

一种基于时序窗口的动态热点话题提取模型^①

马慧芳^{②*} 尹 旻^{***} 何 清^{*} 史忠植^{*}

(^{*}中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

(^{**}中国科学院研究生院 北京 100049)

(^{***}联想(中国)研究院网络服务研究室 北京 100085)

摘 要 针对新闻领域的专题组织进行了研究,提出了一种基于时序窗口的动态热点话题提取模型。该模型整合了热点话题的两个特点。一方面关注主题词在新闻文本中的广泛性,衡量标准为多频道播报特征项的频率综合,词频越高其广泛性越高;另一方面考虑新闻流主题词的突发性,表现为特定时间段内主题词出现频率显著异常于其它时间段。引入时序窗口进行上升和下降突发模式提取,并结合 TF-DF 作为主题词赋权值依据。实验结果表明,这种基于时序窗口的动态热点话题提取模型对新闻文本进行主题抽取具有很好的性能。

关键词 话题提取,时序窗口,广泛性,突发性,TF-PDF

0 引 言

话题自动提取在文本挖掘和信息检索领域中有着重要用途。主题提取的概念最早由 Luhm^[1]于1956年提出,基本思想是利用词语出现频率和分布等统计信息计算词语及句子的相对重要程度,提取并输出重要度最高的主题词或句子来获取文本主题。传统提取主题的方法通常基于词频统计法,这种方法简单朴素,认为文本中相对使用频率较高的词反映了文章的主题。该方法的优势在于实现简单,然而由于对词的处理过于简化,完全没有考虑词语之间的语义联系,这种处理方式显然会影响提取主题的质量。基于语义分析和理解的主题提取方法需要对文本进行深层句法和语义分析。一种典型的理解式方法是建立预定制框架,从原文中提炼重要的信息填入框架中,从而生成主题。这种方法的优点是考虑了深层语义信息,但需要较成熟的人工智能技术和大型专家知识库的支持。随着信息量的飞速增长,网络上信息越来越丰富,越来越多的研究聚焦于自动探测新闻、聚类和组织相关信息的方法。热点话题提取作为话题检测与跟踪(topic detection and tracking, TDT)^[2]最核心和最基本的课题之一,也日益受到重视。本文旨在以新闻领域专题组织为背

景提出一种动态提取热点话题模型,应用场景由传统静态文本集转向动态文本流,除了关心哪些特征主题词出现是频繁的,同时希望能捕获近来哪些特征主题词的出现具有突发性,将广泛性和突发性二者有机地整合成为模型构建的基础。其中前者体现了热点主题词具有多个频道播报且高频的特点,而突发性则是侧重为特定时间段出现频率异常的热点特征主题词赋予高权值,将这两者的整合可以提高动态获取主题词的效率。本文首先回顾了相关的研究工作并简要介绍了话题特点定义分析、已有的话题提取模型以及突发模式提取,然后论述了本文提出的话题突发提取算法以及基于突发与广泛性的话题提取模型,并给出了相关实验及分析,最后给出了结论。

1 话题定义与特点分析

1.1 问题定义与相关研究

定义 1(事件) 在话题检测与跟踪(TDT)研究中,定义为发生于特定时间和特定地点的事情^[2]。

定义 2(话题) 在 TDT 研究中,定义为新闻话题由种子事件及其直接相关的事件共同组成,其往往表现为某一新闻事件引发的所有相关报道的综合^[2]。

^① 863 计划(2007AA01Z132),国家自然科学基金(60435010), 973 计划(2007CB311004)和国家科技支撑计划(No. 2006BAC08B06)资助项目。
^② 女,1981 年生,博士生;研究方向:人工智能,数据挖掘与机器学习;联系人, E-mail: mahf@ics.ict.ac.cn
(收稿日期:2009-03-06)

定义 3(热点话题) 在本研究背景下,采用文献[3]的定义,定义为在特定时间段内出现频率高的主题特征词。其热度取决于两点:一是在单个文本内出现的频率;二是包含该话题的文本数。其特点在于动态进化特征,也即热点话题具有生命周期。

1.2 基于 TF-PDF 的赋权模式及其改进

主题词特征项加权通常采用基于向量空间模型的 TF-IDF 模式^[4]。其中 TF(term frequency)代表词频, IDF(inverse document frequency)代表逆文档频率。由于赋权模式整合了特征项的出现频率以及在整个文档中的分布情况,赋权的偏置是对在整个数据集高频且指示度高的特征项赋高权值。然而在热点话题提取算法的应用场景下,整个数据集中大量新闻报道中出现的主题特征词也应该被识别出来。文献[3]提出了一种适用于 TDT 领域的加权模式 TF-PDF(term frequency-proportional document frequency),这种加权方式对多个频道播报也即出现在多个新闻报道中的主题特征项赋予高权值。文献[5]指出该方法的弱点在于没有考虑主题词热度随报道时间动态变化的特征,并提出了一种融合话题广泛性和主题性基于时间线索的热点特征词提取算法。前者与本文意义相同,由 TF-PDF 来体现,后者利用生命周期建模来体现特征词的使用频率随时间变化的特点。与文献[5]不同,本文提出的动态热点话题提取模型引入时序文本流突发模式提取算法来获取具有高突发性的主题特征词,同时整合广泛性来有效捕获热点话题。

1.3 时序文本流突发模式提取

数据流上某段时间内出现了异常数目的元素,这种异常称作突发(burst),也叫突变。它与数据流不断变化与难以预测的特性相对应。在数据流上进行突发检测是数据流的内在、固有问题之一。数据流突发检测在金融、互联网监测、控制及趋势分析、传感器网络管理、科学计算等领域有着广泛应用和深入研究。数据流上的突发目前没有统一定义,在不同的应用场景和应用领域,突发也有不同的定义。一般地,突发定义如下:

定义 4(突发) 数据流中某段时间内数据量显著地不同于其它时间段内数据量的异常现象。

依照数据流监测类型和数据流窗口模型的不同,突发检测的模型分为两类:一是基于点监测的自动机模型,二是基于聚合监测的滑动窗口模型。

在互联网新闻报道中提取热点特征词的应用背景下,考虑到的一个思路是提取特征项的突发模式

作为突发特性的赋权依据。对时序文本流的突发监测可以采用两个相邻窗口的能量强度比值来衡量,通过上升突发和下降突发比较,提供突发模式挖掘的度量标准。本文的突发模式提取以文献[6]为基础,这种突发模式提取方法为时序文本流的突发模式提取奠定了基础。以下是用到的基本概念:

定义 5(时序窗口) 设 TW 对应一个新闻文本流子序列,该文本流的长度记为 $tw.size = |TW|$,时间跨度定义为 $tw.span = [st, et]$,其中 st 表示该时间窗口的起始时间, et 表示该时间窗口的结束时间。一个基本窗口作为一次模式提取的基本单元。称 TW 标识一个基本时序窗口。整个时序文本流对应一个时序模式集合,记作 $tw.patternset$,其中每个模式记作 $tw.pattern$ 。此时序模式用作突发模式提取的基本单元。

定义 6(滑动时间窗口) 用 STW 来标识,对应一个连续时序窗口序列。滑动窗口最初用于 TCP 传输控制,滑动窗口的大小意味着接收方还有多大缓冲可以接收数据,发送方可以通过滑动窗口的大小来确定应该发送多少字节的数据。此处表示为 $STW = tw_1, tw_2, \dots, tw_k$ 。

定义 7(能量强度) 用 w 标识时序文本流中选定的窗口, p 是窗口 w 中对应的一个主题特征词,则 p 在 w 中的能量强度定义为

$$E_{pw} = \sum_{c \in C} \chi_{s,c}^2 \quad (1)$$

其中 C 表示新闻报道的频道集合; $\chi_{s,c}^2$ 表示在时序窗口 w 中主题词 p 与时序窗口 w 的相关度,用式

$$\frac{(A + B + E + F) * (AF - BE)^2}{(A + B) * (E + F) * (A + E) * (B + F)} \quad (2)$$

计算^[4]。其中 A 是在 tw 时间段频道 c 播报 p 的次数; B 是在 tw 时间段内其他频道播报 p 的次数; E 是不在 tw 时间段内频道 c 播报 p 的次数; F 是不在 tw 时间段内其他频道播报 p 的次数。

定义 8(能量比率) 设 TTW 是数据流中选定的考察窗口, CTW 是比照窗口, P 是 TTW 中任意模式,则 TTW 中的模式 P 相对于比照窗口 CTW 的能量比率为

$$Ratio(P_{TTW, CTW}) = \frac{E_{p, TTW}}{E_{p, CTW}} \quad (3)$$

定义 9(突发模式) 设 RSW 为数据流中选定的考察窗口, RFW 为参考窗口, P 是 RSW 中任意模式,再设 $[ls, hs]$ 为边界参数阈值,则 RSW 中的模式 P 相对于参考窗口 RFW 出现突发当且仅当

$$\text{Ratio}(P_{RSW, RFW}) \geqslant hs \text{ 或 } \text{Ratio}(P_{RSW, RFW}) \leqslant ls \quad (4)$$

则突发模式定义为所有给定考察窗口中相对于参考窗口满足设定阈值条件的模式集合。

2 基于时序窗口的热点话题提取算法

2.1 确定提取特征项表

把所有在新闻文本流中出现的词语作为主题词提取的来源显然计算代价太大,因而仅关心一个特征项表。对整个词表首先做预处理,去除停用词,并利用波特词干算法,也即波特词干器(port stemmer)^[7]除去英文单词分词变换形式的结尾,提取出单词词干。对得到的特征项按频率高低排序,大于设定阈值的特征项作为进一步处理的特征项表。

2.2 基于时序窗口的动态热点话题提取算法

特征项的突发权重定义为该特征项在整个滑动窗口内突发值之和;广泛性则定义为 TF-PDF 对应权值。算法 1 给出了突发模式提取突发权重的过程,算法 2 则给出了计算最终权重的伪码。

算法 1 突发模式提取

输入: D : 由预处理 2.1 得到的特征项集;

ls : 边界参数阈值;

hs : 边界参数阈值。

输出: 对应每个特征项的突发权重 t_{burst} 。

步骤:

1. 初始化所有特征项的突发权重 $t_{burst} = 0$;

2. 对特征项集 D 中每个特征项 t , 作如下循环:

计算特征项 t 在时序窗口 tw_0 的能量强度;

For $i = 1$ to k do

计算特征项 t 在时序窗口 tw_i 的能量强度;

计算 $t_{ratio}(tw_{i-1}, tw_i)$;

If ($t_{ratio}(tw_{i-1}, tw_i) \geqslant hs$)

$t_{burst} = t_{burst} + t_{ratio}(tw_{i-1}, tw_i)$

Else if

$t_{burst} = t_{burst} + 1/t_{ratio}(tw_{i-1}, tw_i)$

3. 输出每个特征项的突发权重 t_{burst} 。

算法 2 动态话题提取算法

对于特征词表的所有特征词:

1. 用下面的公式计算特征词 t 的原始 TF-PDF 值:

$$|F_{jc}| = \frac{F_{jc}}{\sqrt{\sum_{k=1}^{k=K} F_{kcr}^2}} \quad (5)$$

$$W_j = \sum_{c=1}^{c=|C|} |F_{jc}| \exp\left(\frac{n_{jc}}{N_c}\right) \quad (6)$$

其中 W_j 表示特征项 t_j 的权重; F_{jc} 表示特征项 t_j 在频道 c 的播报频率; n_{jc} 表示频道 c 播报中包含特征项 t_j 的新闻文档个数; N_c 表示频道 c 播报的新闻文档总数; k 表示频道 c 播报的所有特征项; $|C|$ 表示播报新闻的频道总数。

2. 规范化特征项 t 对应的 TF-PDF 权值 W_j 与突发权重 t_{burst} 。

3. 计算特征项 t 的最终权值:

$$t_{new} = \alpha \times W_j + (1 - \alpha) \times t_{burst} \quad (7)$$

TF-PDF 加权公式具有三大特点:一是特征项的最终权值是各个播报频道获取的权值综合;二是用 F_{jc} 表示特征项 t_j 频率,由于不同新闻频道的播报规模不尽相同,这种规范化 F_{jc} 的方法可以对源于不同频道的相同特征项给予同样的赋权机会;三是作为 PDF 的实现, $\exp(n_{jc}/N_c)$ 体现了相对于某个频道的整个播报文档集包含某个特征词的文档数目幂级增长的特点。与传统的 IDF 赋值相比,这种基于 TF-PDF 的赋权模式可以将整个数据集中大量新闻报道中出现的主题特征词识别出来。

按照由算法 2 得到的新权重降序排列特征项和该序列的前 k 个特征项构建热点特征项表。这里 k 仅仅是一个可调参数用于限制提取过程的话题数目, k 越大则提取结果热度越高,反之亦然。依照特征项权重进而可以通过句子聚类获得摘要生成主题。句子的权重为其包含所有特征项的平均权重之和。与特征项类似的,通过抽取前 k 个句子作为热点句子来作为聚类的前提,每个聚类的结果则为某特定话题。文献[8]提出的构建多文档摘要的文本关系图作为热点句子的聚类的基础。

3 实验与结果分析

3.1 数据集

LDC^[1] 为 TDT 研究提供了 TDT 预研语料、TDT2、TDT3、TDT4 和 TDT5 共 5 期语料。这些语料均源于各新闻媒体, TDT5 只包含文本形式的新闻报道,而其他语料同时包含文本和广播两种形式的新闻报道。

本文实验利用 TDT5 的标注英语语料。TDT5 语料都来自于新闻专线,包含着英语、汉语、阿拉伯语 3 种语言的语料,时间跨度为 6 个月(2003-04-01 至 2003-09-30)。其中包括 278109 篇英语报道, 56486

篇汉语报道和 72910 篇阿拉伯语报道。语料中一共标注了 250 个话题。由于 TDT5 语料的时间跨度较大,而本文算法关注于识别短时期内的突发情况,所以实验中将 TDT5 的整个语料按月份划分为 6 个子集,并以月份名称命名各个子集。以数据集 *April* 为例,该集合由来源于 4 个报道机构标注了 38 个主题的 733 篇报道组成,然后将少于 10 篇报道的主题去除,最终数据集 *April* 由 633 篇报道组成,共标注 15 个类别。表 1 为数据集 *April* 的具体组成说明。

表 1 数据集 *April* 话题组成

主题编号	报道数目	报道比率	机构数目
55004	10	1.51%	2
55010	21	3.17%	3
55041	45	6.79%	4
55056	22	3.32%	2
55058	19	2.87%	3
55063	33	4.98%	2
55066	25	3.77%	3
55076	56	8.45%	3
55077	39	5.88%	3
55087	18	2.71%	3
55098	136	20.51%	4
55143	17	2.56%	3
55196	38	5.73%	4
55200	130	19.61%	4
55240	54	8.41%	4

3.2 热点特征项赋权分析

与文献[5]一致,本文设计了如下两个实验分析热点特征项的赋值性能。实验 1 比较了本文提出的模型与 TF-PDF 模型的赋值结果,通过该实验证实了本文提出的模型的有效性。实验 2 追踪了热点特征项的突发提取过程。

3.2.1 与 TF-PDF 的比较

表 2 与表 3 是运用 TF-PDF 模型与本文模型得到的排名前 30 的热点特征项提取实验结果。由实验结果可以看出,TF-PDF 模型仅仅考虑了话题覆盖的广泛性而忽略了新闻报道特征词的突发特性,该赋权模式提取得到的关键词项仅体现了特征项的高频特点。与此相对应的,本文提出的模型更好地反映了热点话题的动态变迁与突发特性。尽管本文模型得到的结果中也有一些噪音,与 TF-PDF 模型相比,结果还是比较令人满意的。此外,本文提出的模型与语料库中报道的真实内容更加一致,在下个实验中将会进一步说明。

3.2.2 突发性能提取分析

以表 3 中的特征项“loot”为例来分析其与特征项“data”在各自生存周期的突发情况。从图 1 可以看到,随着时间的变迁,特征项“loot”在其生命周期内出现明显的突发异常。其能量强度在 12 日出现了一次向上突发跳跃,紧接着 13 日出现了下突发,此后 18 日和 19 日到达了能量强度的顶峰,最后在

表 2 TF-PDF 提取的热点特征项

特征项	权值	特征项	权值	特征项	权值	特征项	权值	特征项	权值
Bush	0.792	nation	0.784	told	0.734	House	0.693	there	0.683
official	0.632	U.S.	0.645	British	0.573	kill	0.532	death	0.476
will	0.463	Iraq	0.461	military	0.438	people	0.432	building	0.430
year	0.427	Hussein	0.424	ancient	0.419	country	0.417	solider	0.413
vessel	0.411	why	0.401	attack	0.391	Taipei	0.384	Blair	0.377
medical	0.357	library	0.351	expert	0.348	staff	0.337	Tiger	0.329

表 3 时序窗口模型提取的热点特征项

特征项	权值	特征项	权值	特征项	权值	特征项	权值	特征项	权值
Iraq	0.812	budget	0.804	loot	0.754	Hussein	0.723	nation	0.644
museum	0.618	post-war	0.532	law	0.521	George	0.520	Baghdad	0.473
marathon	0.452	war	0.446	u.s	0.437	House	0.432	ferry	0.397
summit	0.386	Tuesday	0.382	Bush	0.371	Prime	0.362	weapon	0.355
troop	0.347	construction	0.347	pilot	0.336	Mosul	0.319	hospital	0.318
farewell	10.287	Wood	0.276	funeral	0.272	SARS	0.256	whale	0.242

29 日再次出现突发,所有的突发都增强了该特征项的最终突发权重。相反地,特征项“data”在对应的生命周期内没有突发的变化。由于热点话题在一段时期内会频繁出现,因而该实验证明了本文的模型可以为热点话题特征项赋予合理的权重。在实验语料中,有许多关于伊拉克国家博物馆遭到洗劫的报道(对应于主题 55200),特别是在 17 日到 18 日,该类新闻报道达到了顶峰。这证明了本文的算法可以很好地将此突发识别出来。

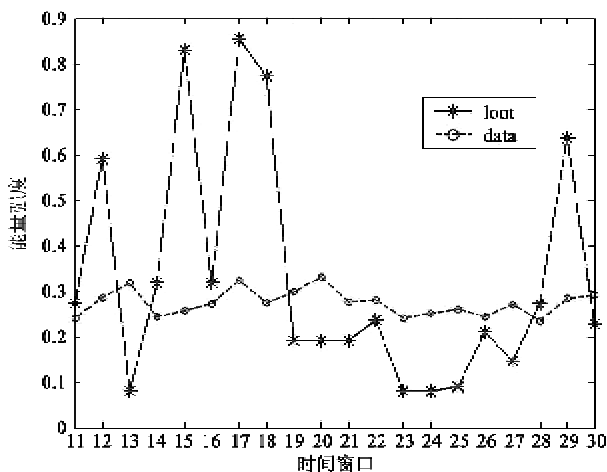


图 1 “loot”与“data”突发示意图

3.3 主题句提取分析

3.3.1 评价指标

为进一步量化分析主题句提取效果,选择特定评价指标进行结果评价。下面定义两个主题句提取评价的指标:正确性和覆盖率。

(1) 正确性(correctness):为了检验最终聚类效果,按文献[3]标准定义如下评价尺度:

- a) 错失句子(MS):错误的聚入某类话题的句子;
- b) 错误句子(FS):未正确聚入某类的句子;
- c) 未聚类句子(NC):未被聚入任何一类的句子。

(2) 覆盖率(coverage rate):该指标用来体现在某时间段内热点话题的识别率,定义为

$$\text{覆盖率} = \frac{\text{提取的热点话题}}{\text{实际热点话题}} \times 100\% \quad (8)$$

3.3.2 正确性

表 4 比较了 TF-PDF 模型与本文提出的模型在正确性方面的结果。其中 SL 表示聚类句子个数。从表中可以看出,本文得到实验结果通常好于 TF-PDF 的结果。特别的,当聚类句子个数多于 200 时

候,错失句子比率下降到最低。

表 4 正确性比较

	MS%		FS%		NCS%	
SL=30	4.2%	2.1%	14.2%	11.2%	27.6%	31.4%
SL=50	7.1%	3.2%	18.2%	14.3%	39.1%	28.1%
SL=100	3.6%	2.7%	24.3%	22.7%	26.9%	21.8%
SL=200	8.3%	1.2%	15.8%	12.0%	13.2%	18.0%
SL=300	2.3%	1.6%	13.6%	13.2%	10.1%	11.0%

3.3.3 覆盖率

本实验的目的是比较本模型提取的热点话题与实际语料中的热点话题的比率。由于聚类过程是从热度最高的句子开始,各个类别中的句子个数可能会不一样多。表 5 比较了 TF-PDF 模型与本文模型的覆盖率,可以看出,随着聚类句子的增多,越来越多的句子被识别出来。当参与聚类的句子个数达到 3%时,覆盖率接近了 90%。由表 5 看到,在大多数情况下,本文提出的模型可以得到更好的结果。

表 5 覆盖率比较

	>5		>1%		>3%	
SL=30	28.2%	33.2%	35.2%	36.2%	47.6%	51.2%
SL=50	41.1%	40.3%	47.2%	44.2%	59.3%	58.4%
SL=100	53.6%	57.2%	59.3%	62.2%	59.3%	86.2%
SL=200	68.3%	72.6%	71.8%	77.2%	69.2%	86.2%
SL=300	72.3%	77.1%	74.6%	78.2%	70.1%	86.2%

4 结 论

本文提出了一种基于时序窗口的动态热点话题提取模型,它可以从新闻语料中提取热点特征项与特征主题句。与传统的为特征项赋权方式不同,该模型一方面考虑了主题词在新闻文本中的广泛性,多频道播报的主题词频越高其广泛性越高;另一方面考虑了新闻流主题词的突发性,特定时间段内的主题词的出现频率显著异常于其它时间段。而且将时序窗口引入进行突发模式提取,并结合 TF-PDF 作为主题词赋权依据。实验结果证明,该模型对于动态新闻热点话题提取具有很好的性能。

该模型存在的不足在于实验中选取的时间段为一个月,并将时间窗口固定为一天,这并不一定适用于各类文本数据流。此外,突发检测中边界参数阈值难于确定,今后的工作将考虑如何解决这些问题。

参考文献

- [1] Luhn H P. The Automatic Creation of Literature Abstracts. *Advances in Automatic Text Summarization*. MIT Press, Cambridge, Massachusetts, USA, 1956. 15-22
- [2] The 2004 topicdetection and tracking task definition and evaluation plan. <http://www.nist.gov/speech/tests/ttdt/ttdf2004/evalplan.htm>; NIST. 2004
- [3] Bun K K, Ishizuka M. Topic extraction from news archive using TF-PDF algorithm. In: *Proceeding of the 3rd International Conference on Web Information Systems Engineering*, Singapore, 2002. 73-82
- [4] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 1989, 4(5): 513-523
- [5] Chen K Y, Luesukprasert L, Chou S T. Hot topic extraction based on timeline analysis and multidimensional sentence modeling. *IEEE Trans TKDE*, 2007, 19(8): 1016-1026
- [6] 宋国杰, 唐世渭, 杨冬青等. 数据流中异常模式的提取与监测. *计算机研究与发展*. 2004, 41(10): 1754-1759
- [7] Porter M. An algorithm for suffix stripping. *Program*, 1980, 14(3): 211-218
- [8] Ma H F, Zhao W Z, Shi Z Z. An approach of multi-document summarization based on text relationship map. In: *Proceeding of the International Conference on Advanced Intelligence*, Beijing, China, 2008, 1536-1542

A dynamic hot topic extraction model based on time window

Ma Huifang^{* **}, Yin Min^{***}, He Qing^{*}, Shi Zhongzhi^{*}

(^{*} The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

(^{**} Graduate University of Chinese Academy of Sciences, Beijing 100049)

(^{***} Internet Service Lab, Lenovo Corporate Research, Beijing 100085)

Abstract

This paper gives a description of a study of topic organization in the news domain, and presents a novel dynamic hot topic extraction model based on the time window. The model combines two characteristics of hot topics together. One is the pervasiveness of topic terms in news texts, which is evaluated by the occurrences of the topic terms reported by different channels, and the more frequent the occurrence of the topic terms reported, the higher the pervasiveness of topic terms. The other one is the burst of topic terms in the news stream, which can be assessed by the abnormal occurrence frequencies of topic terms in a specific interval compared with other different time intervals. The time window is introduced to make burst detection and the term frequency-proportional document frequency (TF-PDF) is combined to weigh the terms. The experimental results demonstrate that this model is effective in topic extraction for news texts.

Key words: topic extraction, time window, pervasiveness, burst, TF-PDF