一种基于TF·IEF模型的在线新闻事件探测方法*

张 辉,李国辉,贾 立,孙博良
(国防科技大学 信息系统与管理学院,湖南 长沙 410073)

摘 要:为了提升在线新闻事件探测的性能,提出一种基于TF·IEF模型的在线新闻事件探测方法。该方法受TF·IDF思想的启发,直接计算特征词表征事件的权重,建立新的增量事件模型,并将探测过程分为两个阶段:第一阶段利用Single-Pass将一定时段内收集到的报道聚成微簇;第二阶段将微簇与已有事件进行相似性匹配,然后通过重新计算事件向量实现模型更新。实验结果表明,该方法运算速度快,特征信息丢失少,提高了探测的效率和准确率。

关键词:在线新闻事件探测;TF·IEF模型;增量事件模型;Single-Pass聚类;
中图分类号:TP391 **文献标志码**:A **文章编号**:1001-2486(2013)03-0055-06

On-line news event detection based on TF · IEF model

ZHANG Hui, *LI Guohui*, *JIA Li*, *SUN Boliang*
(College of Information System and Management, National University of Defense Technology, Changsha 410073, China)

Abstract: According to the characters of web news stream, an on-line news event detection (ONED) method, based on the two-stage clustering, is proposed to solve the problem of repeated matching. A novel incremental event model was established by calculating terms weighting of events directly. Two stages are involved in our method. In the first stage, the similar reports collected in a certain period were clustered into micro-clusters. In the second, the micro-clusters were matched with existed events, and then this method updated the event model. Experiment shows that the proposed method improves the efficiency and accuracy of ONED with lower complexity and less feature information loss.

Key words: on-line news event detection; TF · IEF model; incremental event model; Single-Pass clustering

在线新闻事件探测(ONED)所要解决的问题就是如何从网络新闻报道流中自动发现最新发生的新闻事件。利用在线新闻事件探测技术,重要信息可以免于被大量的无序新闻所淹没,用户能够快捷地了解近期内发生的重大事件。

近年来,许多学者对新闻事件探测进行了研究。Allan[1]、Papka[2]、Yang[3]、Lam[4]等使用TF·IDF(Term Frequency & Inverse Documentation Frequency)对报道建模,利用Single-Pass聚类方法进行探测。Brants[5]等用改进的增量式TF·IDF方法建立报道的单向量模型,而Stocks[6]、Giridhar[7]、张阔[8-9]等则用TF·IDF建立报道的多向量模型,这些改进研究主要侧重在报道模型以及充分利用报道的语义特征两个方面。付艳[10-11]等提出一种基于命名实体匹配技术的快速探测方法,这种方法主要侧重减少报道相似性计算的时间开销。张小明[12]等提出增量聚类的自动话题探测,准确率和效率有一定提升。王灿辉[13]等使用TF·IWF(Term Frequency & Inverse Word Frequency)建立报道模型,进而建立新闻专题。文献[15]使用新闻要素建立报道模型,提出加窗的在线新闻事件探测方法,其中窗口内报道使用凝聚层次聚类建立候选事件集,然后再将候选事件与已有事件进行相似性比较。总的来说,以往的在线新闻事件探测主要采用TF·IDF建立报道的向量模型[1-5,9,14-15],以单篇报道作为统计单元,计算报道中的特征权重,而事件模型则用事件包含的所有报道向量的质心表示。这种事件模型本质是以单篇新闻报道向量作为事件向量模型计算的基本单元,仅仅是将多篇报道的向量特征权重进行求和平均,这种模型不能很准确地反映特征在事件中的重要程度。

因此,为了更准确地表征事件模型,受TF·IDF思想启发,本文在文献[15]的探测策略基础上,提出一种基于TF·IEF(Term Frequency & Inverse Event Frequency)模型的在线新闻事件探测方法,提高探测的效率和准确率。

* **收稿日期**:2013-03-05
基金项目:国家部委资助项目;国家自然科学基金资助项目(61170158);湖南省自然科学基金资助项目(12JJ5028)
作者简介:张辉(1983—),男,湖南湘潭人,博士研究生,E-mail:zhanghui@nudt.edu.cn;
李国辉(通信作者),男,教授,博士,博士生导师,E-mail:guohli@nudt.edu.cn

1 TF · IEF 模型

1.1 基于 TF · IEF 的事件模型

借鉴 TF · IDF 思想[8],本文提出一种基于事件的特征权重计算模型,称为 TF · IEF,TF(Term Frequency)表示特征在事件中的频次,IEF(Inverse Event Frequency)表示特征的逆事件频次。

设 $E=\{E_i|i=1,2,\cdots,l\}$ 表示一个事件集合,$\omega(t_k,E_i)$ 表示第 i 个事件中第 k 个特征的权重值,$E_i=\{(t_k,\omega(t_k,E_i))|k=1,2,\cdots,m\}$ 表示事件向量 E_i 的 m 个特征及其权重,TF · IEF 计算事件特征权重如式(1)、(2)所示:

$$\omega(t_k,E_i)=\frac{[1+\log_2 TF(t_k,E_i)]\cdot IEF(t_k)}{\sqrt{\sum_{k=1}^{m}\{[1+\log_2 TF(t_k,E_i)]\cdot IEF(t_k)\}^2}} \tag{1}$$

$$IEF(t_k)=\log_2\frac{|E|+1}{|EF(t_k)|+0.5} \tag{2}$$

其中,$TF(t_k,E_i)$ 是特征 t_k 在事件 E_i 包含的各报道中出现次数之和,$|EF(t_k)|$ 是出现特征词 t_k 的事件数;$|E|$ 是总的事件数。

从上式可以看出,TF · IEF 与以往权重计算方法不同[1-15],它将事件作为特征权重计算的基本单元,直接计算特征在事件中的权重值。TF · IEF 的建模思想:如果一个特征在一个事件中频繁出现,而在其他事件中很少出现,则认为此特征具有很好的事件区分能力,适合用来进行事件分类。从式(1)的 TF 部分可看出:当 TF 越大,特征权重就越大,但这种情况不是绝对的,当特征在每个事件中均频繁出现时,则该特征对区别不同事件能起的作用就不大,所以使用 IEF 对 TF 进行加权。从式(1)的 IEF 部分看出,当特征在较少的事件中出现,则特征的 IEF 越大,特征有较好的区别不同事件的能力。

本文使用 TF · IEF 计算事件包含的特征权重,并排序,选取 Top-N 个特征建立事件模型。过滤掉权重小的特征,有利于保留主要的特征信息,弱化噪声对事件模型的干扰,使模型具有较准确的事件描述能力,从而有利于提高事件探测的准确性。

1.2 增量式 TF · IEF 事件模型

在线新闻事件探测针对报道流进行检测,新事件不断达到,事件数以及事件模型不是静态的,需要对事件模型以及事件数进行更新,因此,本文提出增量式 TF · IEF 事件模型。在增量 TF · IEF 模型中,特征的事件频次和事件总数依固定时段动态更新。动态更新策略是将第 n 个时段内收集到的报道全部探测完成后再对系统的事件模型库进行更新,TF、EF 更新方式如式(3)、(4)所示:

$$TF_n(t,E_i)=TF_{n-1}(t,E_i)+TF(t,E_i,MC) \tag{3}$$

$$EF_n(t)=EF_{n-1}(t)+EF(t) \tag{4}$$

其中,$TF_{n-1}(t,E_i)$ 表示在第 $n-1$ 时段结束时特征 t 在事件 E_i 中的频次;$TF(t,E_i,MC)$ 表示第 n 时段加入事件 E_i 的报道集(MC)中特征 t 的频次;$EF(t)$ 表示在第 n 时段结束时新探测到包含特征 t 的事件个数,$EF_{n-1}(t)$ 表示在第 $n-1$ 时段结束时包含特征 t 的事件数。因此根据式(1)~(4),在第 n 时段结束时,特征 t 的权重更新如式(5)所示:

$$\omega(t_k,n,E_i)=\frac{[1+\log_2 TF_n(t_k,E_i)]\cdot IEF_n(t_k)}{\sqrt{\sum_{k=1}^{n}\{[1+\log_2 TF_n(t_k,E_i)]\cdot IEF_n(t_k)\}^2}} \tag{5}$$

1.3 模型相似性计算

本文使用经典的余弦公式对事件向量相似性进行计算,对事件 E_i 与 E_j 的相似性计算如式(6)所示:

$$\mathrm{sim}(E_i,E_j,n)=\sum_{t\in E_i\cap E_j}\omega(t,n,E_i)\cdot\omega(t,n,E_j) \tag{6}$$

其中 $t\in E_i\cap E_j$ 表示两个事件的共有特征,如果不是共有的特征,则两事件向量中该特征的权重乘积为零,相似性计算时该特征可以不予考虑。

2 两阶段探测(TSD)方法

2.1 基本思想

新闻事件发生后,不同新闻网站都会对该事件进行相关报道,固定时段内关于同一事件(特别是热点事件)的相似或相同报道比较多。因此,本文提出一种 TSD(Two Stage Detection)方法,该方法先对固定时段内收集到的相似报道进行聚类,生成微簇[16](MC,Micro-Cluster),并用微簇的向量模型表示簇中多篇新闻报道,再将微簇与已探到的事件匹配,这样避免了簇中多篇报道重复与事件进行匹配。

2.2 算法流程

TSD 方法流程包括预处理、第一阶段聚类、第二阶段微簇匹配三部分,算法流程如图 1 所示。

预处理:预处理包括分词、停用词过滤及词频统计。文中利用中科院开发的 ICTCLAS[17] 软件

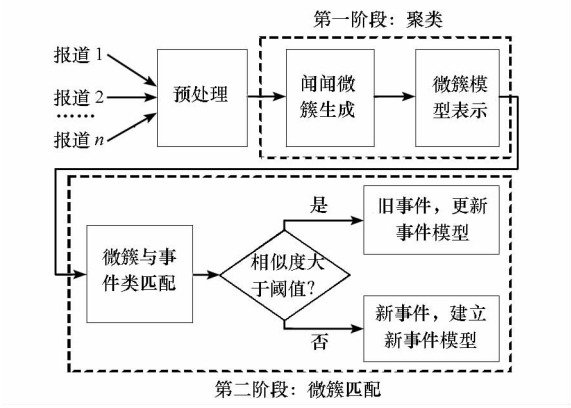

图 1　两阶段探测流程

Fig. 1　Flow chart of two-stage detection

实现分词等预处理,最后得到候选特征集,并统计出特征词频。

第一阶段:利用 Single-Pass 及最大簇间距离度量对固定时段内收集的新闻报道进行初聚类,将关于同一事件的相似报道聚成微簇,最后形成微簇集(候选事件集),并用微簇特征(*MCF*)描述微簇,*MCF* 用向量表示如下:

$$MCF = [Num_{docs},\ Time_{MC},\ Title_{docs},\ E_{MC}] \quad (7)$$

其中 Num_{docs} 表示簇中报道数,$Time_{MC}$ 表示簇时间,用簇中首篇报道发布时间代替,$Title_{docs}$ 表示簇中报道名索引数组,E_{MC} 表示用 TF · IEF 建立的簇模型(候选事件模型)。第一阶段聚类的特点是报道数量较少,信息干扰也较小,通过调整 λ 能较好地区分不同新闻事件,λ 的最优取值通过实验对比分析得出,见第 3.2 节。

第二阶段:将第一阶段生成的微簇集按 $Time_{MC}$ 的顺序与已有的事件类逐一进行相似度计算,选取最大相似值与阈值 θ 进行比较,如果大于(等于)阈值,则微簇与最大相似值事件类是同一类,微簇是旧事件;如果小于阈值,则微簇是一个新事件。

2.3　算法描述

假设 D_n 表示第 n 时段内按发布时间顺序收集的报道集;λ 和 θ 分别表示第一、二阶段阈值;MC_s 表示微簇集;*MaxSim* 表示最大相似值;E 为事件集; ONED 的目标是检测出事件的首次报道,用 D_{New} 表示新事件的首篇报道集合,作为算法的输出结果。TSD 算法设计如下:

第一阶段:

Step 1:输入第 n 时段内收集到的报道集 D_n,将 D_n 中的报道进行预处理;

Step 2:将 D_n 中 d_1 作为 MC_1,用式(1)和(2)计算特征向量,建立 MCF_1,将 MC_1 添加到 MC_s 中;

Step 3:顺序读取 D_n 中下一篇 d_i($i \neq 1$),将 d_i 当作一个事件,用式(1)和(2)计算 d_i 向量,比较 d_i 与 MC_s 中所有簇的相似性,获得最大相似值并赋予 *MaxSim*,并标记最大相似簇为 MC_{max};

Step 4:将 *MaxSim* 与 λ 进行比较,如果 *MaxSim* 大于 λ,则 d_i 属于 MC_{max},将 d_i 加入 MC_{max},使用式(1)和(2)重新计算簇 MC_{max} 的特征权重;如果 *MaxSim* 小于或等于 λ,则 d_i 为新的 *MC*,建立 MC_{New} 和 MCF_{New},将 MC_{New} 添加到 MC_s 中;

Step 5:重复 Step3 至 Step4,直到 D_n 中所有报道全部完成匹配,输出 MC_s,第一阶段聚类结束。

第二阶段:

Step 1:按簇的时间顺序输入第一阶段生成的 MC_s;

Step 2:按顺序读取一个 MC_i,用式(1)和(2)计算 MC_i 的向量特征,并与 E 中每个事件 E_i 进行相似匹配,将最大相似值赋予 *MaxSim*,并标记对应的事件类 E_{max};

Step 3:将 *MaxSim* 与 θ 进行比较,如果 *MaxSim* 大于 θ,则 MC_i 属于 E_{max},将 MC_i 加入 E_{max};如果 *MaxSim* 小于或等于 θ,则 MC_i 为新事件,建立 E_{New},该 MC_i 加入 E_{New},将 E_{New} 添加到 E 中,并将 MC_i 中首篇报道名加入 D_{New} 中;

Step 4:重复 Step2 与 Step3,直到 MC_s 中所有微簇匹配完成,然后使用式(3) ~ (5)更新事件模型,并输出 D_{New},探测过程结束。

2.4　算法复杂度分析

假设 N_E、N_{D_n}、N_{MCs} 分别表示聚类完成后总的事件数、D_n 中报道数、第一阶段聚类后的微簇数。将在线新闻事件探测中公认的 Single-Pass 聚类算法[1-5,9] 及本文中提出的 TSD 算法的渐进复杂度进行对比,说明两个算法的优劣,对比结果如表 1 所示。

表 1　算法复杂度比较

Tab. 1　Comparison of algorithm complexity

探测算法	渐进复杂度上界
Single-Pass	$O(N_E * N_{D_n})$
TSD	$O(N_{D_n} * N_{MCs}) + O(N_E * N_{MCs})$

根据算法复杂度的加法规则[16] 有如式(8)的变换:

$$O(N_{D_n} * N_{MCs}) + O(N_E * N_{MCs})$$

$$= O(max(N_{D_n} * N_{MCs}, \quad N_E * N_{MCs})) = \begin{cases} O(N_E * N_{MCs}) & N_E \gg N_{D_n} \\ O(N_{D_n} * N_{MCs}) & N_E \ll N_{D_n} \end{cases} \quad (8)$$

通过式(8)可以看出,Single-Pass 算法与 TSD 算法的渐进复杂度比较主要考虑 N_E 与 N_{D_n} 的量级关系。当 $N_E \gg N_{D_n}$ 时,TSD 的渐进复杂度为 $O(N_E * N_{MCs})$,因通常有 $N_{MCs} < N_{D_n}$($N_{MCs} = N_{D_n}$ 几乎不可能),所以 TSD 的渐进复杂度小于 Single-Pass;当 $N_E \ll N_{D_n}$ 时,TSD 的渐进复杂度为 $O(N_{D_n} * N_{MCs})$,因通常有 $N_{MCs} < N_E$,所以 TSD 的渐进复杂度小于 Single-Pass。当 N_E 与 N_{D_n} 属于同一量级时,TSD 算法的复杂度为 $O(N_E * N_{MCs})$ 或 $O(N_{D_n} * N_{MCs})$,分析同上。所以,从理论上分析,在不考虑单次运算时间消耗的情况下,仅从算法的渐进复杂度来看,TSD 算法的探测效率将优于 Single-Pass 算法,文中接下来将用实验进行验证分析。

3 实验及分析

3.1 实验数据及性能指标

实验采用 TDT4 中文语料[19]进行评测。TDT4 语料涵盖了多个新闻源,有 70 个话题类,共计 1257 篇新闻报道,话题包含新闻报道篇数从 1 到 140 不等,具有良好的时序信息,主要用来对 TF · IEF 模型和在线探测策略进行性能测试。

实验基于 TDT 发布的评测指南[20],采用探测错误代价 C_{Det}、丢失率 P_{Miss}、误报率 P_{False} 三个指标对系统准确率进行评测;同时,通过比较系统的时间消耗,对系统的效率进行评测。C_{Det} 的计算如式(9)所示:

$$C_{Det} = C_{Miss} P_{Miss} P_{target} + C_{False} P_{False} P_{non-target} \quad (9)$$

其中,C_{Miss} 和 C_{False} 表示丢失和误报的代价系数,P_{target} 表示先验目标概率,三个参数设置参考 TDT 评测[20]。探测错误代价 C_{Det} 的规范化形式 $Norm(C_{Det})$ 如式(10)所示。同时,实验采用可视化的评测工具 DET(探测错误权衡图)曲线[21]来绘制系统的性能曲线,DET 是根据丢失率和误报率随阈值 θ 的变化趋势来绘制的。由于系统丢失率与误报率越低,性能越好,因此 DET 曲线越靠近坐标系的左下角,代表系统性能更优。DET 曲线上的最小 $Norm(C_{Det})$ 指标代表探测系统的最佳性能,简写为 $Min(C_{Det})$。

$$Norm(C_{Det}) = \frac{C_{Det}}{\min(C_{Miss} P_{Miss}, C_{False} P_{False})} \quad (10)$$

3.2 第一阶段阈值 λ 估计

在 TSD 算法中,第一阶段的阈值 λ 设置很关键。λ 设置的好坏直接关系到第二阶段聚类的效果。λ 设置过高,将本该属于一个微簇的报道分开,增加第二阶段的计算开销;设置过低,会将不属于一个微簇的报道聚到一个微簇,从而形成噪声,影响第二阶段的聚类质量。

实验设置 λ 从 0.06 变化至 0.13,步长为 0.01,绘制 DET 曲线。最优 λ 值估计是通过观察 DET 曲线随 λ 的变化情况而得出。为方便观察,选取三组较好实验结果绘制 DET 曲线,如图 2 所示,图中横轴表示误报率,纵轴表示丢失率;二维坐标系内的一条曲线代表一个固定的 λ 值对应的系统探测事件的整体性能,单一曲线的变化过程表示系统在 λ 值不变的情况下随阈值 θ 变化的一组误报率和丢失率指标;不同曲线表示在不同 λ 值时系统随阈值 θ 变化的性能指标;图中三个几何形状的标识分别表示三个不同的 λ 值对应的系统性能曲线的最小规范化探测错误代价 $Min(C_{Det})$。

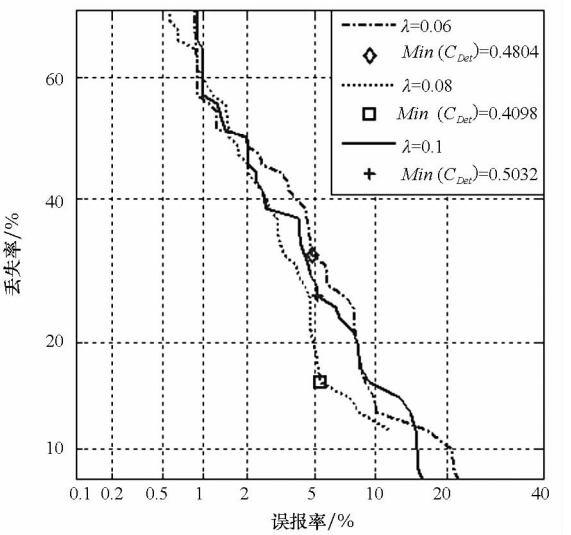

图 2 不同阈值 λ 的 $Norm(C_{Det})$ 曲线

Fig. 2 DET curves with different threshold λ

通过观察 DET 曲线,$\lambda = 0.08$ 时,曲线较靠近坐标左下角,且 $Min(C_{Det})$ 也最小。基于这一结果,实验取最优 λ 值为 0.08,在进行系统的性能对比实验时选用最优 λ 值。

3.3 探测效率及分析

系统探测效率分析主要通过系统执行的时间消耗来说明,使用统一的 TF · IEF 模型对事件建模,分别用 Single-Pass 算法和本文的 TSD 算法对语料进行实验,结果如表 2 所示。参与对比的三个时间指标是:系统在取不同阈值 θ 时的平均时间消耗、最大时间消耗以及 $Min(C_{Det})$ 对应的时间消耗。同时,根据不同阈值 θ 对应的系统时间消耗绘制系统的时间消耗曲线,如图 3 所示。

从表 2、图 3 均可以看出:TSD 算法的效率明显高于 Single-Pass 算法,各指标提升的效率均在 31% 以上。因为相对于 Single-Pass 算法,TSD 算法通过第一阶段的聚类,多篇报道聚集成微簇集,再将微簇与事件类匹配,减少了相似报道的重复匹配次数,进而提高了系统的效率。同时,通过表 1 的理论分析可知,这种效率提升在事件类较多且一定时段内收集的相似报道越多的情况下,会表现得更为明显。

表 2　实验探测效率对比

Tab. 2　Efficiency comparison

探测效率指标	TSD	Single-Pass
Average Time Cost(s)	9.353	15.305
Max Time Cost(s)	11.649	16.989
Min (C_{Det}) Time Cost(s)	8.467	12.982

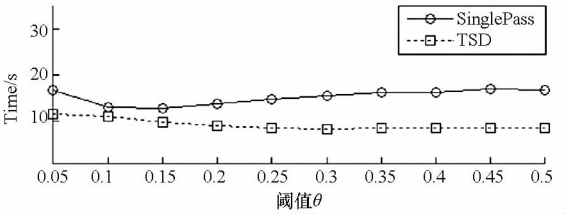

图 3　系统运行的时间消耗曲线

Fig. 3　Time consumption curve of systems

3.4　探测准确率及分析

准确率分析主要通过绘制 DET 曲线图进行说明,根据探测实验结果绘制 DET 曲线如图 4、图 5 所示。图 4 是使用 TF · IDF 和 TF · IEF 两种模型在 Single-Pass 的基础上进行实验绘制的 DET 曲线,主要验证 TF · IEF 模型的性能。图 5 是用 TF · IEF 模型和 TSD 探测策略构成的综合系统进行实验获得的性能曲线,主要是对本文研究内容的总体评价,参与对比的系统是 Single-Pass + TF · IDF。

从图 4 可以看出,基于相同的探测策略,TF · IEF 模型绘制的 DET 曲线更趋近于坐标的左下角,同时,TF · IEF 模型的 $Min(C_{Det})$ 要小于 TF · IDF 模型的 $Min(C_{Det})$ 值,因此,TF · IEF 模型优于 TF · IDF 模型。原因主要是: TF · IEF 将事件作为计算单元,所计算的权重直接反映特征表征事件的贡献大小,获得了较好的事件表示模型,同时利用 Top-N 过滤掉部分噪声特征,模型对事件描述更准确。

从图 5 的 DET 曲线同样可以看出,相对于 Single-Pass + TF · IDF 系统的 DET 曲线, TSD + TF · IEF 系统对应的曲线总体上离坐标系左下角比较近。同时,从图 5 的最佳性能指标 $Min(C_{Det})$ 来看, TSD + TF · IEF 系统的 $Min(C_{Det})$ 优于 Single-Pass + TF · IDF 系统。因此,Single-Pass + TF · IDF

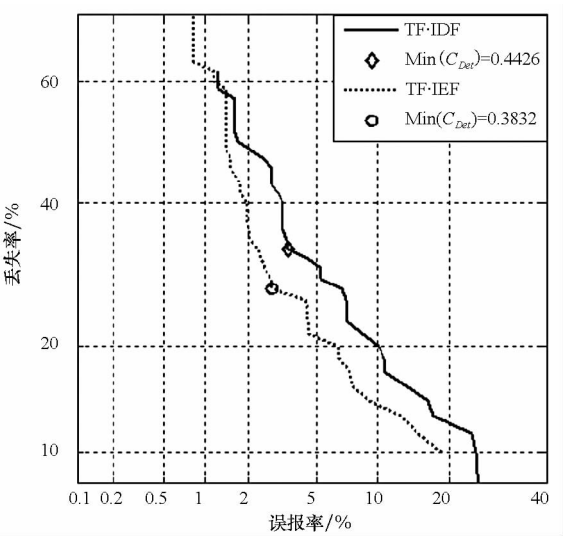

图 4　不同模型的 DET 曲线

Fig. 4　DET curves for different event model

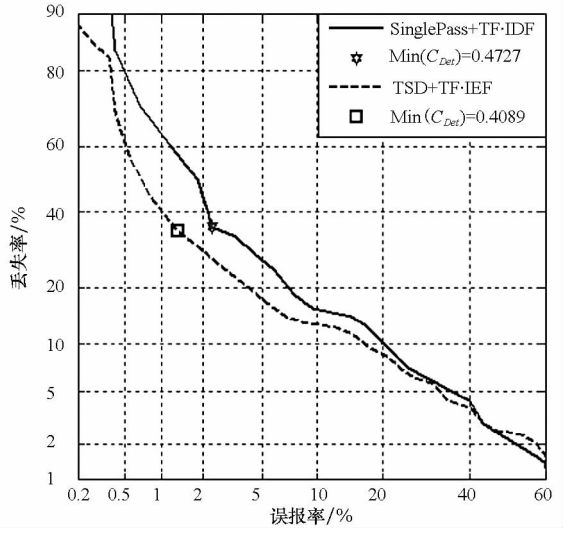

图 5　不同探测系统的 DET 曲线

Fig. 5　DET curves for different systems

系统的整体性能要优于 TSD + TF · IEF 系统。分析主要有两个方面的原因:一方面是利用 TF · IEF 建模;另一方面是时间上越接近的报道,更有可能讨论同一新闻事件,因此第一阶段对固定时间段内的报道进行聚类,能够获得较好的候选事件集。

4　结论

本文利用增量式 TF · IEF 模型建立事件的特征权重向量,并在此基础上,使用分阶段探测策略对网络新闻报道流进行聚类,判断待测新报道是否描述了一个新事件,最后通过实验验证了方法的有效性。通过分析、实验可知,这种方法建立的事件模型比较准确,噪声干扰较小,算法复杂度相对较小,能够较好地适应新闻报道流实时、海量、时序的特性。

参考文献(References)

[1] 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, New York: ACM Press, 1998:37 – 45.

[2] Papka R, Allan J. On-line new event detection using single pass clustering TITLE2[R]. University of Massachusetts, Amherst, MA, 1998.

[3] Yang Y, Pierce T, Carbonell J. A study on retrospective and on-line event detection[C]//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York: ACM Press, 1998: 28 – 36.

[4] Lam W, Meng H, Wong K, et al. Using contextual analysis for news event detection[J]. Int'l Journal on Intelligent Systems, 2001, 16(4):525 – 546.

[5] Brants T, Chen F, Farahat A. A system for new event detection[C]//Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York: ACM Press, 2003: 330 – 337.

[6] Nieola S, Joe C. Combining semantic and syntactic document classifiers to improve first story detection[C]//Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York: ACM Press, 2001: 424 – 425.

[7] Kumaran G, Allan J. Text classification and named entities for new event detection[C]//Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York: ACM Press, 2004: 297 – 304.

[8] Zhang K, Li J Z, Wu G. New event detection based on indexing-tree and named entity[C]//Proceeding of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York: ACM Press, 2007:215 – 222.

[9] 张阔, 李涓子, 吴刚, 等. 基于词元再评估的新事件探测模型[J]. 软件学报, 2008, 19(4):817 – 828.
ZHANG Kuo, LI Juanzi, WU Gang, et al. A new event detection model based on term reweighting [J]. Journal of Software, 2008, 19(4):817 – 828. (in Chinese)

[10] 付艳, 杨冬青, 唐世渭, 等. 基于实体识别的在线主题检测方法[J]. 北京大学学报(自然科学版), 2009, 45(2): 227 – 232.
FU Yan, YANG Dongqing, TANG Shiwei, et al. On-line topic detection using named entity recognition [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2009, 45(2):227 – 232. (in Chinese)

[11] 付艳,周明全,王学松,等. 面向互联网新闻的在线事件检测[J]. 软件学报,2010,21(增):363 – 372.
FU Yan, ZHOU Mingquan, WANG Xuesong, et al. On-line event detection from web news stream [J]. Journal of Software, 2010, 21(Supplement):363 – 372. (in Chinese)

[12] 张小明, 李舟军, 巢文涵. 基于增量聚类的自动话题检测研究[J]. 软件学报, 2012, 23(6):1578 – 1587.
ZHANG Xiaoming, LI Zhoujun, CHAO Wenhan. Research of automatic topic detection based on incremental clustering[J]. Journal of Software, 2012, 23(6): 1578 – 1587. (in Chinese)

[13] Wang C H, Zhang M, Ma S P. Automatic online news issue construction in web environment[C]//Proceedings of the 17th International Conference on World Wide Web, Beijing, 2008: 457 – 466.

[14] Xu R F, Peng W H, Xu J. On-line new event detection using time window strategy [C]//Proceedings of the 2011 International Conference on Machine Learning and Cybernetics, Guilin, 2011:1932 – 1937.

[15] Zhang H, Li G H. One Method For On-line News Event Detection Based On The News Factors Modeling [C]// Proceedings of the Sixth International Conference on Intelligent Systems and Knowledge Engineering, Shanghai, 2011: 427 – 434.

[16] Han J W, Kamber M. Data mining concepts and techniques [M]. Second Edition, Diane Cerra Press, 2007.

[17] Zhang H P, Liu Q. Calculation of the Chinese lexical analysis system LCTCLAS [CP/OL]. Institute of Computing, Chinese Academy of Sciences, 2002. http://sewm. pku. e du. cn/QA/reference/LCTCLAS/FreeICTCLAS/

[18] Chen M, Algorithm analysis preliminary [EB/OL]. http://blog. csdn. net/hnzmdzcm/article /details/7626183 2012

[19] 张晓燕. 新闻话题表示模型和关联追踪技术研究[D]. 长沙:国防科学技术大学, 2010.
ZHANG Xiaoyan. Research on the representation model and technologies of link detection and tracking on news topic[D]. Changsha: National University of Defense Technolog, 2010. (in Chinese)

[20] NIST. The 2003 topic detection and tracking task definition and evaluation plan[R]. http://www. itl. nist. gov/iaui/894. 01/tests/tdt/tdt2003/evalplan. htm: National Institute of Standards and Technology(NIST), 2003.

[21] 洪宇,张宇,范基礼,等. 基于子话题分治匹配的新事件检测[J]. 计算机学报,2008,31(4):687 – 695.
HONG Yu, ZHANG Yu, FAN Jili, et al. New event detection based on division comparison of subtopic [J]. Chinese Journal of Computers, 2008, 31(4):687 – 695. (in Chinese)