# 基于改进的 $\chi^2$ 检验的热点词突发性度量研究*

翟东海[1,2] 聂洪玉[1] 崔静静[1] 于 磊[1] 杜 佳[2] 王佳君[2]
(1. 西南交通大学信息科学与技术学院 成都 610031)(2. 西藏大学工学院 拉萨 850000)

**摘 要** 采用原始 $\chi^2$ 检验公式进行突发性度量时存在低频词偏袒问题,论文提出了结合 TF 的改进的 $\chi^2$ 检验方法能有效克服该问题。该方法将词频累加和作为文档统计篇数的影响因子 $\beta$ 引入原始 $\chi^2$ 检验公式从而解决了低频词偏袒问题,提高了度量热点词突发性的精确度。动态突发性热点词库依据改进后的 $\chi^2$ 检验公式得到的突发性度量值来建立,并将该词库运用在动态突发性向量空间模型中来发现与追踪网络突发性热点话题。实例验证表明,利用该文的方法进行话题发现与追踪,可以获得有更高的准确率、召回率以及 $F$ 度量。

**关键词** 突发性热点词; $\chi^2$ 检验; 词频; 动态突发性词库
**中图分类号** TP339.4 **DOI**:10.3969/j.issn1672-9722.2013.11.023

# Bursty Measurement of Hot Term Based on Improvement $\chi^2$ Test Combined with TF

ZHAI Donghai[1,2] NIE Hongyu[1] CUI Jingjing[1] YU Lei[1] DU Jia[2] WANG Jiajun[2]
(1. School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031)
(2. Engineering School, Tibet University, Lhasa 850000)

**Abstract** Original $\chi^2$ test formula favors low frequency words when it measures bursty of hot term. To overcome this problem, the improved $\chi^2$ test formula combined with TF is proposed. In this approach, the term frequency summary, an impact factor $\beta$ to the document statistics, is introduced into the original $\chi^2$ test formula. The experimental results show the dynamic bursty vector space model achieved higher precision, recall and F-measure in online bursty topic detection and tracking, when dynamic bursty lexicon is constructed according to the bursty measurement using the improved $\chi^2$ test.

**Key Words** bursty of hot term, $\chi^2$ test formula, term frequency, dynamic bursty lexicon
**Class Number** TP339.4

## 1 引言

随着网络的普及,网络搜索逐渐成为人们生活中获取信息的主要渠道,而网络搜索特别是网络突发性热点话题的搜索主要是通过突发性热点词来获取相关信息。突发性热点词是指一定时间、一定范围内,公众最为关心的热点词语,它具有当前未知、未来发生、集中时间段存在的特点。同时,突发性热点词也是发现与追踪网络突发性热点话题的启发性知识,因此,对于网络中大规模的文本流,要有效地实现网络突发性热点话题的发现与追踪,就必须从中获取突发性热点词并建立动态突发性词库[1]。

建立动态突发性热点词库,首先要对文本流中的词进行突发性度量,其度量值是从网络文本流中发现并提取出突发性热点词的依据。Fung 等根据词语在时间上的分布来度量词语的突发性[2~3]。Wang 等[4]和 Suba 等[5]通过在文本流中挖掘突发性模式来发现突发性热点词。Xue 等根据 $\chi^2$ 检验公式计算出每个词语的 $\chi^2$ 值即突发值,基于突发值来进行突发性热点词选择[1]。然而,$\chi^2$ 检验公式并不是很完美,因为基于 $\chi^2$ 检验公式来进行突发性热点词选择时只统计了文档中是否出现词语 $w$,忽略其在文档中的出现次数,这种方法容易导致低频词的 $\chi^2$ 值偏大从而误判突发性热点词。本文针对 $\chi^2$ 检验公式对处理低频词时存在的偏袒问题,以网络文本流为研究对象,提出了一种结合 TF 的改进的方法。

## 2 $\chi^2$ 检验方法对低频词的偏袒问题

为了更好地表示热点词 $w$ 与时间段 $t$ 的关系,将时间段 $t$ 与 $t$ 之前(记作 $\bar{t}$)表示为两个特征属性,同时,将包含词语 $w$ 和不包含词语 $w$(记作 $\bar{w}$)表示为另两个特征属性。考虑在时间段 $t$ 及 $t$ 之前一段时间内出现的所有文档数,将在时间段 $t$ 内出现并包含词语 $w$ 的文档数用 $A$ 表示,在时间段 $t$ 之前出现并包含词语 $w$ 的文档数用 $B$ 表示;在时间段 $t$

* **收稿日期**:2013 年 5 月 11 日,**修回日期**:2013 年 6 月 27 日
**基金项目**:国家语委"十二五"科研规划项目(编号:YB125-49);教育部科学技术研究重点项目(编号:212167);中央高校基本科研业务费专项资金科技创新项目(编号:SWJTU12CX096);国家级大学生创新创业训练计划项目(编号:201210694017)资助。
**作者简介**:翟东海,男,博士,副教授,研究方向:海量数据挖掘、数字图像处理。聂洪玉,女,硕士研究生,研究方向:海量数据挖掘。崔静静,女,硕士研究生,研究方向:海量数据挖掘。于磊,男,硕士研究生,研究方向:海量数据挖掘。杜佳,男,研究方向:海量数据挖掘。王佳君,男,研究方向:海量数据挖掘。

万方数据

内出现但不包含词语 $w$ 的文档数用 $C$ 表示，在时间段 $t$ 之前出现且不包含词语 $w$ 的文档数用 $D$ 表示。那么，对于要研究的词语 $w$ 与时间段 $t$ 之间的关系可以用表 1 所示的四格表来表示[1,6]。

**表 1　词语 $w$ 与时间段 $t$ 之间的关系(四格表)**

| | $w$ | $\bar{w}$ |
|---|---|---|
| $t$ | $A$ | $C$ |
| $\bar{t}$ | $B$ | $D$ |

根据 $\chi^2$ 检验公式，词语 $w$ 在时间段 $t$ 内的突发值由以下公式计算得到：

$$\chi^2_{w,t}=\frac{(A+B+C+D)(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)} \tag{1}$$

最后选取突发值大的词语 $w$ 作为时间段 $t$ 内的突发性词语[5~6]。

通过研究文档统计数 $A$ 后发现其大小对 $\chi^2$ 值有很大的影响。这是由于在统计过程中只统计了文档中是否出现词语 $w$ 而忽略了其词频的做法会导致特征词的 $\chi^2$ 值有很大的误差，从而引起突发性热点词的误判。例如，在时间段 $t$ 内有 100,000 篇文档，词语 $w_1$ 在 10,000 篇文档中各出现了 1 次，词语 $w_2$ 在 9,999 篇文档中各出现了 10 次以上，而文档统计结果 $A_1>A_2$，在进行突发性热点词判断的时候，很可能会选择前者而抛弃后者。但显然在时间段 $t$ 内词语 $w_2$ 的突发性应该大于 $w_1$。

因此，针对 $\chi^2$ 检验公式对低频词的偏袒问题，本文提出一种结合 TF 方法的改进 $\chi^2$ 检验公式来进行词语的突发性度量，实例验证结果表明，本文提出的改进方法能够有效改善突发性度量的精度。

## 3　结合 TF 的改进的 $\chi^2$ 检验方法

目前，用 $\chi^2$ 检验公式进行突发性度量存在低频词偏袒问题，它的根源在于没有充分考虑词频对突发性的影响。如果把热点词的词频累加和作为文档数 $A$ 的影响因子 $\beta$ 代入 $\chi^2$ 检验公式进行突发性度量则可以有效克服低频词偏袒问题，而 TF-IDF 算法中词频统计公式 TF 是目前最简便有效的词频统计方法，因此，本文提出的方法是一种结合 TF 的改进的 $\chi^2$ 检验公式的方法，其主要流程包括四个方面(流程图如图 1 所示)。

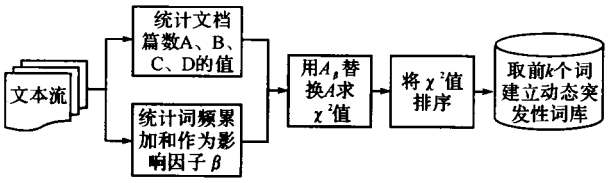


**图 1　改进后方法的流程图**

1) 对于进入系统的文本流并行的进行两方面的工作：(1)四格表中文档篇数 $A$、$B$、$C$、$D$ 值的统计，(2)统计 $A$ 类文档中词频值的总和作为影响因子 $\beta$。

2) 用 $A+\beta$ 代替 $A$ 的值(记为 $A_\beta$)并代入 $\chi^2$ 检验公式计算热点词的突发性。

3) 根据各词的突发性由大到小进行排序得到一个序列。

4) 取该序列的前 $k$ 个词加入动态突发性词库。

在本文提出的改进方法中，先用 TF-IDF 算法中的 TF 公式计算 $A$ 篇文档中的各个词的词频值 $tf_w$，并将它们的累加和作为影响因子 $\beta$，其计算公式分别如下：

$$tf_{w,j}=\frac{n_{w,j}}{\sum_i n_{i,j}} \tag{2}$$

$$\beta=\sum_j tf_{w,j} \tag{3}$$

$tf_{w,j}$ 表示词语 $w$ 在第 $j$ 篇文档中的词频；$\beta$ 表示影响因子，其值为 $A$ 篇文档的词频总和。针对原始的 $\chi^2$ 检验公式中文档篇数 $A$ 在统计过程中不考虑词频会引起对低频词的偏袒问题，本文提出的算法在进行文档数统计过程中考虑到词频 $tf_w$ 的影响将每篇文档的统计个数由 1 修正为 $(1+tf_w)$，相应的文档篇数 $A$ 修正为 $A_\beta$，其计算公式如下：

$$A_\beta=\sum_j(1+tf_{w,j})=A+\sum_j tf_{w,j}=A+\beta \tag{4}$$

由于引入词频累加和作为影响因子 $\beta$，需对原始四格表作如下修改(如表 2 所示)。

**表 2　改进后的四格表**

| | $w$ | $\bar{w}$ |
|---|---|---|
| $t$ | $A_\beta$ | $C$ |
| $\bar{t}$ | $B$ | $D$ |

在求得修正后的 $A_\beta$ 后，本文中用到的改进的 $\chi^2$ 公式如下：

$$\chi^2_{w,t,\beta}=\frac{(A_\beta+B+C+D)(A_\beta D-BC)^2}{(A_\beta+C)(A_\beta+B)(B+D)(C+D)} \tag{5}$$

改进后的公式中，词语 $w$ 在文档中出现的次数越少则 $A_\beta$ 的值越小，反之，$A_\beta$ 的值随之增加，即达到了克服低词频偏袒问题的效果。

## 4　实验与分析

采用本文提出的结合 TF 改进后的 $\chi^2$ 检验公式来进行热点词语的突发性度量，将该度量值作为参数代入动态突发性向量空间模型进行热点话题发现与追踪[5]。采用文本聚类的基本评测指标精确率 $P$、召回率 $R$ 和 $F$ 度量来检测其效果，其中，实验数据是从网络上获取的与“云南彝良地震”有关的 1300 篇文档。

数据中有 $k$ 个话题 $T_1,T_2,\cdots,T_k$ 通过话题发现方法[4]得出 $m$ 个类 $C_1,C_2,\cdots,C_m$，被聚类到类 $C_i$ 属于话题 $T_j$ 的文本数记为 $n_{ij}$(如表 3 所示)。

**表 3　话题类别表**

| | $T_1$ | $T_2$ | … | $T_j$ | … | $T_k$ |
|---|---|---|---|---|---|---|
| $C_1$ | $n_{11}$ | $n_{12}$ | … | $n_{1j}$ | … | $n_{1k}$ |
| $C_2$ | $n_{21}$ | $n_{22}$ | … | $n_{2j}$ | … | $n_{2k}$ |
| $C_i$ | $n_{i1}$ | $n_{i2}$ | … | $n_{ij}$ | … | $n_{ik}$ |
| … | … | … | … | … | … | … |
| $C_m$ | $n_{m1}$ | $n_{m2}$ | … | $n_{mj}$ | … | $n_{mk}$ |

关于话题 $j$ 和类 $i$ 的准确率和召回率分别定义为

$$P(i,j)=\frac{n_{ij}}{n_i} \tag{6}$$

$$R(i,j)=\frac{n_{ij}}{n_j} \tag{7}$$

其中，$n_i=\sum_{j=1}^{k} n_{ij}$，$n_j=\sum_{i=1}^{m} n_{ij}$。那么，类 $i$ 和话题 $j$ 的 $F$ 度量由以下公式计算

万方数据

$$F(i,j)=\frac{2P(i,j)R(i,j)}{P(i,j)+R(i,j)} \tag{8}$$

要计算数据集的准确率 $P$、召回率 $R$ 及 $F$ 度量,对数据中的所有话题的 $P(i,j)$、$R(i,j)$及 $F(i,j)$求平均值。即

$$P=\sum_j \frac{n_j}{n}P_i \tag{9}$$

$$R=\sum_j \frac{n_j}{n}R_i \tag{10}$$

$$F=\sum_i \frac{n_j}{n}F_i \tag{11}$$

为了更好地表示文本与时间之间的关系,将获取的 1300 篇文档用时间-文本坐标轴表示,如图 2 所示。

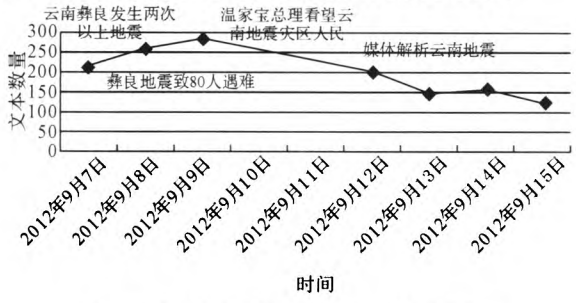


**图 2 云南彝良地震事件的文本-时间分布图**

从图 2 中可以看出,有关云南地震的网络文本主要分为四个阶段,将各个阶段的文本数量按照改进后算法的流程图(如图 1)计算出各突发词语的突发值,把建立的动态突发性词库运用到动态突发性向量空间模型中对热点话题进行发现和跟踪得到的准确度、召回率、$F$ 度量与改进前进行比较,结果如图 3 所示。

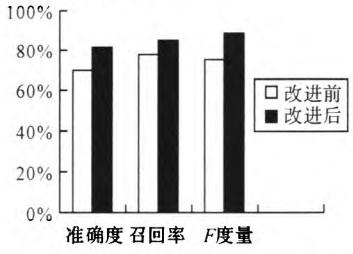


**图 3 改进前后的 $\chi^2$ 检验公式对突发性话题发现的效果比较**

通过实验分析可以发现经过改进的 $\chi^2$ 检验公式能更精确度量突发值,从而更好地对话题进行发现和跟踪,其召回率、准确率和 $F$ 度量都明显优于改进之前。

## 5 结语

本文针对 $\chi^2$ 检验公式进行突发性度量存在的低频词偏袒问题,将词频作为计算文档篇数的影响因子 $\beta$ 从而将文档篇数 $A$ 修正为 $A_\beta$,在此基础上提出了改进的 $\chi^2$ 检验公式的方法,该方法能够更准确地度量热点词的突发性。实例验证表明,由改进后的 $\chi^2$ 检验公式得到的度量值作为建立动态突发性热点词库的参数,将在此基础上构建的动态突发性向量空间模型运用在网络突发性热点话题发现与追踪上有更高的准确率、召回率以及 $F$ 度量。

**参 考 文 献**

[1] 薛峰,周亚东. 一种突发性热点话题在线发现与跟踪方法[J]. 西安交通大学学报,2011,12(45):64-70.
XUE Feng, ZHOU Yadong. An Online Detection and Tracking Method for Bursty Topics[J]. Journal of Software,2011,45(12):64-70.

[2] FUNG G P C, YU J X, LIU H C. Time-dependent event hierarchy construction[C]//BERKHIN P, eds. Proceedings of the 21st Annual International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM,2007:300-309.

[3] FUNG G P C, YU J X, YU PS, et al. Parameter free bursty events detection in text streams[C]//Proceedings of the 31st International Conference on Very Large Data Bases. Trondheim, Norway: VLDB Endowment,2005:181-192.

[4] WANG X H, ZHAI C X, HU X, et al. Mining correlated bursty topic patterns from coordinated text streams[C]//BERKHIN P, et al. Proceedings of the Thirteenth ACM International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM,2007:784-793

[5] SUBA I. From bursty patterns to bursty facts: the effectiveness of temporal text mining for news[C]//Proceedings of 19th Artificial Intelligence. Fairfax, VA, IOS Press,2010:517-522.

[6] SWAN R, ALLAN J. Automatic generation of overview timelines[C]//Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM,2000:49-56.

(上接第 1735 页)

[7] Martin T H, Howard B D, Mark H Beale. 神经网络设计[M]. 北京:机械工业出版社,2002:239-243.
Martin T H, Howard B D, Mark H Beale. Neural Network Design[M]. Beijing: China Machine Press,2002:239-243.

[8] 金保明. LMBP 算法在闽江上游十里庵站洪水流量预测中的应用[J]. 福州大学学报(自然科学版),2012,40(4):527-530.
JIN Baoming. Application of LMBP Arithmetic in Flood Flow Forecast of Shili'an Station in the Upper Minjiang River[J]. Journal of Fuzhou University(Natural Science Edition),2012,40(4):527-530.

[9] 吕京虎,陆君安,陈士华. 混沌时间序列分析及其应用[M]. 武汉:武汉大学出版社,2002:58-60.
LV Jinghu, LU Jun'an, CHEN Shihua. The Analysis of Chaotic Times Series and Its Application[M]. Wuhan: Wuhan University Press,2002:58-60.

[10] 周开刊,康耀红. 神经网络模型及其 MATLAB 仿真程序设计[M]. 北京:清华大学出版社,2002:82-83.
ZHOU Kaikan, KANG Yaohong. Netrual Network Model and Its MATLAB Simulation Programming[M]. Beijing: Tsinghua University Press,2002:82-83.

万方数据