划分完成新闻事件发现，提供了一种可参考的信息推送技术新思路。

1　方法框架

为获得更好的新闻事件发现和代表性新闻抽取结果，引入分层增量思想和信息支撑度的信息约简策略，提出基于分层增量代表点采样的事件发现及表示的集成分析方法，其基本框架如图 1 所示。

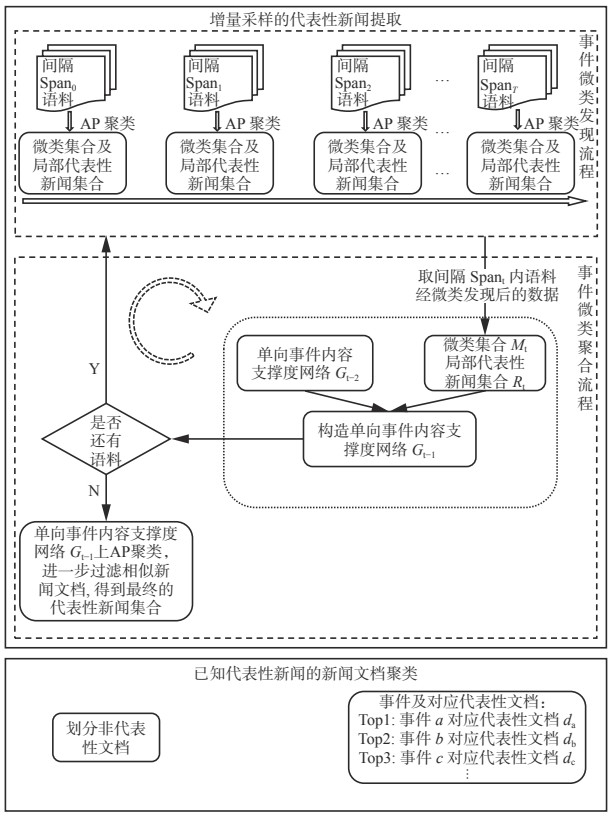


图 1　方法框架

Fig. 1　Framework of the proposed method

方法框架包含两个方面的处理:

1) 增量采样的代表性新闻提取。本文提出一种分层增量的代表性新闻采样方法, 如图 1 所示, 将新闻依据时间进行划分, 对每个时间片内的新闻进行 AP 聚类获得时间片上的事件微类及局部代表性新闻, 在此过程中增量的依据信息支撑度关系权重计算方法构建单向事件内容支撑度网络, 约简非局部代表性新闻。通过在单向事件内容支撑度网络上的 AP 聚类合并同一事件, 提取出最终的代表性新闻。

2) 已知代表性新闻的新闻文档聚类。在分层增量采样获取的代表性新闻基础上, 依照最大相似度划分策略, 将非代表性新闻分配给一个代表性新闻, 完成新闻文档聚类。

2　增量采样的代表性新闻提取

如图 1 所示, 增量采样的代表性新闻提取包含事件微类发现阶段和事件微类聚合阶段, 事件微类发现的目的在于通过聚类算法找到一个时间段内的局部代表性新闻, 事件微类聚合的目的在于将各时间段推选的局部代表性新闻进行合并, 从而推选出整个数据集上的代表性新闻。

2.1　事件微类发现

事件微类发现的目的是将一段时间内发表的新闻文档按照相关性组织为与事件相关的文档集合, 每一个文档集合称为该时间片内的一个事件微类, 并给出如下事件微类代表性新闻概念。

定义 1　在一个事件微类中, 其文档间内容高度相似, 从事件微类中选择一篇文档作为该事件微类的代表性新闻, 代表性新闻在内容上与事件微类中的其他新闻文档高度相似, 是事件微类的中心。

为了更方便及准确地找到事件微类的聚类中心, 采用 AP 聚类[18] 完成事件微类发现的工作。AP 算法通过节点间吸引度消息和归属度消息的传递收敛将相似文档划分到同一类别, 其对数据的初始值不敏感, 且 AP 算法能选择真实存在的数据点作为聚类中心, 从而在寻找事件微类的代表性新闻时不再需要额外处理手段, 对应的文档可以直接作为事件微类的代表性新闻文档。对应 AP 算法输入的相似度矩阵的要求, 此阶段采用余弦相似度表征新闻文档间的相似性。

2.2　事件微类聚合

事件微类的聚合主要包含 3 个步骤: 1) 定义单向文档关系网络, 以及定义引入建构主义信息支撑度概念的单向文档内容支撑度网络; 2) 在单向文档内容支撑度网络的基础上对单向事件内容支撑度网络进行增量构建; 3) 单向事件内容支撑度网络上的相似事件聚合。

2.2.1　单向文档内容支撑度网络

一般来讲, 热点事件的热度高, 持续时间长, 因此在相邻的时间片上可能存在内容非常相似的两个事件微类, 但随着时间的增长, 相隔越远的两个报道越不可能同属于一个事件。为此, 引入单向文档关系网络去解释这一现象。单向文档关系网络将新闻流中的每一篇新闻文档看作网络中的节点, 节点间的边权重由文档间余弦相似度确定, 边方向由发表时间上的前序文档指向后序文档, 且规定只有相邻时间片上的节点之间存在联系, 同一时间片以及不相邻时间片上的节点之间不直接相联系, 相应的网络结构如图 2 所示。

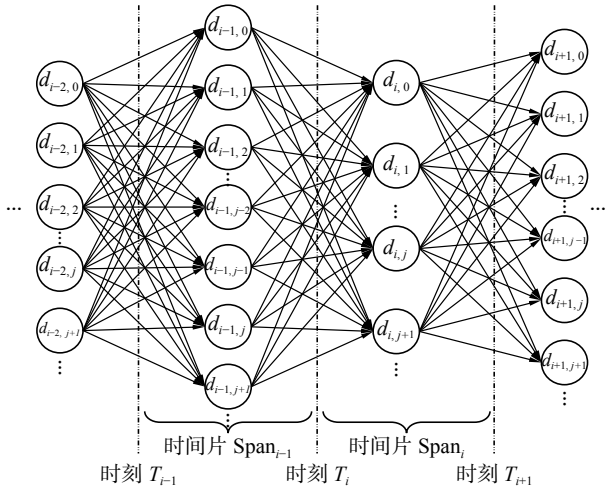


图 2　单向文档关系网络

Fig. 2　One-way document network

进一步地, 在单向文档关系网络上引入信息支撑度概念, 信息支撑度源于建构主义学习理论中的知识支撑概念[19], 建构主义认为知识是在其他基础知识上通过它们的相互作用而主动构建产生的。迁移到文档产生上, 一篇新文档的产生通常需要借鉴和引用已有文档内容, 新文档会从旧文档中汲取知识完成自身的知识构建, 基于此引入关于信息支撑度的概念定义。

定义 2　在文档产生中时间上的后序文档会对前序文档有一定的借鉴和引用, 后序文档中内容的构建会一定程度上参考前序文档。信息支撑度描述了这一知识传递关系, 表示有直接关联的前序文档对后序文档在内容上的直接支持程度。信息支撑度是一个单向的指向关系, 由前序文档指向后序文档, 表示由前序文档将知识传递给后序文档。

考虑新文档更可能借鉴和引用时间上更近的文档，规定两个相邻时间片上的文档间有直接关联性。这样可考虑所有前一时间片内的文档构成了对下一时间片某一文档 $\boldsymbol{d}$ 的内容支撑集合，内容支撑集合对某一文档 $\boldsymbol{d}$ 的支撑度总和设为 1。文档间的支撑度定义为

$$\begin{cases} \sup_{j,a} = \mathrm{sim}(\boldsymbol{d}_j, \boldsymbol{d}_a) \times \dfrac{1}{U} \\ \boldsymbol{d}_j \in \mathrm{Span}_{i-1}, \quad \boldsymbol{d}_a \in \mathrm{Span}_i \\ U = \displaystyle\sum_{\boldsymbol{d}_l \in \mathrm{Span}_{i-1}} \mathrm{sim}(\boldsymbol{d}_l, \boldsymbol{d}_a) \end{cases} \tag{1}$$

式中：$\sup_{j,a}$ 表示直接关联的属于时间片集合 Span_{i-1} 的文档 $\boldsymbol{d}_j$ 对属于时间片集合 Span_i 的文档 $\boldsymbol{d}_a$ 的支撑度；$\mathrm{sim}(\boldsymbol{d}_j, \boldsymbol{d}_a)$ 为两文档的余弦相似度；U 为 Span_{i-1} 内所有文档与文档 d_a 的余弦相似度总和。以信息支撑度为关系权重的单向文档关系网络称为单向文档内容支撑度网络。

2.2.2　单向事件内容支撑度网络

为了在获取代表性新闻文档的过程中不断约简掉非代表性新闻文档以减少计算量和所需存储空间，考虑用事件微类发现阶段产生的局部代表性新闻去代表微类。相应地，需要将单向文档内容支撑度网络转化为单向事件内容支撑度网络。同单向文档内容支撑度网络的特征相似，单向事件内容支撑度网络中只有相邻时间片上的局部代表性新闻间存在联系，同一时间片以及不相邻时间片上的局部代表性新闻间不直接相联系。为此，引入复合支撑度作为单向事件内容支撑度网络的关系权重。

定义 3　单向事件内容支撑度网络构建中，两个相邻时间片上的任意两个局部代表性新闻 $R_{i,a}$、$R_{i+1,b}$ 间的复合支撑度 $\mathrm{csup}_{R_{i,a},R_{i+1,b}}$ 定义为

$$\mathrm{csup}_{R_{i,a},R_{i+1,b}} = \exp\left(-\frac{\min(|M_{i,a}|,|M_{i+1,b}|)}{\displaystyle\sum_{d_{i,x}\in M_{i,a}}\sum_{d_{i+1,y}\in M_{i+1,b}} \sup_{x,y}}\right) \tag{2}$$

式中：$M_{i,a}$ 为区间 $[T_i, T_{i+1})$ 的微类；$R_{i,a}$ 为 $M_{i,a}$ 产生的局部代表性新闻；$M_{i+1,b}$ 为区间 $[T_{i+1}, T_{i+2})$ 的微类；$R_{i+1,b}$ 为 $M_{i+1,b}$ 产生的局部代表性新闻；$d_{i,x}$ 为微类 $M_{i,a}$ 中的文档；$d_{i+1,y}$ 为微类 $M_{i+1,b}$ 中的文档，$\sup_{x,y}$ 意义同式 (1) 中 $\sup_{j,a}$ 的含义。

单向事件内容支撑度网络是增量构建的，如图 1 所示，其构建过程与事件微类发现流程同步。在图 3 中，当 Span_i 时间片上的事件微类发现完成后，与 Span_{i-1} 时间片上的事件微类做复合支撑度的计算，将 Span_i 上的微类代表性新闻文档作为节点加入到 $\{\mathrm{Span}_0, \mathrm{Span}_1, \cdots, \mathrm{Span}_{i-1}\}$ 上微类代表性新闻文档构成的单向事件内容支撑度网络 supNet_{i-2} 上，构建单向事件内容支撑度网络 supNet_{i-1}。在这一构建过程中，事件微类发现产生的非局部代表性新闻在完成复合支撑度的计算后被约简，其包含的特征被整合为复合支撑度赋予局部代表性新闻，本身不需要参与第二层聚类运算。当新闻流中的所有文档均处理完成后，形成所有局部代表性新闻构成的单向事件内容支撑度网络，完成对所有非局部代表性新闻文档的约简。

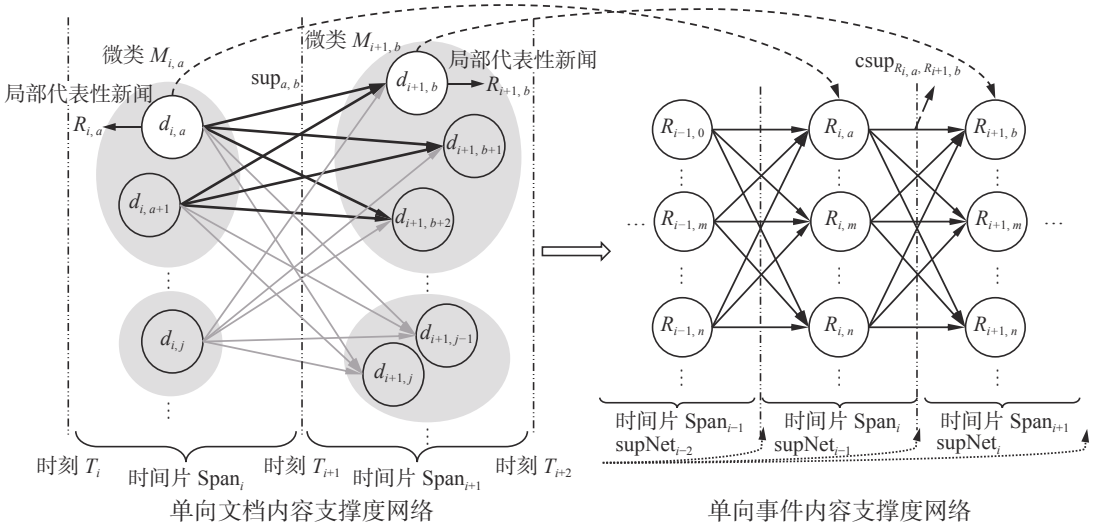


图 3　单向文档内容支撑度网络到单向事件内容支撑网络的转化

Fig. 3　Transformation from one-way document content support network to one-way event content support network

2.2.3　相似事件聚合

考虑到单向网络的不对称性和代表点获取的需求，同样采用 AP 算法进行事件微类的聚合。给定 $M_{t-2} \times M_{t-2}$ 维相似度矩阵 $\boldsymbol{S}_{t-2}$，$M_{t-1} \times M_{t-1}$ 维相

似度矩阵 $\boldsymbol{S}_{t-1}$, 依据上述单向事件内容支撑度网络构建过程, 相应的 t 时刻网络对应相似度矩阵 $\boldsymbol{S}_t$ 定义为

$$\boldsymbol{S}_t(i,j)=\begin{cases}\boldsymbol{S}_{t-1}(i,j), & i\leqslant M_{t-1}, j\leqslant M_{t-1}\\ \text{csup}_{i,j}, & M_{t-2}<i\leqslant M_{t-1}, j>M_{t-1}\\ 0, & i>M_{t-1}\text{或者}i\leqslant M_{t-2}, j>M_{t-1}\end{cases}\tag{3}$$

基于 AP 算法将相似局部代表性新闻文档划分为一类, 即微类聚合为大的事件类, 同样依照 AP 聚类结果选择聚类中心作为最终的代表性新闻, 只有局部代表性新闻才有可能被选为最终的代表性新闻。此外, 在事件微类发现阶段, 期望得到更为广泛的事件微类, 有助于事件微类聚合过程中得到代表性更强、重要性更高的新闻文档, 因此在事件微类发现阶段, 使用文档间相似度的中位值作为 AP 算法的偏向参数, 但在事件微类聚合阶段需要设置合适的偏向参数。

算法 1　增量采样的代表性新闻提取

输入　新闻子集流 $\{D_1,D_2,\cdots,D_n\}$, 偏向参数 preference;

输出　代表性新闻集 globalR。

init lastR ← Ø, lastCluster ← Ø, $\boldsymbol{G}$ ← []
for i ← 1 to n do
if i = 1
根据D_1获取相似度矩阵 $\boldsymbol{S}$
lastR, lastCluster ← AP($\boldsymbol{S}$, $\boldsymbol{S}_{\text{median}}$)
扩展$\boldsymbol{G}_{0\times0}$ 到 $\boldsymbol{G}_{|\text{lastR}|\times|\text{lastR}|}$
else
根据D_i获取相似度矩阵 $\boldsymbol{S}$
localR, localCluster ← AP($\boldsymbol{S}$, preference)
扩展$\boldsymbol{G}_{|G|\times|G|}$ 到 $\boldsymbol{G}_{(|\text{localR}|+|G|)\times(|\text{localR}|+|G|)}$
for x ← 1 to $||\text{localR}|+|G||$ do
for y ← 1 to $||\text{localR}|+|G||$ do
依据式 (3) 计算$\boldsymbol{G}[x][y]$
(base lastR, lastCluster, localR, localCluster)
end for
end for
lastR ← localR, lastCluster ← localCluster
end if
end for
globalR ← AP($\boldsymbol{G}$, preference)
return globalR

2.3　已知代表性新闻的新闻文档聚类

通过增量分层采样得到代表性新闻文档后, 依据最大相似度策略将非代表性新闻文档分配给一个代表性新闻文档所代表的类, 完成所有文档的全局划分, 划分的结果即为事件发现的结果。

算法 2　已知代表性新闻的新闻文档聚类

输入　新闻子集流 $\{D_1,D_2,\cdots,D_n\}$, 代表性新闻集 globalR;

输出　聚类结果 $\left\{C_1,C_2,\cdots,C_{|\text{globalR}|}\right\}$。

$D \leftarrow D_1\cup D_2\cup\cdots\cup D_n$
for i ← 1 to $\left|\text{globalR}\right|$ do
delete globalR[i] from D
create C_i ← Ø
end for
for j ← 1 to $|D|$ do
for i ← 1 to $\left|\text{globalR}\right|$ do
calculate sim($D[j]$, globalR[i]))
end for
label ← $i|\max(\text{sim}(D[j],\text{globalR}[i]))$
add $D[j]$to C_{label}
end for
rerutn $\{C_1,C_2,\cdots,C_{|\text{globalR}|}\}$

假设新闻流文档总数为 N, 时间片数为 T, 微类发现阶段的局部代表性新闻提取率为 p, 在文档规模为 N 的情况下, 标准 AP 算法复杂度[18] 为 $O(N^2\lg N)$。本文方法复杂度主要考虑 AP 算法部分和复合支撑度部分, 微类发现阶段 AP 算法时间复杂度为 $O(N^2/T^2\lg(N/T))$, 微类聚合阶段 AP 算法时间复杂度为 $O(N^2p^2\lg Np)(0<p\ll1)$, 复合支撑度计算的事件复杂度为 $O(N^2/T^2)$。考虑 N/T 的规模较小, 而 p 正常情况是一个远小于 1 的数, 相比于直接使用 AP 算法, 文中方法的时间复杂度可以远远降低。

3　实验研究

3.1　实验方案

以新闻文档为研究对象, 为更好地检验算法性能, 实验中采用了以下几种事件发现方法作为研究:

1) k-means++方法[20]。使用标准 k-means 方法对所有文档统一进行聚类, 初始聚类中心依照相互较远的准则进行选择。划分的每一个类作为一个事件, 选择离聚类中心最近的文档作为代表性新闻。

2) 标准 AP 方法[18]。使用标准 AP 方法对所有文档统一进行聚类, 选择聚类中心对应的文档作为代表性新闻。

3) 改进 single-pass 方法[17]。使用时间片分层 single-pass 在线聚类方法增量地对文档进行处理, 不同时间片间事件依据话题的相似度进行合并,

选择离聚类中心最近的文档作为代表性新闻。

4) IAPNA 方法[21]。使用增量式扩充吸引度和归属度矩阵的增量 AP 方法对文档进行处理，选择聚类中心对应的文档作为代表性新闻。

实验中采用向量空间模型[22]对文档进行表示，TF-IDF 方法计算词项权重，为消除不同文档长度对计算的影响，需要同时对权重做归一化处理[23]。在对比方法中均使用余弦相似度表征文档间关系。

文中实验所用平台配置为：Intel(R) Core (TM) i7-6700k CPU @4.00 GHz、内存为 16 GB、操作系统为 Windows10、所有代码基于 Python 语言实现。

3.2　评价指标

考虑主要从聚类质量和选取新闻的代表性程度两个方面评价算法。聚类评价标准采用标准化互信息 (normalized mutual information, NMI) 以及 F_1 值。标准化互信息是被广泛使用的评价聚类效果的方法，可以度量方法输出结果和真实结果之间的相似程度，NMI 的取值范围为 [0, 1]，越靠近 1 表示方法输出结果与真实结果越相似。本文采用文献 [24] 中 NMI 的计算方式。

F_1 值是分类问题的一个衡量指标，也常用于文本聚类评价，是一种同时考虑准确率 (precision) 和召回率 (recall) 的综合性评价方法，是两者的调和平均数，本文采用文献 [17] 中准确率、召回率、F_1 值的计算方式。

针对新闻代表性程度，新定义新闻文档接近度 (document proximity, DP) 指标来评价提取的代表性新闻对非代表性新闻的内容覆盖程度，其定义为

$$\mathrm{DP}=\exp\left(-\frac{N-n_r}{n_r}\right)\cdot\frac{\sum_{j=1}^{N-n_r}\min(\mathrm{dist}_{j,1},\mathrm{dist}_{j,2}\cdots,\mathrm{dist}_{j,n_r})}{N-n_r}\tag{4}$$

式中：N 是文档总数；n_r 是提取的代表性新闻数；dist 表示两文档间归一化的欧式距离。新闻文档接近度考虑在合理的代表性新闻文档数的情况下，代表性新闻在内容上对整体新闻流的覆盖程度，DP 值越小，说明提取的代表性新闻在内容上的覆盖程度越高。

3.3　实验结果与分析

通过爬取人民日报 2019 年 6 月初到 7 月中旬的时政报道，筛选出 530 篇长文档，按照报道发表的时间排序，精确到天。对该新闻数据集人工标记了 43 个事件话题，其数量分布如图 4 所示。

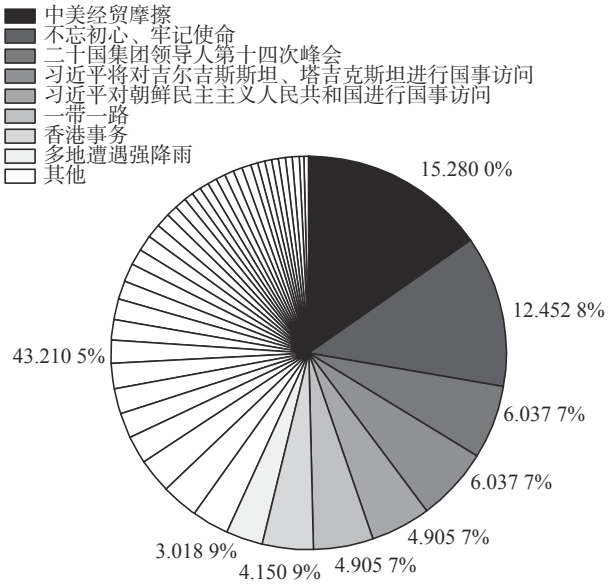


图 4　人民日报数据集话题分布

Fig. 4　Topic distribution of People's Daily Online dataset

采用北大开源词法分析工具 pkuseg[25]对新闻文档进行分词、去停用词以及 TF-IDF 权重计算。因为人工标记的话题数为 43，实验过程中设定 k-means++的类数为 43，标准 AP 方法偏向参数取默认中位值。改进 single-pass 方法相似度阈值为 0.05，话题合并阈值为 0.15。IAPNA 方法初始偏向参数为 0.0018。

首先分析 AP 算法的偏向参数 (preference) 对文中方法的性能影响。以 10 天为时间间隔，阻尼系数设定为 0.9，在 [0.01,0.1] 的范围内，以 0.001 为步长，得到不同偏向参数下的 F_1 值和 NMI 值，如图 5 所示。

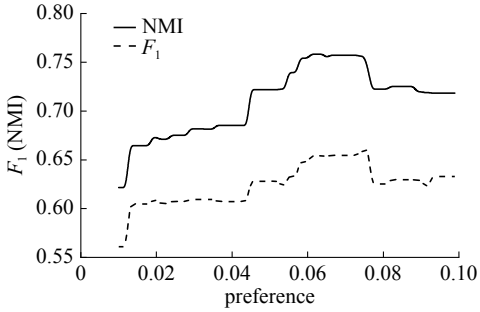


图 5　不同 preference 参数对性能结果的影响

Fig. 5　Performance results on with different preference parameters

从图 5 中可以看到，开始时随着偏向参数的增加，F_1 和 NMI 值均不断增加，此时偏向参数过小生成的类数较少，许多不属于同一个事件的文档被划分到了同一个类中，聚类的准确率较低，因而一开始 F_1 值与 NMI 值较小。偏向参数增加后，不同的文档开始被分到不同的类中，聚类的准确率也开始增加，F_1 值与 NMI 值也随之增大。但当偏

向参数增加到一定的值时, 因为所有数据被选为聚类中心的可能性增大了, 所以有许多同一类的文档被分到不同的类中, 聚类的召回率降低, 此时 F_1 值和 NMI 值开始减小。在本实验中, 偏向参数为 0.076 时, 结果的 F_1 值和 NMI 值达到最大。

本文方法与 k-means++、标准 AP、改进 single-pass、IAPNA 方法对比实验结果如表 1 所示, 可以看出, 本文方法在聚类质量上均优于其他 4 种方法, 选取文档的重要性程度方面则位于中间位置。本文方法在 F_1 值和 NMI 值均高于其他 4 种方法的同时 DP 值较小, 保证了较好的新闻内容覆盖度。

表 1　5 种对比方法的实验性能结果

Table 1　Performance comparison of the five methods

方法	F_1	NMI	DP/10^{-4}
本文方法	**0.6583**	**0.7561**	0.349
k-means++	0.4832	0.7047	0.781
标准AP	0.5609	0.7218	0.145
改进single-pass	0.6505	0.7166	0.404
IAPNA	0.5843	0.7457	**0.135**

本文方法、k-means++、标准 AP、改进 single-pass、IAPNA 的准确率、召回率、F_1 值、NMI 值和 DP 值对比结果如图 6 所示。5 种方法的准确率较为相近, 本文方法和改进 single-pass 方法的召回率要好于其他方法, 综合准确率和召回率, 本文方法具有最高的 F_1 值。此外, 本文方法的 NMI 值也是最高的, NMI 值较高的还有 IAPNA 方法。而从代表性新闻的内容覆盖度来看, 本文方法的 DP 值处于中间位置。AP 及基于 AP 的方法在 DP 值上有较好表现, 标准 AP 方法计算与存储均不增量, IAPNA 则是计算增量、存储不增量, 因此在代表点选择上效果较好, 本文方法计算与存储均是增量的, 不可避免地有信息缺失的影响, 但 AP 方法的良好性能使得本文方法所得 DP 值小于 k-means++和改进 single-pass 方法。这也从另一个角度反映了本文所提 DP 性能指标的合理性, 可为相关研究提供一个十分有价值的参考。

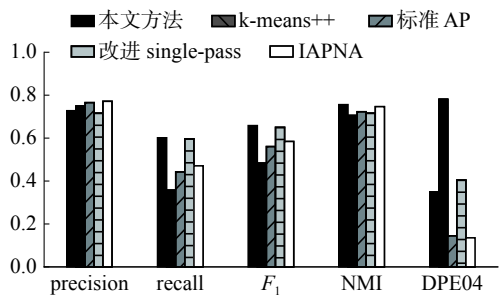


图 6　5 种方法的详细性能对比

Fig. 6　Detailed performance comparison of the five methods

5 种方法的运行时间结果如表 2 所示, 为了进一步分析数据集规模对不同算法运行时间的影响, 除了在人民日报数据集上做运行时间的比较之外, 还选取中国新闻网 2018 年 1 月国内时政新闻报道, 共 2 747 篇新闻作为实验数据集。如表 2 结果所示, 随着数据规模的增大, 本文方法的运行时间增长倍数最低, 尤其是相比于其他两种在线方法, 本文方法有着更好的运行效率。

表 2　5 种对比方法的运行时间结果

Table 2　Comparison of the running times of the five methods

方法	T_1(530)/s	T_2(2 747)/s	T_2/T_1
本文方法	4.59	65.75	14.32
k-means++	0.76	15.13	19.91
标准AP	1.43	38.88	27.19
改进single-pass	27.49	641.87	23.35
IAPNA	4.56	275.80	60.48

3.4　事件发现应用实验

考虑将本文方法应用于年度新闻热点事件发现中, 与权威机构发布的 2018 年新闻评选结果[26] 进行比较, 权威机构评选结果如表 3 所示。

表 3　人民网评选出的 2018 十大时政新闻

Table 3　Top ten political news published by People's Daily Online in 2018

编号	人民网评选出的2018十大时政新闻
1	《中华人民共和国宪法修正案》通过
2	庆祝改革开放40周年大会举行
3	党和国家机构改革顺利推进
4	首届中国国际进口博览会举行
5	民营企业座谈会召开
6	全国生态环境保护大会召开
7	海南全境建设自贸区并探索实行自由贸易港政策
8	港珠澳大桥正式通车运营
9	嫦娥四号探测器成功发射
10	个税起征点上调

爬取中国新闻网 2018 年国内时政新闻报道, 删除篇幅过短的文章, 保留共 40 106 篇 2018 年时政新闻报道。以 10 天为时间片划分数据集, 用本文方法获得全年划分的事件类, 经过规模排名得到排名最靠前的 10 个类, 从每个类中选取本文方法所得代表性新闻以及最靠近代表性新闻的 30 篇新闻作为该类事件的推荐新闻, 对新闻文档进行关键词提取, 本实验对

2018 年新闻数据集用本文方法进行 5 次热点事件发现工作，最终展示结果为 5 次实验结果的并集。本文方法得到的新闻事件发现结果如表 4 所示。

表 4　本文方法提取的代表性新闻文档对应关键词及其事件

Table 4　Representative documents and their corresponding events extracted by our method

本文方法得到的热点事件类的重要文档关键词	对应事件
改革、建设、机构、社会主义、贯彻、执行…	党和国家机构改革
脱贫、攻坚、扶贫、精准、发展、贫困人口…	脱贫攻坚精准扶贫
代表、委员、宪法、推进、修正案、使命…	《中华人民共和国宪法修正案》
污染、生态、督查、整改、治理、环保部门…	全国生态环境保护大会
国际、贸易、金融、海南、发展、试验区…	海南全境建设自贸区并探索实行自由贸易港政策
上合、组织、峰会、会见、青岛、国际…	上合组织青岛峰会举行
媒体、高考、服务、组织、新闻、情况…	高考
国家、合作、全面、阿联酋、中阿、金砖…	习近平主席开启非洲之旅
医疗、违法、长生、发布、审批、涉案…	吉林长春长生公司问题疫苗
遭遇、游船、倾覆、救援、调查、旅游…	游船在泰国普吉岛附近海域倾覆
个人、退休、上调、文件、加减、改革…	养老金上调
合作、中非、论坛、会见、峰会、北京…	中非合作论坛北京峰会
大桥、建设、仪式、港澳、启动、活动…	港珠澳大桥正式通车运营
国际、博览会、中国、推动、世界、开放…	首届中国国际进口博览会
经济、问题、民营企业、对话、政策、探索…	民营企业座谈会
中国、改革、庆祝、成立、时代、振兴…	庆祝改革开放40周年大会

通过与权威机构发布的 2018 年十大时政新闻评选结果进行比较可以发现，本文方法提取的 16 个热点事件中有 8 个相似事件包含在 2018 年十大时政新闻评选榜中，分别为党和国家机构改革、《中华人民共和国宪法修正案》、全国生态环境保护大会、海南全境建设自贸区并探索实行自由贸易港政策、港珠澳大桥正式通车运营、首届中国国际进口博览会、民营企业座谈会以及庆祝改革开放 40 周年大会，对应于表 3 序号 1~8。而提取结果中的另外 8 个事件，除脱贫攻坚精准扶贫外，均被人民日报盘点为 2018 年每月大事件[27]。

作为对比，对 2018 年新闻数据集用 k-means++方法进行事件发现工作，结果如表 5 所示。表 5 中序号 1~3 对应的事件为党和国家机构改革、2 位院士获国家最高科技奖、天安门升旗仪式，序号 6~7 对应事件为脱贫攻坚精准扶贫、全国生态环境保护大会，序号 10 对应事件为一带一路，序号 12 对应事件为港珠澳大桥正式通车运营。序号 4、5、8、9、11、13、14、15 对应于科学、文化、教育、军事、监察等领域，但没有特殊的关键词能推测出具体的事件。此外，k-means++方法提取的热点事件中只有 3 个相似事件包含在 2018 年十大时政新闻评选榜中。

表 5　k-means++提取出的代表性新闻文档对应关键词

Table 5　Key words of representative documents extracted by k-means++

编号	k-means++得到的热点事件类的重要文档关键词
1	代表、改革、推荐、机构、中国、组织…
2	研究、院士、科学家、基因、病毒、科研…
3	天安门广场、仪式、北京、升国旗、国旗…
4	中国、科学家、治疗、研究、量子、突破…
5	代表、委员、文化、政策、呼吁、传承…
6	问题、扶贫、政府、农村、整改、预算…
7	水质、保护、环境、环境监测、治理、管理…
8	高校、中国、人才、大学生、大赛、创新…
9	回应、中国、国际、南海、关系、制裁…
10	中国、合作、一带一路、建设、国家…
11	案件、监察、监察机关、检方、公安部、专项…
12	大桥、完成、工程、港澳、开通、世界…
13	精神、工作、红色、时代、举办、发展…
14	科学家、突破、太空、载人、执行、黑洞…
15	专项、执法、整治、违法、生产、违规…

相比之下, 本文方法得到的代表性新闻粒度更细, 提取的关键词中能够包含热点事件的特殊名词, 事件发现质量更高。而 k-means++方法初始点的选择对结果影响较大, 虽然能在一定程度上划分事件的领域, 但是在提取细化的热点事件上结果较差。

4 结束语

为获得更好的事件发现和代表性新闻抽取性能, 从数据集代表点采样聚类的视角, 提出了增量采样聚类驱动的事件发现新方法。新方法通过引入两层聚类策略、单向文档内容支撑度网络和单向事件内容支撑度网络, 结合代表性新闻间基于内容支撑度的关系权重计算方法, 有效地实现了按照时间顺序组织新闻文档。其中基于信息支撑度的事件关系权重表示方法, 完成了从文档相关性网络到事件相关性网络的转换, 有效地约简了相似性高的新闻文档流, 减少了算法所需的存储空间和计算量。

实验中使用真实新闻数据集进行性能分析, 结果表明, 本文方法在文档聚类质量上优于 k-means++、标准 AP、改进 single-pass、IAPNA 方法, 代表性新闻提取质量上处于中间位置。相比于改进 single-pass、IAPNA 在线聚类方法, 本文方法有计算时间上的优势。在中国新闻网 2018 年全年国内时政新闻文档集上的事件发现实验结果表明, 新方法可获得非常接近于人工评选的结果。

参考文献:

[1] QU Xiaoting, YANG Juan, WU Bin, et al. A news event detection algorithm based on key elements recognition[C]// Proceedings of 2016 IEEE First International Conference on Data Science in Cyberspace. Changsha, China, 2016: 394–399.

[2] YAN Danfeng, HUA Enzheng, HU Bo. An improved single-pass algorithm for Chinese microblog topic detection and tracking[C]//Proceedings of 2016 IEEE International Congress on Big Data. San Francisco, USA, 2016: 251–258.

[3] 路荣, 项亮, 刘明荣, 等. 基于隐主题分析和文本聚类的微博客中新闻话题的发现 [J]. 模式识别与人工智能, 2012, 25(3): 382–387.
LU Rong, XIANG Liang, LIU Mingrong, et al. Discovering news topics from microblogs based on hidden topics analysis and text clustering[J]. Pattern recognition and artificial intelligence, 2012, 25(3): 382–387.

[4] GENG Xiao, ZHANG Yanmei, JIAO Yuhang, et al. A novel hybrid clustering algorithm for topic detection on Chinese microblogging[J]. IEEE transactions on computational social systems, 2019, 6(2): 289–300.

[5] GUAN Renchu, SHI Xiaohu, MARCHESE M, et al. Text clustering with seeds affinity propagation[J]. IEEE transactions on knowledge and data engineering, 2011, 23(4): 627–637.

[6] SHRIVASTAVA S K, RANA J L, JAIN R C. Text document clustering based on phrase similarity using affinity propagation[J]. International journal of computer applications, 2013, 61(18): 38–44.

[7] ALLAN J, PAPKA R, LAVRENKO V. On-line new event detection and tracking[C]//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Canberra, Australia, 1998: 37–45.

[8] 赵旭剑, 杨春明, 李波, 等. 一种基于特征演变的新闻话题演化挖掘方法 [J]. 计算机学报, 2014, 37(4): 819–832.
ZHAO Xujian, YANG Chunming, LI Bo, et al. A topic evolution mining algorithm of news text based on feature evolving[J]. Chinese journal of computers, 2014, 37(4): 819–832.

[9] YIN Jianhua, WANG Jianyong. A dirichlet multinomial mixture model-based approach for short text clustering[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2014: 233–242.

[10] 周楠, 杜攀, 靳小龙, 等. 面向舆情事件的子话题标签生成模型 ET-TAG[J]. 计算机学报, 2018, 41(7): 1490–1503.
ZHOU Nan, DU Pan, JIN Xiaolong, et al. ET-TAG: a tag generation model for the sub-topics of public opinion events[J]. Chinese journal of computers, 2018, 41(7): 1490–1503.

[11] XU Guixian, MENG Yueting, CHEN Zhan, et al. Research on topic detection and tracking for online news texts[J]. IEEE access, 2019, 7: 58407–58418.

[12] 黄九鸣, 吴泉源, 张圣栋, 等. 基于 AC-Trie 的在线社交网络文本流热点短语挖掘 [J]. 电子学报, 2016, 44(10): 2466–2470.
HUANG Jiuming, WU Quanyuan, ZHANG Shengdong, et al. Mining hot phrases on sociai network text streams based on AC-Trie[J]. Acta electronica sinica, 2016, 44(10): 2466–2470.

[13] CHEN Ling, TU Ding, LV Mingqi, et al. A knowledge-based semisupervised hierarchical online topic detection framework[J]. IEEE transactions on cybernetics, 2019, 49(9): 3307–3321.

[14] SAYYADI H, RASCHID L. A graph analytical approach for topic detection[J]. ACM transactions on internet technology, 2013, 13(2): 4.

[15] CHEN Peixian, ZHANG N L, LIU Tengfei, et al. Latent tree models for hierarchical topic detection[J]. Artificial intelligence, 2017, 250: 105–124.

[16] 柏文言, 张闯, 徐克付, 等. 一种融合用户关系的自适应微博话题跟踪方法 [J]. 电子学报, 2017, 45(6): 1375–1381.
BAI Wenyan, ZHANG Chuang, XU Kefu, et al. A self-adaptive microblog topic tracking method by user relationship[J]. Acta electronica sinica, 2017, 45(6): 1375–1381.

[17] 张斌, 胡琳梅, 侯磊, 等. 基于词向量的中文事件发现及表示 [J]. 模式识别与人工智能, 2018, 31(3): 275–282.
ZHANG Bin, HU Linmei, HOU Lei, et al. Word embedding based Chinese news event detection and representation[J]. Pattern recognition and artificial intelligence, 2018, 31(3): 275–282.

[18] FREY B J, DUECK D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972–976.

[19] 谢振平, 金晨, 刘渊. 基于建构主义学习理论的个性化知识推荐模型 [J]. 计算机研究与发展, 2018, 55(1): 125–138.
XIE Zhenping, JIN Chen, LIU Yuan. Personalized knowledge recommendation model based on constructivist learning theory[J]. Journal of computer research and development, 2018, 55(1): 125–138.

[20] ARTHUR D, VASSILVITSKⅡ S. K-Means++: the advantages of careful seeding[C]//Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. New Orleans, USA, 2007: 1027−1035.

[21] SUN Leilei, GUO Chonghui. Incremental affinity propagation clustering based on message passing[J]. IEEE transactions on knowledge and data engineering, 2014, 26(11): 2731–2744.

[22] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613–620.

[23] DAI Xiangying, HE Yancheng, SUN Yunlian. A two-layer text clustering approach for retrospective news event detection[C]//Proceedings of 2010 International Conference on Artificial Intelligence and Computational Intelligence. Sanya, China, 2010: 364−368.

[24] AMELIO A, PIZZUTI C. Is normalized mutual information a fair measure for comparing community detection methods?[C]//Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. Paris, France, 2015: 1584−1585.

[25] LUO Ruixuan, XU Jingjing, ZHANG Yi, et al. PKUSEG: a toolkit for multi-domain Chinese word segmentation [EB/OL]. [2019−06−22] http://axvio.org/abs/1906.11455.

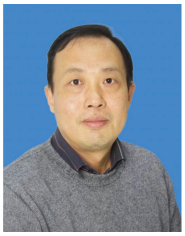
[26] 人民日报社评选. 2018 国内十大新闻 [N]. 人民日报, 2018−12−29(02).

[27] 人民日报. 9 张图速读 12 个月, 带你回顾即将过去的 2018[DB/OL]. 人民日报微博, [2018−12−26]. https://baijiahao.baidu.com/s?id=1620877305413536727.

作者简介：

陈晓琪，硕士研究生，主要研究方向为大数据知识发现。

谢振平，教授，博士生导师，主要研究方向为知识表示与认知学习。主持或参与完成国家、省部级科研项目 6 项，承担产学研合作项目 15 项。获发明专利 5 项，发表学术论文 30 余篇。



刘渊，教授，博士生导师，主要研究方向为网络安全、数字媒体。作为项目负责人完成了省部级科研项目 3 项。发表学术论文 40 余篇，出版专著 1 部。