ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*, Vol.21, Supplement, December 2010, pp.363−372

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

# 面向互联网新闻的在线事件检测*

付 艳+, 周明全, 王学松 , 栾 华

(北京师范大学 信息科学技术学院,北京 100875)

# On-Line Event Detection from Web News Stream

FU Yan+, ZHOU Ming-Quan, WANG Xue-Song, LUAN Hua

(College of Information Science and Technology, Beijing Normal University, Beijing 100875, China)

+ Corresponding author: E-mail: fuyan@bnu.edu.cn

**Fu Y, Zhou MQ, Wang XS, Luan H. On-Line event detection from Web news stream**. ***Journal of Software***, **2010,21(Suppl.):363−372.** http://www.jos.org.cn/1000-9825/10037.htm

**Abstract**: In order to improve the efficiency of event detection from on-line news stream, we propose a new method to accomplish detection task with window-adding, named entity recognition and suffix tree clustering. In our method, we make full use of informative elements extracted from news (such as date, place, person and so on) to help detection task, and accomplish the detection efficiently with news characteristics matching, which decreases text similarity computation greatly. Experimental results show that our method improves on-line event detection performance, without sacrificing detection precision.

**Key words**: on-line event detection; window-adding strategy; named entity recognition; news characteristic; suffix-tree clustering

**摘 要**: 为了提高互联网上新闻事件在线检测的效率,利用加窗策略、命名实体识别及后缀树聚类等技术提出了一种新的检测算法.该算法基于实体识别技术解析出新闻数据特有的信息元素(例如日期、地点、人物等),并在限定的时间窗口内,通过新闻特征的语义匹配实现了新事件的快速识别,从而大幅降低了基于文本相似度计算的检测算法带来的巨大时间消耗.实验结果证明,该算法能够实现在保障检测准确率的同时显著提高检测的效率.

**关键词**: 在线事件检测;加窗策略;命名实体识别;新闻特征;后缀树聚类

互联网的出现极大程度上满足了人们对数据的需求,同时也为我们提供了一种全新的新闻获取方式,更好地满足了日常生活中人们对新闻的浏览需求.在复杂的网络环境中,大量的、分散在世界各地的互联网站点每天都在以异步的方式在互联网上发布新闻.众多信息源对新闻报道的异步发布导致互联网新闻呈现显著的数据流特征,即连续、有序、变化、迅速及海量等.这些因素使得互联网上的事件检测研究愈加复杂并充满挑战.近年来,越来越多的研究者致力于面向互联网新闻流数据进行数据挖掘研究.本文即针对互联网数据流挖掘中的在线事件检测问题展开探讨.

面向互联网新闻的在线事件检测以持续到达的互联网新闻报道流为研究对象,实时地判断新发布报道与历史报道间是否存在内容上的关联,检测每一篇新到达的报道是否涉及新发生的事件,以实时地在某事件发生

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2008AA01Z301 (国家高技术研究发展计划(863))

Received 2010-07-01; Accepted 2010-12-10

时识别出该事件.利用在线事件检测技术,互联网用户可以免于被大量的无序新闻所湮没,同时也使用户能够更快捷地了解到短期内发生的重大事件.然而,由于互联网上的新闻报道往往在很短的时间间隔内呈现井喷式的增长,这使得准确、高效的在线事件检测研究非常具有挑战性.

目前,常用的在线事件检测研究方法主要包括增量聚类[1−3]、加窗策略[2,3]等.传统的在线检测算法大多是利用文本相似度计算实现的,当处理的新闻数量以百万计时,文本相似度计算的时间消耗将是无法估计的,因此现有方法的检测效率并不理想.James Allan 曾提出,完全依赖于文本相似度计算的检测算法性能是有限的,仅通过参数的调整无法帮助检测算法突破这一性能瓶颈[4].

在本文中我们提出了一种新的检测算法,该算法避免了大部分报道的文本相似度计算,因而获得了检测效率的显著提高.为提高在有限范围内完成检测的可能性,算法首先对事件进行加窗,之后利用命名实体识别技术从新闻中解析出重要的新闻要素(例如人物、地点等),以通过新闻要素的匹配高效地实现部分报道内容相关性的判断,其他未判断成功的新闻可利用后缀树聚类算法完成检测.

本文在第 1 节中对在线事件检测问题进行了详细的分析,第 2 节是新检测算法的详细介绍,第 3 节是实验结果报告及效果评估,第 4 节总结了现有的相关研究成果,在第 5 节对本文的研究内容及未来工作进行了小结.

## 1 问题分析

首先,我们对本文中讨论的概念给出具体的定义.

**定义 1(新闻报道(Story))**.

表示与某个特定主题紧密相关的新闻片段,通常由两个或多个报道同一事件的独立语句构成.

历史报道是在某个特定时间点之前发布的报道.在历史报道集合中,每一篇新闻报道均与某一特定的事件相关.

**定义 2(事件(Event))**.

表示某时间段内发生的特定事情.某一事件通常由多个围绕该事件的新闻报道组成、描述.

历史事件集合是指在某特定时间点之前,所有已识别出的事件.通常情况下,每一个事件均有多篇历史报道与其紧密相关,可认为是该事件的相关报道.

新闻报道与事件间的关系及报道顺序发布的过程如图 1 所示.

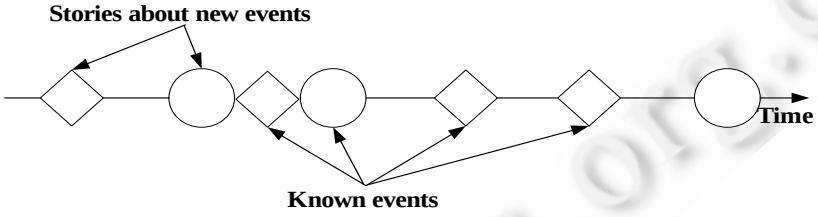



Fig.1 Story and Event

图 1 新闻流中的报道−事件示意图

在图 1 中时间轴上的每一个图形代表一篇新闻报道,不同的形状(如圆形或菱形)分别代表不同的事件.在本文中,假设关于某事件的报道是依时间顺序、串行发布的,报道之间不存在时间上的重叠关系.在线事件检测的任务是在第一时间检测并标识出新闻流中未出现过的、第一篇关于某事件的报道.

假设 $S=\{s_1,s_2,\ldots\}$表示互联网上的新闻数据流,其中,每篇报道 $s_i$ 依发布时间顺序到达.假设 $E=\{e_1,e_2,\ldots,e_m\}$ 表示已知的历史事件集合.其中,若干篇新闻报道可同时被包含在同一个事件 $e_j$ 中.

在线事件检测研究关注连续的互联网新闻,判断新到达或发布的新闻报道是否涉及新发生的事件.检测过程根据每一篇新闻报道到达的时间顺序,依次对其进行分析以判断是否谈及新事件,即在第一时间检测、标识出之前并未出现过的、第一次关于某事件的报道.若判断该报道是关于新事件的第一篇报道,需将其标识为“NEW”并实时反映给用户.如果该新闻报道并不涉及新发生的事件,则将其标识为“OLD”.

因此,本文将在线事件检测任务描述为

$$Label(s_i)=\begin{cases} OLD, \exists(e_j \in E \wedge Sim(s_i,e_j)>sim) \\ NEW, \neg\exists(e_j \in E \wedge Sim(s_i,e_j)>sim) \end{cases}.$$

其中,$Label(s_i)$表示对新闻报道 $s_i$ 的事件标识.$sim$ 是预设的相似度阈值.$Label(s_i)=OLD$ 表示与 $s_i$ 内容相似或相关的内容已出现过,$s_i$ 并非关于某一事件的首篇新闻报道.$Label(s_i)=NEW$ 表示 $s_i$ 的描述内容是新出现的事件.

$Sim(s_i,e_j)$表示新闻报道 $s_i$ 与事件 $e_j$ 间的相似度,计算方式如下:

$$Sim(s_i,e_j)=\max Sim(s_i,s_j'),(s_j' \in e_j).$$

通常,围绕着同一新闻事件往往会出现多篇内容非常相似的报道,但其中只有一篇可称为“第一篇报道”.互联网上新闻报道流的连续、有序、变化、迅速及海量特征,对新闻流中“第一篇报道”的检测研究提出了很高的性能要求.因而,如何在短时间内迅速、准确地判断出报道内容是否首次出现是本文研究的关键问题.

## 2 在线事件检测算法

新闻报道是关于发生在特定时间、特定地点事件的详细报告.与互联网上其他普通的文档不同,互联网新闻仅在有限的时间范围内有效,具有很强的时效性.某一事件发生后,互联网上会涌现大量来自不同信息源的相似报道,但经过一段时间后,关于该事件的报道就会越来越少,直至消失.由此可知,发布时间过久的新闻报道对事件检测工作并无帮助,更有很大可能造成计算开销的浪费及检测误差.

此外,有些新闻内容具有明显的周期性特征,会不断地重复发生,例如美国总统选举及奥林匹克运动会等.如果事件检测单纯依赖于文本相似度的计算,不考虑新闻报道自身的时间特性,很容易导致关于新近事件的报道被湮没在大量的历史事件中,人们很难在第一时间了解到新事件的发生,同时,大量的历史新闻报道也会导致检测速度过慢、效率过低的情况.基于上述考虑,本文提出了一种新的在线事件检测方法,充分利用新闻报道的时效性特征,在限定的时间窗口内利用报道的时间、地点、人物等新闻特征来提高检测效率.该方法包含三个步骤,一是事件的加窗及更新策略,二是基于实体识别技术的新闻特征获取,三是基于实体及增量聚类的在线事件检测.

### 2.1 事件加窗及更新策略

在对互联网上的新闻报道流进行事件检测时,如果不考虑新闻报道的时效性,那么新闻流的海量、动态及变化等特征很可能造成检测范围过大、检测效率低、检测出现错误的情况.例如,如果两个地区在相隔几天的时间中相继发生爆炸,若不考虑事件的发生时间,单纯依赖新闻内容的文本相似度计算来进行事件检测,那么关于这两个不同事件的报道极容易被混淆,进而导致对新事件判断的失误,也就是说新发生的事件会被历史报道湮没.

在实际应用环境中,我们发现互联网上的新闻具有很强的时效性,这一特性体现在一是关于某事件的报道在持续一段时间后,就不再有新报道出现了;二是许多事件可能会重复发生.为降低计算的复杂性并提高检测的准确度,本文考虑将检测范围限定在一个较小的时间间隔内.

我们首先引入新闻的生存周期对事件检测的报道集范围进行限定,以突出新闻报道的时效性及缩小检测的范围.为统计每篇新闻报道 $s_i$ 的生存周期 $life\_time_i$,我们利用每篇报道 $s_i$ 发布的时间(从新闻网页或 URL 解析出的发表时间)作为报道 $s_i$ 的时间戳.有研究表明,将新闻生存周期设置为 24~32 天时,检测的效果明显优于其他情况[5].也就是说,发布时间超过 24 天的新闻报道,其新闻价值已大大减弱.因此,本文将 $life\_time_i$ 的阈值设置为 24 天,当 $life\_time_i$>24 时,即将报道 $s_i$ 视为过期报道,并从历史报道、历史事件集合中删除该报道.

基于互联网上新闻流的时序特征,本文为历史事件加窗.窗口中的信息放置在内存中,窗口以外的历史信息可存储在磁盘上.根据事件发生的时间顺序排列,窗口中存储的是最近发生的若干个事件,假设窗口大小为 $window\_size$.时间窗口 $window\_size$ 的大小是根据所处理的新闻报道集规模确定的.本文中将窗口大小设置为 500,即窗口内可以存储新近发生的 500 个事件.

在本文中,我们利用事件中包含的一篇新闻报道来表示该事件.假设当前时刻为 $t$,最近发布的新闻报道为 $s_t$,若 $s_t \in e_i$,则我们使用 $s_t$ 表示事件 $e_i$.随着某一具体事件的发展,相关新闻报道的内容会发生不同程度的迁移.本文中,我们假设内容相关的传递性成立,即若报道 A 与报道 B 相关,报道 B 与报道 C 相关,那么报道 A 与报道 C 也相关.根据内容相关的传递性假设,本文利用每个事件的最新报道来表示该事件,可以在事件内容发生少量迁移的情况下尽可能地保障检测的准确性.因此,窗口中存储的实际上是表示多个不同事件的报道集合,报道集合的大小为 *window_size*.

初始情况下,只有小部分新发布报道被存储在窗口中,此时,报道集的规模小于 *window_size*.随着报道的陆续发布,当新闻流上出现新报道时,我们首先在时间窗口中寻找可能与新报道相关的历史事件.如果没有在当前窗口中找到与新报道相关的内容,再对窗口外存储的历史事件集进行搜索.检测过程中,如果新到达报道 $s_{new}$ 与当前窗口中的某一报道 $s_{old}$ 相关,即表示其与某历史事件 $e_{sold}$ 内容相关,则利用新报道 $s_{new}$ 替换窗口中描述同一事件的历史报道 $s_{old}$,并将 $s_{old}$ 移出窗口;如果新报道 $s_{new}$ 与当前窗口中的所有报道并不相关,即根据所有历史报道的时间戳,将生命周期 $life\_time_i$ 最长、发布最早的报道(即出现最早的事件)从当前窗口移出,将新到达报道 $s_{new}$ 作为新事件 $e_{snew}$ 的种子添加至窗口内的事件集合中.在对历史报道集合及历史事件集合信息进行更新时,需要根据上述步骤对窗口内及窗口外的报道集进行同步更新.

对时间窗口的更新策略,避免了大量相似报道占据窗口空间,造成窗口中事件覆盖范围过小的情况,提高了新到达报道命中当前窗口内容的概率.由于窗口内的报道是放置在内存中的,因此如果在窗口中即可完成检测,则不需要频繁访问磁盘,会使检测效率得到显著提高.

### 2.2 命名实体识别

对于新闻报道而言,报道内容中涉及的时间、地点及人物对事件本身有很好的标识及区分作用.一旦报道的时间、地点、人物等特征确定下来,人们即可迅速确定事件的相关情况,并将其与别的事件区分开来.这些新闻特征正是命名实体,因此,本文引入实体识别技术将这些信息解析出来,期望通过对新闻特征的语义匹配实现事件的高效检测.

命名实体是文本中基本的信息元素,也是正确理解文本的基础.通常,命名实体包括 7 种类别,即:人物、组织、地点、日期、时间、货币、比率.实体识别技术是用来判断一个文本串是否代表命名实体,并确定其类别的技术.本文利用实体识别技术实现日期、地点、人物及组织的新闻特征解析.日期实体的引用形式通常较规则,可采用基于规则的实体识别完成,这一识别过程与利用正则表达式的字符串匹配类似.其他类型的实体,本文采用 ICTCLAS(Institute of Computing Technology Chinese Lexical Analysis System)系统(http://www.nlp.org.cn)完成识别.

在实际应用中我们发现,一篇新闻报道往往包含多个相同及不同类型的命名实体,例如一篇报道中往往会出现多个不同的日期或提及若干不同的人名等.此外,即使是同一个实体也可能在一篇报道中反复出现,如某个人名或地名被多次提及.通常情况下,实体识别过程会从新闻报道中解析出若干个同一类型的新闻特征.然而,在所有相同类型的命名实体中,新闻特征在其重要性及它们对报道内容的表达能力上是存在巨大差异的,也就是说这些特征对报道内容来说并不是同等重要的.以一篇具体的新闻报道为研究对象时,我们显然希望选择出对报道内容具明确限定性的新闻特征来描述该报道.总体而言,新闻特征的选择策略主要基于特征对报道内容的表达及区分能力,对新闻报道内容具有较高标识性及区分能力的实体,其重要性也应该越高,在事件检测中也扮演着越重要的角色.

本文对每一个新闻特征即命名实体的重要性进行评估、衡量,衡量的依据就是其对报道内容的区分或标识能力.在计算权重的过程中,主要考虑两个因素:实体首次出现的位置及其出现的频率.与其他同一类型的实体相比,一般来说出现频率越高,首次出现位置越靠前的实体,其重要性应该越高,权重值也应该越高.

假设 $d$ 是出现在报道 $s$ 中一个具体的日期实体(如 2009 年 12 月 7 日),则其权值计算方法如下:

$$w(d,s)=\log\left[\frac{(N+C_d)\times(N-L_d)}{N^2}\right].$$

其中,$C_d$ 是 $d$ 在报道 $s$ 中出现的次数,$N$ 表示出现在报道 $s$ 中的日期实体个数.将报道 $s$ 中的所有日期实体依照其在报道中出现的先后顺序排成一个有序序列,$L_d$ 即表示 $d$ 在该序列中首次出现的位置序号.

对地点、人物及组织等实体的权值计算方法与日期一致.若某篇报道中人物实体的出现次数 $C_p$ 为 0,则对该报道中的组织实体进行解析及权重计算.若某篇报道中组织实体的出现次数 $C_o$ 为 0,则对该报道中的人物实体进行解析及权重计算.在本文的算法中,不能同时缺少人物及组织实体,两类实体中至少要有一类作为新闻报道的主语出现.

根据日期、地点、人物或组织等新闻特征的加权排序结果,我们选择权值最高的新闻特征来共同表示一篇新闻报道.在本文的实验部分,我们还设计实现了一种 NE_STC 算法,该算法在实体识别部分并不计算、区分实体间的权重差别,仅利用一篇新闻报道同类型实体中的第一个出现的对象来表示报道的内容.在实验部分,我们对区分实体权重及未区分权重的方法进行了比较.

### 2.3 基于新闻特征语义匹配的在线检测

时效性使得第一篇报道新事件的新闻显得尤为重要,因此,如何迅速判断出某报道的内容是否是第一次出现成为在线检测研究的关键问题.基于语义学知识及新闻属性,如果能够利用新闻特征迅速发现与新到达报道相关的事件,即可避免对文本相似度的复杂计算.

假设 *NewsCharacteristic* 表示由日期实体 $D$、地点实体 $Pl$ 及人物实体 $Pe$(或组织机构实体,解析出二者均可,均代表新闻内容的主体)构成的集合{$D$,$Pl$,$Pe$},那么 *NewsCharacteristic* 的子集包括:{$D$,$Pl$,$Pe$},{$D$,$Pl$},{$D$,$Pe$},{$Pl$,$Pe$},{$D$},{$Pl$},{$Pe$},{}.

根据语义分析,仅包含一个实体或不包含任何实体的空集对报道内容没有明确的限定性,因此我们并不考虑后面四个子集.对其他 4 个子集的分析如下:

(1) {$D$,$Pl$,$Pe$}

根据新闻报道的特征,发生在同一日期、同一地点,并涉及相同人物或组织的新闻事件是确定的.考虑到报道间可能出现的微小差别,这样的报道至少是紧密关联的.因此,如果新到达报道与某历史报道具有相同的 $D$, $Pl$ 及 $Pe$ 特征,即可断定新报道的内容并非第一次出现.

(2) {$D$,$Pl$}

从新闻学的角度进行判断,包含同一日期、同一地点的新闻并不一定是对同一事件的报道.因此利用 $D$ 及 $Pl$ 的组合无法判定报道间的关系.但是,当两篇报道的内容非常相似,并且具有相同的日期、地点要素时,即可断定这两篇报道是内容相关的.

(3) {$D$,$Pe$}

如果可以在新闻流中找到在相同日期发生、涉及相同人物或组织的报道,说明新到达报道与历史报道的内容具有关联性,是对已知历史事件的后续报道.

(4) {$Pl$,$Pe$}

如果可以在新闻流中找到在相同地点发生、涉及相同人物或组织的报道,说明新到达报道是已知历史事件的后续报道,并不涉及新的事件.

假设 $s_{new}$ 表示新闻流中新到达的报道,$e_{snew}$ 表示以 $s_{new}$ 为种子建立的新事件.$S$ 表示历史报道集,$S_{window}$ 表示窗口中的报道集合.$s_{old}$ 表示某一历史报道,与其相关的事件是 $e_{sold}$.$s_1(NE)$=$s_2(NE)$表示报道 $s_1$ 与 $s_2$ 具有相同的实体 NE.基于新闻特征语义匹配的事件检测算法见算法 1.

**算法 1**. 基于新闻特征语义匹配的在线事件检测算法.

1: Input: $S$, $S_{window}$, $s_{new}$;

2: Output: $Label(s_{new})$,$e_{snew}$;

3: for all $s_{old}$ in $S_{window}$

4: if($s_{old}(D,Pl,Pe)$=$s_{new}(D,Pl,Pe)$ ||

5: $s_{old}(Pl,Pe)$=$s_{new}(Pl,Pe)$ ||

```
6:          s_old(D,Pe)= s_new(D,Pe)) {
7:          Sim(s_new,e_sold)=1;
8:          Label(s_new)=OLD;        }
9: for all unlabeled s_new
10:     for all s_old in S
11:         if(s_old(D,Pl,Pe)=s_new(D,Pl,Pe) ||
12:            s_old(Pl,Pe)=s_new(Pl,Pe)        ||
13:            s_old(D,Pe)=s_new(D,Pe)) {
14:            Sim(s_new,e_sold)=1;
15:            Label(s_new)=OLD;        }
16:for all unlabeled s_new
17:   similarity=Sim(s_old,s_new);   //s_old∈{S∪S_window}
18: if (similarity>sim){
19:     if (s_old(D,Pl) = s_new(D;Pl)){
20:         Label(s_new)=OLD;
21:         e_snew=e_sold;}
22:     else {Label(s_new)=NEW;
23:            CreateNewEvent(s_new);}
```

在检测的过程中,需要基于检测结果不断地对历史信息进行更新,以保障对后续到达报道检测的准确性.在对 $s_{new}$ 进行事件检测时,如果 $s_{new}$ 与 $s_{old}$ 内容相关并且 $s_{old}$ 在窗口中,利用 $s_{new}$ 替换窗口中的 $s_{old}$,并建立 $s_{new}$ 与事件 $e_{sold}$ 的关联.

基于新闻特征语义匹配的检测方法是针对互联网中多个信息源对相同事件的重复报道提出的.利用这一方法可以迅速发现关于同一事件的相似和相关报道,并迅速建立报道之间、报道与历史事件间的关联.由于命名实体的长度通常远小于报道原文,因此在实体识别及新闻特征语义匹配过程带来额外开销的情况下,利用新闻特征进行的事件检测由于避免了大量的文本相似度计算,其效率仍然应优于传统的聚类算法.

### 2.4 基于后缀树聚类的检测

随着新闻报道的顺序到达,报道集合中词表的统计信息、词频及权重公布等均不断发生变化,若基于传统的文本模型进行主题检测,需要不断地从新闻流中识别出从未出现过的词,并实时地对词汇集、词频等进行更新.为解决检测过程中的这些问题,本文利用后缀树聚类完成剩余的未标识报道的检测工作:

(1) 历史报道集合的后缀树建立.初始情况下,只需完成已有报道间的关系判断并以它们为种子生成事件.建立树的过程包括扫描窗口内外的所有报道并分别建立、维护后缀树结构.

(2) 后缀树聚类.该算法是一个线性、增量的文本聚类算法.它利用文档间的公共子串来建立树结构,之后再动态地将文档聚为若干个利用词组标识的簇[6].

如果我们将一篇新报道归入已有的事件中,即表示 $Sim(s_{new},e_{old})>sim$ ,那么,我们还需要利用解析出的日期、地点判断报道间真正的关联.如果存在 $s_{old}(D,Pl)=s_{new}(D,Pl)$ ,则报道 $s_{new}$ 确实是关于某已知事件的介绍,否则,该报道的内容应是首次出现.这是因为即使内容十分相似的报道,若涉及不同的日期或地点,仍应是针对不同事件的报道.

如果利用后缀树聚类此步骤仍然无法找到某新报道与已知事件的关联,即表示

$$\neg\exists(e_{old}\in E\wedge Sim(s_{new},e_{old})>sim),$$

则 $s_{new}$ 为新事件的首篇报道.根据检测结果,将报道标识为 OLD 或 NEW.

在后缀树聚类中,对相似度阈值 $sim$ 的设定及新闻报道 $s_i$ 与事件 $e_j$ 间相似度 $Sim(s_i,e_j)$ 的计算和表示均与聚

类过程中的合并参数阈值 α 紧密相关.

(3) 后缀树更新.此步骤需要在窗口内外同步进行.若报道被标识为 NEW,则以它为种子建立新事件,根据时间戳更新事件窗口,从窗口中移出最早发布的报道.若报道标识为 OLD,则利用新报道替换 $s_{old}$.

## 3 实验及结果分析

实验算法使用 Java 语言在 JDK 1.5 环境中实现,本文的实验是基于 Pentium 4 3GHz 处理器,1GB 内存,Windows XP 操作系统实现的.实验数据来自于新浪新闻(news.sina.com.cn),网易新闻(news.163.com)及搜狐新闻(news.sohu.com)网站发布的 1000 个新闻页面.进行实验前,为避免 HTML 格式对检测结果造成影响,我们首先消除 Web 新闻报道中包含的 HTML 格式信息.在本文的实验中,设定 $life\_time_i$=24,$window\_size$=500.

本文的实验实现了 3 个检测算法,第 1 种是传统的增量聚类方法——后缀树聚类算法 STC,第 2、3 种均是利用实体识别技术实现的两阶段检测算法.NE_STC 算法利用实体识别技术从新闻报道中解析出第 1 个日期、人物或组织实体来描述报道内容.WNE_STC 算法中则对解析出的同类型实体进行了权重计算及排序,并选择了最高权重的实体作为代表描述报道内容.三种算法中均采用了加窗策略.

**Table 1** Time costs of three methods (ms)

**表 1** 检测算法时间消耗记录 (ms)

| Story Num | 80 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 |
|---|---|---|---|---|---|---|---|---|---|---|
| STC | 27625 | 94968 | 159187 | 191344 | 319875 | 380250 | 439734 | 503454 | 538594 | 725218 |
| NE_STC | 15016 | 66360 | 116859 | 161000 | 214328 | 247281 | 260390 | 338235 | 370688 | 527563 |
| WNE_STC | 13265 | 51954 | 87657 | 106890 | 170266 | 201672 | 232968 | 281094 | 336421 | 440890 |

表 1 记录的是利用 STC、NE STC 及 WNE STC 算法进行事件检测时,随 Web 新闻报道数量的增长,算法时间消耗的记录.其中,后缀树聚类算法中的基础类合并参数阈值 α=0.68.从表 1 中可以看出,在相同的数据集上运行时,3 种方法的时间消耗都是线性增长的,与传统 STC 的方法相比,NE_STC 及 WNE_STC 算法的效率得到了显著提高.

根据图 2 的内容显示,利用 NE_STC 及 WNE_STC 方法进行在线事件检测时,平均有超过 50%的报道通过新闻特征的匹配即可完成检测.也就是说,有超过 50%的报道不需要通过文本相似度的计算,即可判断出是否涉及新发生的事件,因而大幅提高了算法的效率.由于 WNE_STC 方法识别出的新闻特征更精确,降低了与无关报道建立关联的可能性,因而其比例低于 NE_STC 算法.

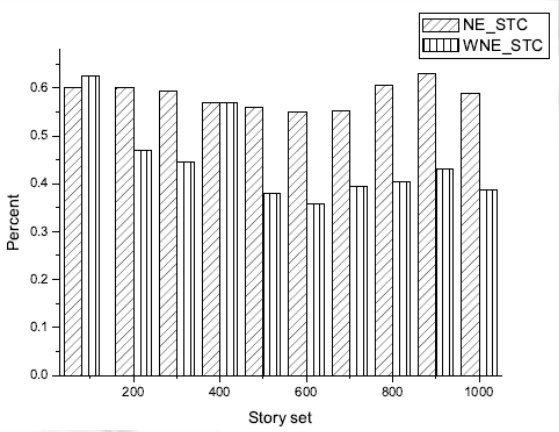



Fig.2 Percentage of detection with news characteristics

图 2 利用新闻特征匹配完成检测的报道比例

我们还利用检测算法的性能衡量指标－丢失率(miss rate)以及误报率(false alarm),对三种在线事件检测方

法的性能进行了比较.当系统未能发现关于新主题的报道时,称为检测结果的丢失(miss).当系统误将已知主题的相关报道标识为新主题时,称为检测结果的误报(false alarm).

根据人工标注的事件信息及手工检测结果,3 种检测方法的平均丢失率及误报率记录如图 3 所示.从图 3 可以看出,WNE_STC 方法的丢失率及误报率均为最低值,优于传统的增量聚类算法——STC,及未区分权重的 NE_STC 方法.

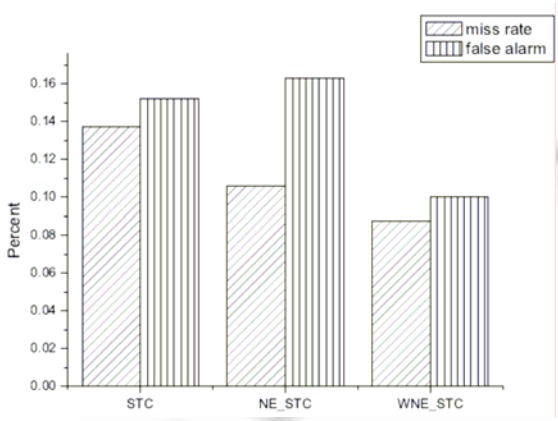



Fig.3 Records of miss rate and false alarm

图 3 STC、NE STC、WNE STC 算法的丢失率、误报率记录

本文还通过改变基础类合并阈值,判断了 STC 方法、NE_STC 方法及 WNE_STC 方法在不同情况下的检测性能.在以下的实验中,以相同的 500 篇 Web 新闻报道为测试集,对$\alpha$分别为 0.58,0.63,0.68,0.73,0.78 的检测情况进行探讨.图 4 和图 5 中记录的是随$\alpha$取值变化,检测事件数量及时间消耗的变化情况.

由图中可以看出,随着$\alpha$的变化,三种算法的运行时间及检测的事件数目均未发生明显的改变.可以认为基于后缀树聚类的事件检测对合并阈值$\alpha$的变化并不敏感.

## 4 相关研究

在线事件检测研究的目标是从实时、连续的新闻报道流中第一时间发现新近事件的出现点.因此,在线检测工作需要对某篇新闻报道是否涉及从未出现过的事件介绍进行判断,并对其进行标识.近年来,众多研究者均在此领域展开了大量的研究工作,其中广泛研究及采用的方法主要包括增量聚类、加窗略及改进的文本模型.

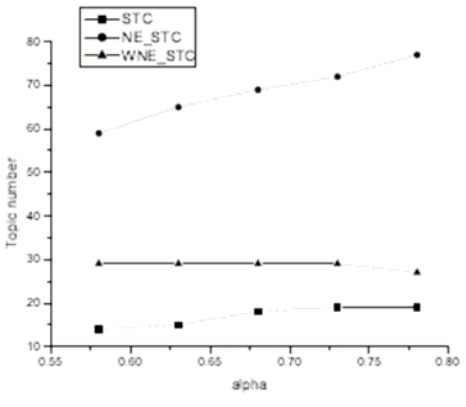



Fig.4 Records of events with different $\alpha$

图 4 检测事件数量随$\alpha$的变化记录

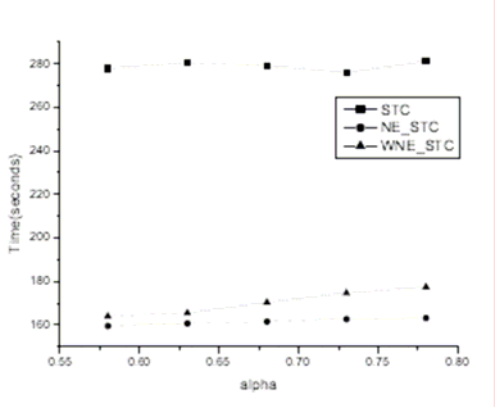



Fig.5 Records of time costs with different $\alpha$

图 5 检测时间消耗随$\alpha$的变化记录

增量聚类算法利用文本特征集合(利用向量或概率分布表示)表示每篇新闻报道,并依次对每篇报道进行处理[1−3].此类算法利用第一篇新闻报道作为种子建立第一个事件.之后,当其他的报道依次到达时,利用特征集合对新报道及历史事件的相似度进行计算.如果新报道与某一历史事件的相似度大于预设的聚类阈值 $e_c$,即将该报道归入该事件中.否则,该报道的内容应该是对某突发事件的首次介绍,即将其作为一个新事件的种子.以此类推完成所有报道的检测工作.

通常情况下,围绕某一具体事件的新闻报道在发布时间上均十分接近,这是由事件本身的特征及新闻报道的特点决定的.因此,加窗着重考虑了新闻报道的发布时间[2,3].常用的加窗策略有两种,一是为报道加窗,另一种是为事件加窗.报道加窗策略中认为发布时间间隔较短的新闻报道比间隔较长的新闻报道更相似.事件加窗为历史事件建立滑动窗口,那么新发布的报道只需要与窗口内的事件进行比较即可确定是否涉及第一次出现的事件.

还有些研究者采用了改进的文本模型完成检测任务[7],使用初始词表及倒文档频率(IDF:inveerse document frequency)进行事件检测.之后,算法收集新发布报道的信息对初始的词表及IDF进行增量更新,因此被称为增量的 TF-IDF 模型.

近年来,越来越多的研究引入更多的属性来实现在线事件检测[8,9].也有研究者对现实应用环境下的检测任务进行了深入探讨[5,10−12].

然而现有的在线事件检测算法主要基于文本相似度计算,因此检测算法的性能有限.

## 5 结 论

基于加窗策略、命名实体识别及后缀树聚类技术,本文提出并实现了一个面向互联网新闻的在线事件检测算法.本文的方法有以下特点:(1) 通过加窗缩小了事件检测的范围及问题规模.这一策略中充分利用了新闻报道的时间特性,以期大幅提高检测的效率.(2) 实体识别技术充分利用了新闻报道的特征,例如地点、人物及组织,保障了算法性能的优越性.(3) 本文的算法相继使用了新闻特征的语义匹配及后缀树聚类实现了事件检测,在不影响检测准确性的情况下大幅提高了检测效率.但是,本文现有的研究内容中尚未对加窗策略对检测性能造成的影响进行数值评估,在未来的工作中,我们将在这一方向进行更深入的探讨和研究.

**References**:

[1] Allan J, Carbonell J, Doddington G, Yamron J, Yang Y. Topic detection and tracking pilot study: Final report. In: Proc. of the DARPA Broadcast News Transcription and Understanding Workshop. 1998. 194–218.

[2] Yang Y, Pierce T, Carbonell J. A study on retrospective and on-line event detection. In: Proc. of the 21st ACM Int'l Conf. on Research and Development in Information Retrieval. 1998. 28–36.

[3] Yang Y, Carbonell J, Brown R, Pierce T, Archibald BT, Liu X. Learning approaches for detecting and tracking news events. IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval, 1999,14(4):32–43.

[4] Allan J, Lavrenko V, Jin H. First story detection in TDT is hard. In: Proc. of the 9th Int'l Conf. on Information and Knowledge Management. ACM, 2000. 374–381.

[5] Luo G, Tang C, Yu PS. Resource-Adaptive real-time new event detection. In: Chan CY, Ooi BC, Zhou A, eds. Proc. of the 2007 ACM SIGMOD Int'l Conf. on Management of data. ACM, 2007. 497–508.

[6] Zamir O, Etzioni O. Web document clustering: A feasibility demonstration. In: Proc. of the 21st ACM Int'l Conf. on Research and Development in Information Retrieval. Melbourne: ACM Press, 1998. 46–54.

[7] Brants T, Chen F. A system for new event detection. In: Proc. of the 26th ACM Int'l Conf. on Research and Development in Informaion Retrieval. ACM Press, 2003. 330–337.

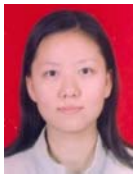
[8] Yang Y, Zhang J, Carbonell JG, Jin C. Topic conditioned novelty detection. In: Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2002. 688–693.

[9] Kumaran G and Allan J. Text classification and named entities for new event detection. In: Proc. of the 27th ACM SIGIR International Conference on Research and Development in Information Retrieval. Sheffield: ACM Press, 2004. 297–304.

[10] Zhang K, Zi J, Wu LG. New event detection based on indexing-tree and named entity. In: Proc. of the 30th ACM Int'l Conf. on Research and Development in Information Retrieval. ACM Press, 2007. 215–222.

[11] Chen CC, Chen MC, Chen MS. An adaptive threshold framework for event detection using HMM-based life profiles. ACM Trans. on Information System, 2009,27(2).

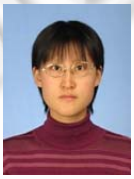
[12] Wang C, Zhang M, Ma S, Ru L. Automatic online news issue construction in Web environment. In: Huai J, Chen R, Hon HW, Liu Y, Ma WY, Tomkins A, Zhang X, eds. Proc. of the 17th Int'l Conf. on World Wide Web. Beijing: ACM, 2008. 457–466.



**付艳**(1980－),女,陕西西安人,博士,讲师,主要研究领域为信息检索,数据挖掘.

**王学松**(1975－),男,博士,工程师,主要研究领域为信息检索,虚拟现实.

**周明全**(1956－),男,博士,教授,博士生导师,主要研究领域为计算机可视化,软件工程,中文信息处理.



**栾华**(1980－),女,博士,讲师,主要研究领域为数据库,数据仓库.