基于 TF * PDF 的热点关键短语提取*

马佩勋[1]，高 琰[2]
(1. 长沙民政学院 软件学院，长沙 410082；2. 中南大学 信息科学与工程学院，长沙 410000)

摘 要：传统的 TF * PDF 方法提取的关键短语可精确地描述话题并进行新闻报道的追踪，但存在误将噪声数据识别为关键短语的情况。提出了一种基于位置权重 TF * PDF 的两段式关键短语提取方法滤除噪声数据。该方法将传统的 TF * PDF 算法与位置权重相结合，计算词汇与短语的权重，获取候选关键短语列表，关键短语的脉冲值则用于过滤列表中的噪声。通过关键短语识别进程根据位置信息、频率信息等将热点词汇组合成短语。TF * PDF 位置权重算法同时也用于为短语分配权重，排名前 K 的短语被认为是热点关键短语。以真实网络数据为基础的实验结果表明，该提取方法与传统的 TF * PDF 提取方法相比，可更好地去除关键词短语中的绝对噪声，较好地改善了热点话题检测的准确度。

关键词：TF * PDF；TDT；提取；脉冲值；关键词短语
中图分类号：TP391 **文献标志码**：A **文章编号**：1001-3695(2013)12-3610-04
doi:10.3969/j.issn.1001-3695.2013.12.024

Hot keyphrase extraction based on TF * PDF

MA Pei-xun[1], GAO Yan[2]
(1. *Dept. of Software, Changsha Social Work College, Changsha* 410082, *China*; 2. *College of Information Science & Engineering, Central South University, Changsha* 410000, *China*)

Abstract: Key phrase extracted by traditional TF * PDF method could represent topic accurately and track reports effectively, while sometimes noise data may be also recognized as key phrase. This paper proposed two-step key phrase extraction method based on improved TF * PDF to filter noise data. The method combined traditional TF * PDF and position-weight to compute weight of words and phrases, it used obtain candidate hot term list and the burst value of term to filter the noise in the list. In the second step, a phrase identification process combined hot terms into phrases using position information, frequency information etc. At last the position-weighted TF * PDF algorithm are also used to weight the phrase, and chose the top K phrases as hot key phrases. The experiments on the real Web data indicate that this extraction method is able to filter noise data completely and provides a solution with improved quality at topic tracking in comparison with traditional TF * PDF.

Key words: TF * PDF; TDT; extraction; burst value; key phrase

互联网已经成为信息传递的主要媒介，网站上每天都发表很多的新闻报道。人们亦乐于在互联网上表达他们对一些热点事件的观点。网络上能够反映社会或公众的观点，有影响力的新闻和话题有助于当局作出正确的响应。话题检测与跟踪(TDT)任务通常将来自新闻专线与广播的新闻故事结构化成主题[1]。由于网络上信息量巨大，手工检测网络上的热点话题并进行跟踪是一项既耗时又费劲的工作。很多学者在该领域作出了大量的努力[2~6]。一些大型的商业公司开发出了用于 TDT 的实用程序。但是 TDT 技术近乎艺术，离人们满意的期望还很遥远。例如，Google 新闻告警产生了超过 50% 以上的虚假警告[2,7]。在大多数的 TDT 方法中，热点词常代表着热点话题。但是单词无法精确描述话题，尤其是中文。短语在语法和语义结构上比单词要更完整，关键短语比热点词有更为清晰的意义，因此更适合于表示话题，如“自然语言处理”与“自然”“语言”“处理”。本文创新性地提出了一种使用 TD * PDF 提取关键短语以及用关键短语代表热点话题的方法。

该算法提取分为两个步骤：a)热点词提取，使用 TF * PDF 位置权重算法获取候选关键短语列表，关键短语的脉冲值则用于过滤列表中的噪声；b)关键短语提取，相邻的热点词将被合并成短语，本文使用最大重复字符串方法寻找候选短语。接下来，分配位置权重的 TF * PDF 用于对候选关键词的热度排名，排在前 K 名的候选短语就是热点关键短语。

1 相关工作

1.1 话题检测与跟踪

一个话题通常由一个种子事件或活动和伴随直接相关的一系列事件与活动组成，热点话题检测就是发现在一定时期内频繁出现的话题。目前话题检测与跟踪算法主要基于无监督学习算法。Yang 等人[3]使用逐次聚类方法检测主题。另一种流行的方法则是单次通过聚类(递增聚类)[8~10]方法使用词汇权重模式捕获文章内容中重要的或具有代表性的词汇。最常

收稿日期：2013-03-18；**修回日期**：2013-05-27 **基金项目**：国家教育部博士点新教师基金资助项目(20090162120087)；湖南省科技计划资助项目(2009FJ3053)

作者简介：马佩勋(1978-)，男，湖南湘潭人，讲师，硕士，主要研究方向为程序验证、数据挖掘(flymanma@126.com)；高琰(1974-)，女，副教授，博士，主要研究方向为数据挖掘、智能信息处理.

用的词汇权重模式有两个:一个是 TF * IDF[3~5,10],另一个是 Bun 等人[11]提出的 TF * PDF。TF * PDF 强调每个词汇的重要性与唯一性,它只识别同时在文集中的几篇文章中出现的词汇。然而,对于热点话题的提取,代表热点话题的词汇应当在大量文档中频繁地出现。TF * PDF 与 TF * IDF 有所区别。它为在多个渠道多个文档中频繁出现的词汇分配更大的权重,反之亦然。因此 TF * PDF 比 TF * IDF 更适合于提取热点词汇。

1.2　关键词提取

关键词提取有两种类型的算法,即监督学习算法和无监督学习算法。Naive Bayes[12]、决策树[13]以及 SVM[14]都是代表性的监督方法,它们具有令人满意的精确度与相当的稳定性,但是需要对文集进行注解[8]。此外,对于监督算法提取未知关键词则是个难题。无监督算法包括字符串频率、特征权重计算[15~17]和语义网络结构分析方法[18]。为提取更多的信息关键字,学者们开始考虑在整个处理周期内词汇的变化[18,19]。

之前的关键词提取算法都是为单个文档设计的。本文提出的提取关键短语的算法与传统词汇权重模式 TF * IDF 不同的是,本文修改了 TF * PDF 在处理热点关键短语提取过程中为热点词汇分配权重的模式。

2　关键短语提取

关键短语由若干个相关且在多个文档中频繁出现的词汇组成,关键短语对主题的表征性意义非凡。本文提出的热点关键短语提取方法有两个阶段,即热点词提取和热点关键短语提取。该方法使用修改后的 TF * PDF 算法计算词汇与短语的权重,既考虑词汇在文档中出现位置的影响,也考虑词汇每次的变化。

2.1　热点词提取

热点词提取基于位置权重的 TF * PDF 模式。传统的TF * PDF 算法中,某个渠道词汇的权重与它在该渠道出现的频率呈线性比,与该渠道包含该词汇的文档比率呈指数比。词汇的总权重为该词汇在每个渠道的权重之和,如下所示。

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

$$|F_{jc}| = \frac{F_{jc}}{\sqrt{\sum_{k=1}^{K} F_{kc}^2}} \tag{1}$$

其中:w_j表示词汇 j 的权重;F_{jc}表示词汇 j 在渠道 c 出现的频率;n_{jc}表示词汇 j 所在渠道包含的文档数量;N_c表示渠道 c 中文档的总数量;k 表示一个渠道词汇的总数量;D 表示渠道的数量。

互联网上的新闻报道通常由标题与内容组成。标题比内容更具代表性,因此它基本可以反映出新闻报道的主题。在文档 i 中词汇 j 的位置权重定义为

$$ps_{ik}(j) = \begin{cases} 3 & j \in \text{the title of } k\text{th document in channel } i \\ 1 & j \notin \text{the title of } k\text{th document in channel } i \end{cases} \tag{2}$$

词汇总的位置权重为其在不同渠道的平均权重之和:

$$pw(j) = \sum_{i=1}^{|D|} \sum_{k=1}^{|N_c|} \frac{ps_{ik}(j)}{n_{jc}} \tag{3}$$

考虑每个词汇位置的影响,本文通过结合 TF * PDF 与位置权重提出了一种新的词汇权重模式。该算法则称做位置权重 TF * PDF。经位置权重 TF * PDF 修改的词汇 j 的权重定义为

$$\text{weight}_j = w_j \times pw(j) \tag{4}$$

根据修改后的权重模式对词汇进行排序,选择排名前 K 位的词汇构造候选热点词汇表。

一些词汇长期在新闻报道中频繁出现且分布不会发生变化,此类词汇称做绝对噪声[20]。TF * PDF 权重模式并不会考虑词汇在一个周期内不同阶段的变化,它更关注词汇出现的频率。因此绝对噪声通常具有很高的 TF * PDF 值。本文对候选热点词汇表中的词汇计算脉冲值。脉冲值在文献[13]中定义为

$$\text{bursty}_{t_0}(j)^2 = \frac{(ad-bc)^2}{(a+b)(a+c)(b+c)(b+d)}$$

$$\text{bscore}(j) = \arg_{t_0} \max \text{bursty}_{t_0}(j) \tag{5}$$

其中:$\text{bursty}_{t_0}(j)$表示单词 j 在时间 t_0 时的脉冲值;a 表示在 t_0 时刻包含单词 j 的文档数;b 表示在 t_0 时刻不包含单词 j 的文档数;c 表示在 t_0 时刻外包含单词 j 的文档数;d 表示在 t_0 时刻外不包含单词 j 的文档数;$\text{bscore}(j)$为单词 j 在整个时间内的脉冲值。如果词汇 j 的脉冲值小于阈值 $b_{\text{threshold}}$,则将其从候选词汇表中删除。

2.2　热点关键短语提取

热点关键短语提取首先要识别包含若干个热点词汇的短语。组合的短语形成候选关键短语集。短语的识别过程如下:

a)初始化关键短语候选集,将提取出来的关键词汇放入集合。

b)扫描每个文档的句子,如果某些热点关键词连续地在同一句话中出现,则将这些相邻的关键词汇组合成候选关键短语,并将该候选关键短语放入关键短语候选集。此时组合条件相对宽松。如果两个相邻的关键词之间存在一个非热点词汇,该短语仍将加入到关键短语候选集中。

c)候选关键短语有两个特征:关键短语应当在多个文档中频繁出现;更长的关键短语能够描述更多信息。根据这两个特征,首先找出重复出现次数最多的字符串,该字符串的频率值比任何包含 p 的字符串都大,然后利用文献[16]提到的共有信息计算 p 的显著值:

$$ss = \frac{TF(p)}{TF(x) + TF(y) - TF(p)} \tag{6}$$

其中:p、x、y 都是字符串,$p = c_1 \cdots c_n$,$x = c_1 \cdots c_{n-1}$,$y = c_2 \cdots c_n$;$TF(p)$是 p 出现的频率。如果 p 的显著值比阈值 $S_{\text{threshold}}$小,则将 p 从关键短语候选集中删除,否则将其子字符串从关键短语候选集中删除。

热点关键短语提取步骤 b)是从候选关键短语集中选出热点关键短语。短语的热度通过位置权重 TF * PDF 算法计算得来。通过对式(4)修改过权重的短语进行排序,排名前 K 的短语可识别为热点关键短语。

3　实验

3.1　数据集与数据预处理

实验在由真实网络环境下构造的数据集上执行。Crawlers

Hetrix 在 2011 年 4 月 5 日 ~2011 年 7 月 5 日间从两大流行门户网站 Sohu 和 Sina 上收集 5 462 条新闻网页。对于文本挖掘的每一个任务而言,第一步就是分词、停止词过滤及词干分析。本实验使用 ICTCLAS segmenter 来进行中文分词。

为确定参数值,先从数据集中随机选取 500 篇文档作为测试文档集。本文采用文献[13]中提到的方法,通过手动调整参数的方式在测试文档集上进行了实验。图 1 的测试结果显示了根据“动车”出现频率计算出的 x^2 值。文献[13]论证了 average link 聚类法可以产生较好的结果以及较好的权重容错性,故本文将最佳脉冲阈值设为 11.875。同理,该测试文档上的实验结果显示,$S_{\text{threshold}}$ 最佳阈值应当为 0.47。

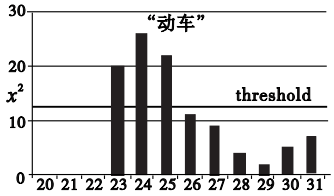

图1 由名词短语“动车”确定x^2及阈值

3.2 结果与分析

为评估本文提出的方法,进行了两次实验:

实验 1 对本文提出的位置权重 TF-PDF 和词汇周期相结合的热点词提取算法与使用传统的 TF-PDF 权重模式的热点词提取算法进行了对比。除此之外,本文亦对两段式关键短语提取方法与热点关键词提取方法进行了对比。

表 1 列出了两种方法得到的热点关键词。从表 1 可以看出,“人”“6 月”“http”“2011 年”都出现在了通过传统 TF-PDF 方法计算得到的前 20 的热点词汇中,但是并未出现在本文提出的方法计算结果中。这些词汇都有着同样的特征:在整个时间周期内有着固定的分布,在数据集中出现的频率异常地高。因此这些词汇都是绝对噪声。表 1 说明本文提出的方法可以去除绝对噪声。从表 1 中还可以发现,传统 TF-PDF 方法获得的 20 个热点关键词仅覆盖了三个话题,即“华南暴风雨”“京沪高铁”“台湾塑化剂事件”。而本文提出的方法获取的前 20 个热点词汇覆盖了除上述三个话题以外,还包含“CNOOC 原油泄漏”。这是因为本文的方法考虑了位置的影响,为标题中的词汇赋予更高的权重,而标题更能代表新闻报道的内容。

表 1 用本文方法和传统 TF * PDF 方法获得的热点关键词列表

序号	hot terms list		序号	hot terms list	
	traditional TF * PDF	our proposed method		traditional TF-PDF	our proposed method
1	高	铁	11	降雨	漏
2	铁	高	12	产品	等
3	油	油	13	强	事故
4	一	京沪	14	2011 年	水位
5	京沪	暴雨	15	站	发生
6	等	强	16	中	中海
7	大	塑化剂	17	水位	事件
8	人	降雨	18	列车	铁路
9	http	大	19	发生	列车
10	6 月	产品	20	市	受灾

表 2 列出了本文提出的算法获得的排名前 10 的热点词汇与热点关键短语。从表 2 可以发现,热点关键短语比热点词汇包含更多信息,关键短语比热点词汇更适合描述话题的某个方面。例如,通过热点关键短语“强降雨”“暴雨橙色预警”“多地受灾”等可以更多地了解关于“华南暴风雨”的主题。

表 2 本文算法获得的排名前 10 的热点词汇与热点关键短语

序号	our proposed method based on TF * PDF				
	hot term	hot key phrases		hot term	hot key phrases
1	铁	京沪高铁	6	强	灾害
2	油	暴雨橙色预警	7	塑化剂	漏油
3	高	强降雨	8	降雨	正式开通运营
4	京沪	塑化剂产品	9	大	台湾
5	暴雨	多地受灾	10	产品	暂停进口

实验 2 执行本文提出的热点关键短语提取方法及基于传统 TF * PDF 热点提取方法,并对两种方法的检测能力进行了对比。在这两种方法中,排名前 K 的热点关键短语用来表示发生在 2011 年 4 月 5 日 ~2011 年 7 月 5 日间的热点话题(实际发生在该期间的热点话题由审核员手工选取)。文献[21]中的覆盖率用于评估热点话题的检测能力。

$$\text{CoverageRate}(CR)=\frac{\text{Extracted Hot Topic}}{\text{Actual Hot Topic}} \tag{7}$$

从表 3 可以看出,本文提出的方法在检测热点话题领域比基于传统的 TF * PDF 方法具有更高的性能。

表 3 两种方法的热点话题检测能力对比

K	our proposed method/%	the extracted method based on traditional TF * PDF/%	K	our proposed method/%	the extracted method based on traditional TF * PDF/%
10	26.7	13.3	40	73.3	66.7
20	53.3	20	50	93.3	80
30	60	46.7			

4 结束语

本文创新性地提出了一种从互联网上发布的消息报道中提取热点关键短语的方法。该方法包括热点关键词提取与热点关键短语提取两个步骤,每个步骤都使用本文提出的位置权重 TF * PDF 模式对词汇与关键短语赋权重。本文最主要的贡献在于:

a)在热点关键词提取步骤,本文提出了新的词汇权重模式——位置权重 TF * PDF 算法,该模式将 TF-PDF 与位置权重相结合。TF * PDF 为在多个文档中出现频率高的词汇赋予更大的权重,且不考虑在不同阶段词汇的变化。为克服 TF * PDF 的缺陷,本文使用脉冲值滤除噪声。

b)在关键短语提取步骤,热点关键词首先根据相邻位置被合并成关键短语,通过最大重复字符串方法寻找候选关键短语,然后使用位置权重 TF * PDF 对候选短语的热度进行排名,排名前 K 的候选短语即热点短语。

通过在真实网络数据上进行实验,其结果表明,本文提出的方法能更有效地检测热点话题。在话题更新时对话题关键短语变化的分析、被同一个话题的关键短语所代表的短语关系的影响将是本文后续要做的工作。

参考文献:

[1] MYKHALOVSKIY E, WEIR L. The global public health intelligence network and early warning outbreak detection: a Canadian contribution to global public health [J]. Canadian Journal on Public Health, 2006, 97(1): 42-44.

[2] HE Qi, CHANG K, LIM E P. A model for anticipatory event detection [C]//Lecture Notes in Computer Science, Vol 4215. Berlin: Springer-

Verlag, 2006:168-181.

[3] YANG Yi-ming, PIERCE T, CARBONELL J. A study of retrospective and online event detection[C]//Proc of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1998: 28-36.

[4] BRANTS T, CHEN F, FARAHAT A. A system for new event detection[C]//Proc of the 26th Annual International ACM SIGIR Conferenceon Research and Development in Information Retrieval. New York: ACM Press, 2003: 330-337.

[5] ZHANG Kuo, ZI Juan, WU Li-gang. New event detection based on indexing-tree and named entity[C]//Proc of the 30th Annual International ACM SIGIR Conference on Research and Development inInformation Retrieval. New York: ACM Press, 2007: 215-222.

[6] ZHANG Zhong-feng, LI Qiu-dan. QuestionHolic: hot topic discovery and trend analysis in community question answering systems[J]. Expert Systems with Applications, 2011, 38(6): 6848-6855 .

[7] HE Qi, CHANG Kui-yu, LIM E P. Analyzing feature trajectories for event detection[C]//Proc of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2007: 207-214.

[8] ALLAN J, CARBONELL J, DODDINGTON G, *et al*. Topic detection and tracking pilot study: final report[C]//Proc of DARPA Broadcast News Transcription and Understanding Workshop. 1998: 194-218.

[9] 张小明,李舟军,巢文涵. 基于增量型聚类的自动话题检测研究[J]. 软件学报, 2012, 23(6): 1578-1587.

[10] WANG C, ZHANG M, MA S, *et al*. Automatic online news issue construction in Web environment[C]//Proc of the 17th International Conference on World Wide Web. 2008: 457-466.

[11] BUN K K, ISHIZUKA M. Topic extraction from news archive using TF * PDF algorithm[C]//Proc of the 3rd International Conference on Web Information Systems Engineering. [S. l.]: IEEE Press, 2002: 73-82.

[12] ZHANG K, XU H, TANG J. Keyword extraction using support vector machine[C]//Proc of WAIM. 2006: 85-96.

[13] SWAN R, ALLAN J . Automatic generation of overview timelines [C]//Proc of SIGIR. 2000: 49-56.

[14] TURNEY P. Learning algorithms for keyphrase extraction[J]. Information Retrieval, 2000, 2(4): 303-336.

[15] CHIEN L F. PAT-tree-based keyword extraction for Chinese information retrieval[C]// Proc of SIGIR. 1997: 50-58.

[16] YANG W , LI X. Chinese keyword extraction based on max-duplicated strings of the documents[C]//Proc of SIGIR. 2002: 439-440.

[17] WITTEN I, PAYYNTER G, FRANK E , *et al*. KEA: practical automatic keyphrase extraction[C]// Proc of the 4th ACM Conference on Digital Libraries. 1999: 254-255.

[18] HE Q, CHANG K, LIM E P. Using burstiness to improve clustering of topics in news streams[C]//Proc of the 7th IEEE International Conference on Data Mining. 2007: 493-498.

[19] FUNG G P C, YU J X, YU P S , *et al*. Parameter free bursty events detection in text streams[C]//Proc of the 31st VLDB Conference. 2005: 181-192.

[20] LI H X, ZHANG H P, QIN P. *et al*. Keywords based hot topic detection on Internet[C]//Proc of the 5th CCIR. 2009: 134-143.

[21] CHEN Kuan-yu, LUESUKPRASERT L, CHOU S C T. Hot topic extraction based on timeline analysis and multidimensional sentence modeling[J]. IEEE Trans on Knowledge and Data Engineering, 2007, 19(8): 1016-1025.

(上接第 3588 页)

模型,并构建模型验证规则,通过在推理机中进行推理,检验模型内容是否符合验证规则,以查找概念模型中不符合验证规则的内容,进而对概念模型进行修改和完善。该方法的实质是利用验证规则和推理机代替领域专家,在计算机上实现概念模型语义验证的自动化。该方法提高了验证效率,减少了专家验证的主观性和不确定性,降低了形式化验证方法的复杂性。本文只是给出了该方法的总体框架和具体实施流程,尚未进行实例验证,下一步将以军事概念模型验证实例对其进行检验。

参考文献:

[1] 樊浩,黄树彩. 基于 Petri 网的概念模型验证方法研究[J]. 计算机应用研究,2010,27(3):999-1002,1005.

[2] YANG Hui-zhen, HAO Li-li. Colored Petri nets-based formal modeling and validation for federation conceptual model[J]. Journal of System Simulation, 2012, 24(7): 1361-1365.

[3] 岳增坤,陈炜,夏学知,等. 基于 xUML 的 C4ISR 系统可执行对象模型设计[J]. 系统仿真学报,2009,21(8):2190-2194.

[4] 杨惠珍,康凤举. HLA 联邦形式化校核方法初探[J]. 系统仿真学报,2008,20(22):6039-6041.

[5] 夏薇,姚益平,慕晓冬. 基于 TLA 的事件图模型形式化验证方法[J]. 计算机应用研究,2011,28(11):4171-4173,4187.

[6] 何晓晔,徐培德,沙基昌. 任务空间概念模型轻量级形式化校核方法初探[J]. 系统仿真学报,2006,18(5):1108-1109.

[7] 陈丽娜,赵建民. 一种新的 Statechart 模型验证方法[J]. 计算机科学, 2011,38(2):144-148.

[8] 骆翔宇,苏开乐,顾明. 一种求解认知难题的模型检查方法[J]. 计算机学报,2010,33(3):406-414.

[9] JAMES C. XSL transformations (XSLT) version 2.0 [EB/OL]. (2002-05-30) [2009-10-10]. http://www.w3.org/TR/xslt2.0.

[10] EVERMANN J. A UML and OWL description of Bunge's upper-level ontology model[J]. Software and Systems Modeling, 2009, 8(2): 235-249.

[11] PARREIRAS F S, STAAB S. Using ontologies with UML class-based modeling: the two use approach[J]. Data and Knowledge Engineering, 2010, 69(11): 1194-1207.

[12] FRIEDMAN-HILL E. Jess in action: Java rule-based systems[M]. [S. l.]: Manning Publications Co, 2003.

[13] Sandia National Laboratories. Jess, the rule engine for the Java platform[EB/OL]. (2008-11-10) [2012-01-12]. http://Herzberg.ca.sandia. gov/jess.

[14] W3C. SWRL: a semantic Web rule language combining OWL and ruleML[EB/OL]. (2004-12-05) [2012-01-12]. http://www.daml.org/rul es/proposal.

[15] MA Li, YU He-long, WNAG Yue, *et al*. The knowledge representation and semantic reasoning realization of productivity grade based on ontology and SWRL[C]//Proc of the 5th International Conference on Computer and Computing Technologies in Agriculture. Beijing: China Agriculture University, 2011: 381-389.