

动态增量式子主题事件演化分析

李凤环 郑德权 赵铁军
(语言语音教育部-微软重点实验室(哈尔滨工业大学) 哈尔滨 150001)
(fhli@mtlab.hit.edu.cn)

Dynamic Incremental Analysis of Sub-Topic Evolution

Li Fenghuan, Zheng Dequan, and Zhao Tiejun
(MoE-MS Key Laboratory of Natural Language Processing and Speech (Harbin Institute of Technology), Harbin 150001)

Abstract There has been increasing interest in the follow-up progress of events because of sustainability and mutual influence of events. Meanwhile, more and more emergent events make it necessary to follow events in an intuitive and efficient way. However, the majority of traditional event analysis is sentence-oriented or topic-oriented which is event extraction or topic detection and tracking. A hierarchical structure of the topic event is constructed according to the research objects and the scope. A dynamic incremental model is presented for analyzing sub-topic dynamics in the topic event, which borrows the ideas of single-pass clustering, multi-category and dynamic incremental model. It is document-oriented and built on temporal property of the topic event, including dynamic threshold selection, similarity smoothing and dynamic incremental strategy. Meanwhile, overall evaluation criteria combined with χ^2 -test is served for performance analysis. The algorithm is effective and facilitates users to follow the topic event explicitly. Experimental results reported for four well-known topic events in China show that the performance of sub-topic evolution analysis is improved significantly.

Key words evolution analysis; sub-topic; dynamic; incremental; topic event

摘 要 事件发展的持续性和相互影响性使人们对事件的后续进展越来越感兴趣,而传统的事件分析大多是针对基于句子的事件.针对专题事件,结合 single-pass 聚类方法、兼类思想以及动态增量思想,提出了一种增量式子主题动态演化分析模型.该模型基于专题事件的时序特征提出,包括动态阈值的设定、相似度平滑、子主题动态增量策略等过程,以及运用 χ^2 统计的思想来综合评价模型性能的方法.该模型可以有效地对专题事件进行子主题分析,进而使人们能够更直接和快速地了解主题事件的进展.实验结果表明提出的方法使子主题演化分析的性能有了显著的提高.

关键词 演化分析;子主题;动态;增量式;主题事件

中图法分类号 TP391;TP18

随着互联网的发展,人们越来越多地从互联网上和快捷的方式来了解事件的信息和进展,因此关于上获知自己感兴趣的事件以及迫切地需要以更直观信息抽取和事件抽取的研究应运而生.当前对事件

分析^[1-3]的研究大多是基于句子或短语的,也就是说研究的是一个动作或状态的变化^[4].但是人们更多的是想关注一个专题事件整体的信息框架和发展过程,因此很多门户网站都推出了专题栏目.为了能够快速了解并掌握专题事件中每篇报道的信息以及后续进展情况,研究者提出了许多解决策略^[5-7].目前,关于专题事件的代表性研究有话题检测与跟踪(topic detection and tracking, TDT)^[8-11]和时间序列文本挖掘方法(temporal text mining, TTM)^[12-17]. He 等人^[18]利用物理中的质量、速度、动量、加速度等动量模型来构造主题的动态性,并把该模型运用到 PubMed 数据库中分析医学文献中的突发话题. Leskovec 等人^[19]基于到达率的突发模型,提出了可扩展的聚类算法来识别在线文本中特定短语. 陈伟^[20]对时序新闻主题分解与摘要(topic decomposition and summarization, TDS)进行了研究,提出了基于非负矩阵分解(non-negative matrix factorization, NMF)的方法,其将时序新闻表示成关键词-句子矩阵,利用 NMF 将矩阵分解成一个非负的子主题矩阵和一个非负的编码矩阵,这样原始文档集合就被分解成基本的子主题. Xie 等人^[21]基于 TextTiling 算法自动划分中文新闻报道中的话题,其核心思想是通过句子间相关度的下降坡度识别子话题的边界.以上方法的局限在于,它们是基于句子的,而人们更多的是关注一篇文档描述的子主题,而且句子只是按照子主题划分,并没有给予主题明确的类别标签.

针对上述问题,本文提出了面向专题事件的子主题切分,即子主题动态演化分析,也就是说判断某篇新闻是描述该专题事件的哪个子主题. 本文结合 single-pass 聚类方法、兼类思想以及动态增量思想,针对文档,结合话题检测与跟踪和时序文本挖掘技术实现增量式子主题动态演化分析,包括基于时序特征的向量空间模型、相似度的平滑和阈值的选择,并运用 χ^2 统计的思想综合评价模型性能. 通过多组实验对比详细论述了该方法的有效性.

1 任务描述及相关定义

专题事件是对一个事件的完整描述,包括事件的发生、后续跟踪报道以及和它相关的活动,对应于 TDT 中一个话题的具体事件,即专题栏目中的某个具体专题,如“2008 年汶川地震事件”、“2011 年温州动车事件”等. 一个专题事件在不同的发展阶段有不

同的侧重点,如图 1 所示的上海地铁事故事件,在整个过程中总共有 4 个侧重点,即事发情况、救援、社会反响、原因调查. 开始时几乎所有的报道都是描述事故发生时的情景,这时频繁出现的关键词是“相撞”、“追尾”等;在救援阶段,关键词是“救援”、“伤者”等;紧接着是社会反响,关键词是“感受”、“采访”等;最后是对事故原因的推测和调查,这时“调查”、“责任”等词频繁出现. 随着事件发展和人们关注的侧重点不同,相应的关键词也随着变化.

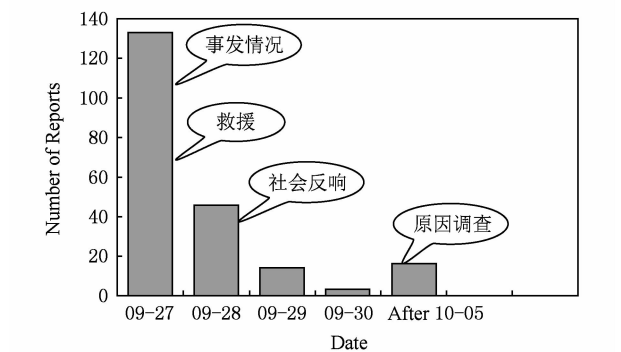


Fig. 1 Progress of Shanghai subway accident in 2011.
图 1 2011 年上海地铁事故发展过程

在对专题事件分析时,子主题即不同阶段的侧重点,是了解事件进展的关键. 所谓子主题动态演化分析即从一组新闻文档集中检测当前报道所属的子主题并发现新的子主题,且将同一子主题的新闻组织在一起. 为了更明确地理解本文研究内容,首先介绍一些术语^[20,22],它们之间的关系如图 2 所示.

1) 元事件(meta event). 表示一个动作的发生或状态的变化,往往由动词或能表示动作的名词等作为触发词触发. 如图 3 中的加粗字体“遇难”、“受伤”等.

2) 主题事件(topic event). 包含一类核心元事件,以及所有与之直接相关的元事件或活动. 可以由多个主题事件片段组成,通常对应于新闻网站专题栏目中的专题报道,如“2008 年汶川地震事件”等.

3) 主题事件片段(event segment). 一个主题事件片段通常由多个元事件组成. 多个主题事件片段完整地表示一个主题事件,通常一个主题事件片段对应一篇报道,如图 3 中 2008-05-15T01:24 对汶川地震的报道.

4) 子主题(sub-topic). 一个主题事件不同的发展阶段和人们关注的侧重点,如“救援”、“纪念活动”等.

5) 主题事件片段的时间. 是指报道中事件情景发生的时间,被指定为关键元事件的时间. 比如,在

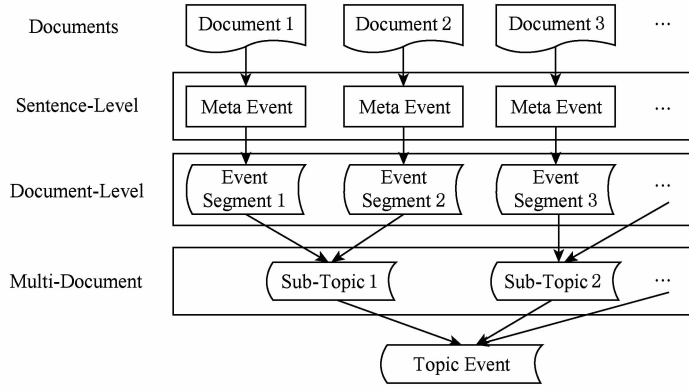


Fig. 2 Framework of topic event.

图2 主题事件知识框架

图3的报道中,关键元事件是“遇难”,主题事件片段的时间为“昨日中午”,其规范化后的时间是“2008-05-14”。

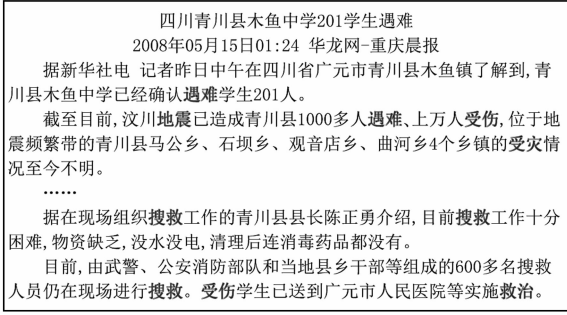


Fig. 3 One report about Wenchuan earthquake.

图3 关于汶川地震的一篇报道

2 基于时间序列的向量空间

2.1 文档的向量空间

一般情况下,文档由向量空间模型表示,由一系列关键字以及关键字的权重表示。每个关键字的权重表示对该文档的支持度,根据空间向量模型,文档被表示成:

$$\{(term_1, weight_{term_1, d}), (term_2, weight_{term_2, d}), \dots, (term_i, weight_{term_i, d}), (term_{i+1}, weight_{term_{i+1}, d}), \dots, (term_n, weight_{term_n, d})\}, \quad (1)$$

其中, $\{term_i | 1 \leq i \leq n\}$ 表示该主题事件中的关键字, n 表示该主题事件中关键字的个数; $weight_{term_i, d}$ 表示 $term_i$ 在该文档 d 中的权重,其由词频逆文档频率 (term frequency inverse document frequency, TFIDF) 公式以及平滑值获得,计算公式如下:

$$I_i^{TFIDF} = tf_i \times \ln \frac{N}{n_i}, \quad (2)$$

$$weight_{term_i, d} = \frac{I_i^{TFIDF} + \mu}{\max_{1 \leq i \leq n} \{I_i^{TFIDF}\}}, \quad (3)$$

其中, I_i^{TFIDF} 表示词语 $term_i$ 对该文档 d 的支持度, tf_i 表示词语 $term_i$ 在该文档 d 中出现的频率, N 表示整个主题事件中文档总数, n_i 表示主题事件中含有该词语 $term_i$ 的文档数, μ 为平滑因子。

由于算法的复杂度随着向量空间的规模呈指数级增长,本文只采用对表达文档语义比较重要的名词、动名词、动词作为文档的关键字来降低文档空间的维度,并选用停用词表进行过滤。通过分析大量的实验数据发现,该方法能够降低 90% 的复杂度。

因为事件报道的动态性,同时主题事件的侧重点随时间的演化而发生变化,因此,仅仅利用关键字来表示文档是不合理的。一般来讲主题事件和子主题会持续一段时间,如果文档与某个子主题中的主题事件片段时间一样,该文档属于这个子主题的可能性就越大。因此本文将主题事件片段的时间加入到文档空间中,具有时序特性的文档空间模型如下:

$$\{(term_1, weight_{term_1, d}), (term_2, weight_{term_2, d}), \dots, (term_n, weight_{term_n, d}), (time_1, weight_{time_1, d}), (time_2, weight_{time_2, d}), \dots, (time_m, weight_{time_m, d})\}, \quad (4)$$

其中, $\{time_j | 1 \leq j \leq m\}$ 表示第 j 个主题事件片段时间,以规范化的时间表达式形式“年-月-日”表示; m 是该主题事件的持续时间,一般以天为单位,选为新闻报道该主题事件的天数;文档的向量空间模型大小为 $m+n$; $weight_{time_j, d}$ 表示 $time_j$ 在该文档 d 中的权重; $time_d$ 表示文档 d 的主题事件片段时间,计算公式如下:

$$weight_{time_j, d} = \begin{cases} 1, & time_j = time_d; \\ 0, & time_j \neq time_d. \end{cases} \quad (5)$$

根据以上描述,文档就表示成了具有时间特性的向量空间,该空间包含了能表达文档特征的关键字和时间信息.在进行子主题演化分析时,我们将这些文档按照它们的时间信息进行升序排序,将文档作为具有时序性的数据流处理,这样就形成了文档的时间序列,为后续的动态增量式模型提供了保障.

2.2 子主题事件的向量空间

由于子主题是由多个主题事件片段组成的文档集,子主题的向量空间依赖于文档的向量空间,根据2.1节文档空间的建立,将式(1)~(4)中的文档 d 换成子主题 sub_k ,就表示成了子主题 sub_k 的向量空间模型.其中, $\{term_i | 1 \leq i \leq n\}$ 表示该主题事件中的关键字, n 表示该主题事件中关键字的个数, $weight_{term_i, sub_k}$ 表示 $term_i$ 在该子主题 sub_k 中的权重, I_i^{TFIDF} 表示词语 $term_i$ 对该子主题 sub_k 的支持度, tf_i 表示词语 $term_i$ 在该子主题 sub_k 中出现的频率, N 表示整个主题事件中文档总数, n_i 表示主题事件中含有该词语的文档数, μ 为平滑因子.

由于子主题可能会持续一段时间,也就是说 $time_{sub_k}$ 可能是包含了多个主题事件片段的时间集合,因此计算时间在该子主题中的权重时,而不是简单地用式(5)中的条件判断 $=$ 和 $\neq$,而是 $\in$ 和 $\notin$,计算公式如下:

$$weight_{time_j, sub_k} = \begin{cases} 1, & time_j \in time_{sub_k}, \\ 0, & time_j \notin time_{sub_k}, \end{cases} \quad (6)$$

其中, $weight_{time_j, sub_k}$ 表示 $time_j$ 在该子主题 sub_k 中的权重, $time_{sub_k}$ 表示子主题 sub_k 中所有文档的主题事

件片段的时间集合.同时,本文采用动态演化策略,文档以及子主题的向量空间模型的大小 $m+n$ 会随着算法的执行而发生相应的变化.

3 增量式子主题动态演化模型

3.1 动态演化策略

随着子主题的演化分析,该主题事件的新闻报道数据会不断增加,同时关于该事件的子主题随事件的发展而发生变化,如图1所示.在不同的时间段,事件的进展和人们的关注点会发生变化,因此为了子主题动态演化分析的准确性,子主题模板需要不断地修正,随着算法不断执行,子主题模板需要重新计算,这体现了子主题的动态性和时序性.子主题的动态演化算法思想如图4所示,图4中的“ $\cdots$ ”表示子主题模板和模板集的更新和增量,以便后续报道处理时用到的是实时更新的模板集; θ 为阈值.在整个过程中,动态性和增量式思想主要体现在:

- 1) 当前报道的子主题的动态演化结果要加入到相应的子主题模板中,以此为基础进行后续报道的处理,同时子主题模板要实时进行更新.
- 2) 在算法的执行过程中,可能会有新的子主题产生,则整个主题事件中的子主题模板的个数也会增加,相对应的参数(比如阈值 θ)也会随着发生改变.
- 3) 对所有主题事件指定相同的阈值也是不合理的,具体的文档和分析应该与其所在主题事件

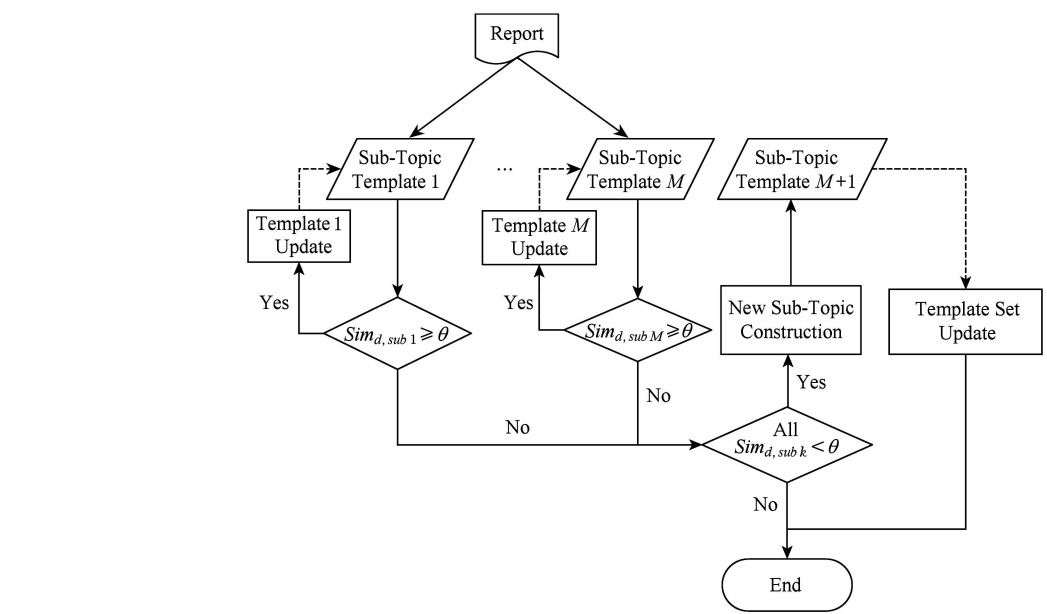


Fig. 4 Process of sub-topic dynamic evolution analysis.
图4 子主题动态演化分析过程

相关,因此每个主题事件应该根据该事件的具体情况具体分析,动态选择相应的参数。

3.2 子主题演化分析算法

本文借鉴 single-pass 聚类算法思想来实现子主题的动态演化分析,该算法具有较快的执行速度,缺点是它的算法执行结果依赖于语料被处理的顺序,这在本文中不会成为问题,因为在第 2 节将主题事件片段时间作为文档和子主题事件空间模型的重要部分.同时本文结合兼类和增量式思想进行子主题的动态演化分析,因为一篇报道中很可能描述了不少 1 个子主题,如图 3,不但描述了地震受灾情况,还描述了救援情况.对实验数据分析,平均一篇报道描述的子主题数为 1.78 个.因此本文采用兼类思想来进行子主题的确 定.具体步骤如算法 1 所示:

- 算法 1. 增量式子主题动态演化.
- 输入:主题事件片段;
- 输出:主题事件片段所属子主题.
- ① 对该主题中所有文档按照主题事件片段时间排序,文档按照时间顺序依次被处理;

② 对训练语料按照 2.1 节方法获得主题事件中的关键词,按照式(3)和式(5)计算关键字和主题事件片段时间的权重,形成文档空间,按照 2.2 节的方法组成子主题事件模板;

③ 文档与所有子主题模板计算相似度,分给超过阈值 θ 的子主题,并把该文档加入到相应的子主题中,转步骤②,子主题模板重新计算,该主题事件片段所属子主题为超过阈值的所有子主题,执行下一篇文档;

④ 如果所有相似度都低于某个阈值 θ ,则把该文档作为种子创建新的子主题,转步骤②,计算新子主题的模板,该主题事件片段所属子主题为新子主题,执行下一篇文档;

⑤ 所有主题事件片段都完成子主题的动态演化分析,算法结束.

从算法 1 可以看到,该算法的核心问题是文档与子主题的相似度计算和阈值的选择.这里采用传统的余弦距离公式来计算文档和第 k 个子主题的相似度,如式(7):

$$Sim'_{d,sub_k} = \frac{\sum_{i=1}^{m+n} (weight_{t_i,d} \times weight_{t_i,sub_k})}{\sqrt{\sum_{i=1}^{m+n} weight_{t_i,d}^2} \times \sqrt{\sum_{i=1}^{m+n} weight_{t_i,sub_k}^2}},$$

(7)

其中, $\{t_i | 1 \leq i \leq n\}$ 为该主题事件中的第 i 个关键

字, $\{t_i | n+1 \leq i \leq m+n\}$ 为第 $i-n$ 个主题事件片段时间, $\{sub_k | 1 \leq k \leq M\}$ 为主题事件中的第 k 个子主题, M 为该主题事件中子主题的个数.

阈值 θ 的选择是很关键的,因为它决定了子主题的选择和新子主题的确 定.每篇文档和各个子主题的相似度范围是不一样的,比如有的在 $[0.1, 0.2]$ 之间,也有在 $[0.4, 0.5]$ 之间,所以当设置阈值 θ 过高时很多文档判断不出所属子主题,设置过低时则会确定出不相关的子主题.因此本文中利用狄利克雷平滑方法^[23]对相似度进行了平滑,最终的相似度如式(8),其中 μ 为平滑因子.同时对于所有的主题事件指定相同的阈值也是不合理的,具体的文档应该与其所在主题事件相关,因此把阈值 θ 根据其所在主题事件中的子主题的个数来确定,如式(9).如果式(9)中的 M 也随着算法的执行而发生变化,那么同一个主题事件中的阈值是变化的,也就可能产生相似度一样、但检测出的子主题是不一样的情况,因此,式(9)中的 M 为开始执行算法时子主题的个数.由于本文采用动态演化策略,在算法的执行过程中,可能会有新的子主题产生,也就是 M 可能会发生变化.

$$Sim_{d,sub_k} = \frac{Sim'_{d,sub_k} + \mu}{\sum_{k=1}^M Sim'_{d,sub_k} + M \times \mu},$$

(8)

$$\theta = \frac{1}{M}.$$

(9)

4 实验设置

4.1 评价标准

表 1 表示实验结果的混淆矩阵,在进行子主题动态演化的效果分析时,采用准确率 (precision, P)、召回率 (recall, R) 和 F 值 (F -Score, F),并借鉴话题检测与跟踪中的评价标准,结合错报率 (fall-out, Fal) 和失报率 (miss, Mis) 来分析性能.我们希望准确率和召回率越高越好、错报率和失报率越低越好,因此本文提出了基于 χ^2 统计 (χ^2 -test) 的综合评价方法. χ^2 值越高性能越好,通常情况下, χ^2 越高

Table 1 Confusion Matrix of Evaluation Results

表 1 实验结果的混淆矩阵

Total Population	Actual Positive	Actual Negative
Predicted Positive	a	b
Predicted Negative	c	d

F 值越高. 本文中准确率和召回率同等重要, 因此选用 $\beta=1$. 各个指标的计算公式如下:

$$P = \frac{a}{a + b}, \tag{10}$$

$$R = \frac{a}{a + c}, \tag{11}$$

$$F = \frac{(\beta^2 + 1)(P \times R)}{(\beta^2 \times P) + R}, \tag{12}$$

$$Fal = \frac{b}{b + d}, \tag{13}$$

$$Mis = \frac{c}{a + c}, \tag{14}$$

$$\chi^2 = \frac{(a \times d - b \times c)^2}{(a + c)(b + d)(a + b)(c + d)}. \tag{15}$$

4.2 实验数据

本文的实验数据是从新浪网和凤凰网上爬取的 4 个专题事件^①, 并手工进行了标注. 表 2 中列出了专题事件的名称、文档数量、子主题信息、持续时间. 子主题信息根据维基百科^②对各个主题事件的描述来确定.

Table 2 Information of Corpora
表 2 语料信息

Topic Event	# Reports	Sub-Topics	Duration
郭美美	458	风波初起、民众围观、郭美美回应、红十字回应、调查、慈善界地震	2011-06-22—2011-11-30
上海地铁	205	事发过程、原因调查、社会反响、救援	2011-09-13—2011-10-10
温州动车	1 338	事发过程、事故调查、救援、善后处理、社会反应、纪念活动	2011-07-23—2011-12-06
汶川地震	1 890	地震详情、救援、震后影响、重建、纪念活动	2008-05-12—2008-09-04

4.3 对照算法

基于时序的子主题分解和摘要是基于句子的, 本文的子主题的动态演化是基于文档的, 而话题检测与跟踪针对的是话题, 本文针对的是子主题. 结合以上技术与本文研究内容, 选用了 4 个实验来研究本文方法的合理性与有效性.

1) 阈值选择的关键性. 根据阈值选择不同, 来验证本文中阈值选择的合理性.

2) 文档向量空间模型中时间的重要性. 与没有加入时间信息的文档向量空间作比较, 以验证时间信息在子主题动态演化中的重要性.

3) 相似度平滑的必要性. 与没有平滑的结果进行效果分析, 以验证本文提出的相似度平滑方法的有效性, 并采用 TopN 原则进行对比. 所谓 TopN 原则就是选择相似度最高的 N 个结果作为最终结果. 本文中研究和分析了 Top1 和 Top2 的实验结果.

4) Baseline 系统. 目前关于子主题的研究中, 只是把句子按照子主题划分, 并没有表明子主题具

体是什么. 因此本文的 baseline 选用 Xie 等人^[21]提出的基于 TextTiling 的子主题划分算法, 并利用支持向量机(support vector machine, SVM)进行分类以确定具体的子主题.

5 实验结果与分析

在进行子主题动态演化实验时, 采用 3 份交叉验证方法, 将 4.2 节中的每个子主题数据平均分成 3 份, 也就是说每个专题事件数据被平均分成 3 份. 把其中 1 份作为测试数据, 并依次把该份数据中的每个子主题作为新子主题, 其余子主题作为已知子主题; 把其他 2 份数据中的已知子主题数据作为训练数据. 交叉验证后的平均结果作为最终结果.

5.1 阈值选择

阈值 θ 的选择在子主题的动态演化过程中是极其关键的, 如果阈值过大, 就会把很多相关的子主题过滤掉; 如果过小, 就会掺杂进太多不相关的子主

① http://news.ifeng.com/society/special/guomeimei/content-3/list_0/0.shtml
http://news.ifeng.com/mainland/special/shanghaiдитиxiangzhuang/content-4/list_0/0.shtml
http://news.ifeng.com/mainland/special/wzdongchetuogui/content-3/list_0/0.shtml
<http://news.sina.com.cn/z/08earthquake/scroll/1.shtml>
② <http://zh.wikipedia.org/wiki/郭美美事件>
<http://zh.wikipedia.org/wiki/2011年上海地鐵追尾事故>
<http://zh.wikipedia.org/wiki/温州动车追尾事故>
<http://zh.wikipedia.org/wiki/汶川地震>

题. 本文针对阈值的选择作了研究,以“上海地铁事故”为例,此专题事件中的子主题数为 4. 不同阈值时的实验结果如表 3 所示,阈值每次以 0.01 为步长变化,图 5 展示了阈值对各个评价指标的影响.

Table 3 Performance of Different Thresholds

表 3 不同阈值时的性能

%

Threshold	<i>R</i>	<i>P</i>	<i>F</i>	<i>Mis</i>	<i>Fal</i>	χ^2
0.23	67.7	71.1	69.4	32.3	8.2	37.9
0.24	66.5	73.9	70.0	33.5	7.1	39.0
0.25	65.4	78.1	71.2	34.6	6.1	40.5
0.26	62.9	80.7	70.7	37.1	5.2	40.3

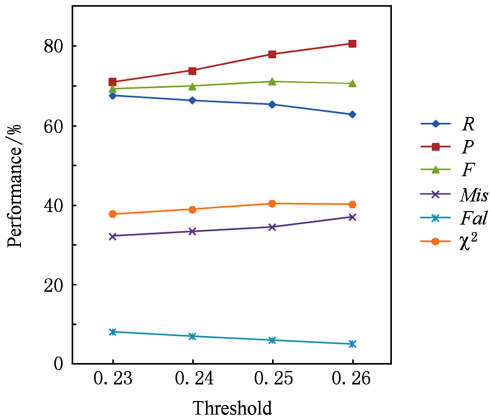


Fig. 5 Variation tendency of every criterion with the threshold changes.

图 5 各个指标随着阈值变化的变化趋势

从表 3 和图 5 可以看出,阈值越靠近 0.25,性能越好,阈值在 0.25 时 χ^2 值最高,同时 *F* 值也达到最大,也就是此时的实验效果最好;召回率随着阈值的增加呈下降趋势,也就是说当阈值越大时,有更多的子主题没有被识别出来,召回率就越低;准确率随着阈值的增大而提高,因为阈值越大,过滤掉的不相关的子主题越多,而识别出来的子主题准确率就越高;*F* 值在 χ^2 值达到最好时也达到了最高;失报率随着阈值的增加而增高,因为阈值越大,过滤掉的子主题就越多,因此失报率就越大;由于阈值越高,不相关的子主题被过滤掉,因此错报率随着阈值的增加而变小.

主题事件是发展变化的,不同事件的发展历程也是不一样的,因此所有的事件选择同一个阈值也是不明智的. 表 4 列出了每个事件在性能达到最好时的阈值,可以看到在性能最好时,不同事件的阈值是不一样的,而此时的阈值都接近或等于式(9)所求结果. 在一个主题事件中,每个子主题在事件发展过

程中的持续时间可能是不同的,对事件造成的影响也可能是不同的. 但是本文阈值的选择方法和优化决策中的平衡决策法是相通的,而使性能达到最好时的点就是各个子主题影响力的等值点,因此本文参考子主题个数设定阈值. 表 3、表 4 和图 5 也表明本文中阈值的选择方法是正确的.

Table 4 Thresholds in Different Events When Performance is Best

表 4 不同事件在性能最好时的阈值

Topic Event	# Sub-Topics	Best Threshold	<i>F</i> /%	χ^2 /%
郭美美	6	0.16	65.6	39.2
上海地铁	4	0.25	71.2	40.5
温州动车	6	0.17	72.9	43.4
汶川地震	5	0.20	78.8	42.8

5.2 相似度平滑

为了体现相似度平滑对实验效果的影响,本文做了相似度平滑和非平滑的研究,并把这 2 个实验与 Top1 和 Top2 效果比较. 实验结果如表 5 所示,表 5 是表 2 中 4 个专题事件结果的平均值. 从表 5 可以看出平滑后实验效果有了明显的提高,并且比 Top1 和 Top2 的效果要好. 当使用 Top1 策略时,准确率达到最高、召回率达到最低,因为此时文档选用相似度最大的子主题作为最终结果,这个结果正确的可能性最大,同时只有 1 个结果;而本文采用兼类思想分析子主题,很多子主题被漏掉,因此召回率相应较低同时失报率最高. 根据 Top*N* 理论,Top2 的准确率比 Top1 的低,而召回率较高. 在没有进行平滑时,主题中所有的文档都用同样的阈值,而文档和子主题的相似度范围不一定相同,因此准确率和召回率都偏低,而失报率和错报率偏高. 因此,非平滑时的效果远不如 Top1,Top2 以及平滑后的效果. 平滑后,文档与子主题的相似度被规范到更相近的区间,因此子主题确定时,效果就得到了显著的提高.

Table 5 Performance Comparison after Smoothing

表 5 平滑性能比较

%

Method	<i>R</i>	<i>P</i>	<i>F</i>	<i>Mis</i>	<i>Fal</i>	χ^2
Non-Smoothing	63.9	73.1	68.2	36.1	7.1	36.2
Smoothing	80.6	79.6	80.1	19.4	7.3	54.2
Top1	50.9	92.7	65.7	49.1	1.7	38.3
Top2	83.4	73.2	78.0	16.6	1.0	50.0

5.3 算法比较

5.1 节和 5.2 节实验没有考虑时间要素对性能的影响,表 6 给出考虑了时间要素时的效果,并与基于 TextTiling 的子主题划分算法的 baseline 系统相对比.表 6 是表 2 中 4 个专题事件结果的平均值,表明了各个指标在 3 种方法下的性能.

Table 6 Performance of Sub-Topic Dynamic Evolution Analysis

表 6 子主题动态演化分析的实验结果 %						
Method	R	P	F	Mis	Fal	χ^2
Without Time	80.6	79.6	80.1	19.4	7.3	54.2
With Time	83.3	78.9	81.0	16.7	4.0	60.3
TextTiling	88.6	68.9	77.5	11.4	15.9	46.3

从表 6 可以看出时间要素对各个指标的影响,当添加了时间要素信息后, χ^2 值得到了明显的提高,同时召回率、 F 值也有了相应的提高,失报率和错报率有了稍许的下降;但是由于加入时间要素信息后,文档的特征维度有了些许的增加,整体的准确率略有下降.

为了探究时间要素在各个专题事件中的影响,图 6 列出了各个评测标准在添加了时间要素后在各个专题事件中的变化幅度,变化幅度大于 0 表示提高,小于 0 代表下降,可以看到各个指标在添加了时间要素后都有了相应的变化. F 值和 χ^2 值在所有的专题中都是大于 0 的;错报率在所有的专题中都是小于 0 的,表明错报率都下降了,也就是说它们在添加了时间要素后,效果都有了明显的改进;召回率大多是提高的;失报率大多是下降的;而准确率提高的

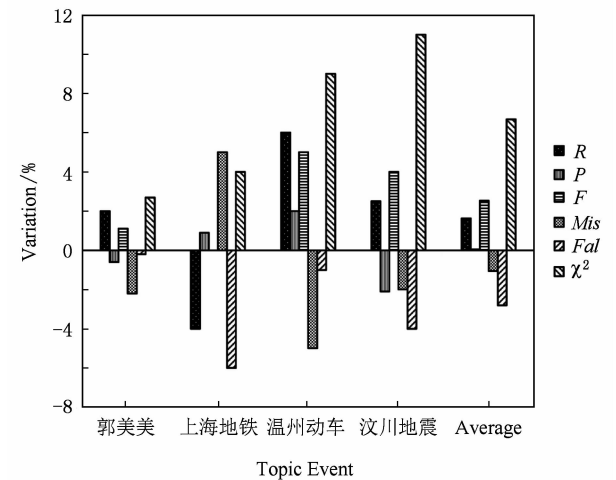


Fig. 6 Variation of every criterion after considering time.
图 6 考虑时间后各个指标的变化趋势

幅度和下降的幅度差不多.因此,在最终的结果中,召回率、 F 值、 χ^2 值都有明显的提高,准确率几乎没变,错报率和失报率明显下降,因此最终效果得到了显著的改进.

同时从表 6 可以看出,我们算法的最终结果比 baseline 系统提高了近 14%.由于 TextTiling 算法是以句子为单位,分析粒度比我们的算法小,更能从细节上划分子主题,从而召回率有很大的提高,失报率得到了显著的降低.但是这样也就不能更好地从整体上把握事件,比如在“汶川地震”的很多报道中,都会首先有“2008 年 05 月 12 日 14 时 28 分 04 秒,在四川汶川县发生 M7.6 级地震”等类似的提及,然后描述当前的进展.这样利用 TextTiling 算法时,就会把“地震详情”子主题识别出来,而相对忽略了该报道真正想要表达的内容,从而使准确率显著降低和错报率显著增加.同时每篇报道划分的子主题较多,在进一步分类时容易产生不相关的子主题结果,并且 baseline 系统中忽略了时间在算法中的重要性,因此本文算法的性能远远超过了 TextTiling 算法.

5.4 时间复杂度

评价一个算法好坏的另一个重要标准就是它的时间复杂度,该实验仍然选择 5.3 节中的 3 个算法,其中考虑时间和没有考虑时间这 2 种算法的时间复杂度是一样的,因为它们的区别是在算法中是否考虑了时间因素,现在主要分析它们与 TextTiling 算法的时间复杂度区别. TextTiling 算法是把文档分割成具有相同词汇的句子,然后将一定数量的句子集组成 1 个块,使用窗口滑动来计算每 2 个块之间相关度的下降坡度,进而识别子主题的边界.现在假设整个主题事件包含 e 个子主题、 d 个文档,每个文档中包含 w 个词语、 s 个句子,使用 TextTiling 算法时,将每 t 个词语组成 1 个句子,每个块中包含 k 个句子.整个算法包含向量空间的组成和子主题的演化分析,因此时间复杂度的分析也从这 2 个方面考虑.

1) 分析组成向量空间的时间复杂度.对于本文中的方法,因为要从整个主题事件中获取关键词,而 2.1 节中关键词的选择方法能降低 90% 的复杂度,所以其时间复杂度是 $f_1(w) = 0.1 \times d \times w = O(w)$.而对于 TextTiling 算法,因为每一次计算相关度都要使用新窗口中新的向量空间,因此它的时间复杂度为

$$g_1(w) = (s - (2 \times k - 1))(k \times t) =$$
$$(s - (2 \times k - 1))(\omega / (s/k)) =$$
$$(1 - (2 \times k - 1)/s) \times k \times \omega = O(\omega).$$

(16)

2) 分析子主题演化分析的时间复杂度. 这部分主要是相关度的计算. 因为本文的方法是面向文档的, 只需要计算该文档和各个子主题的相关度就可以, 因此本文方法的时间复杂度是 $f_2(s) = e = O(1)$. TextTiling 算法需要计算任何 2 个块之间的相关度, 因此它的时间复杂度是 $g_2(s) = s - (2 \times k - 1) = O(s)$.

从以上分析可以看到, 本文方法的时间复杂度远远小于 TextTiling 算法. 图 7 中列出了运用 5.3 节中的 3 个算法在 4.2 中的所有实验数据上的运行时间. 所有算法采用 C++ 编程, 利用 visual studio 2008 编译, 运行于 3.6 GHz 的 4 核处理器, 8 GB 内存的 64 位操作系统上. 从图 7 可以看出, 考虑时间因素和没有考虑时间因素的运行时间相当, 而运用 TextTiling 算法的运行时间远远长于其他 2 种算法. 因此从时间复杂度方面考虑, 我们的方法也是远远好于 baseline 系统的.

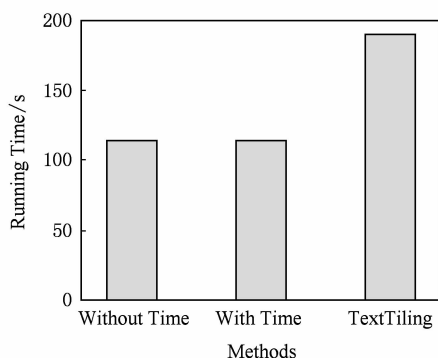


Fig. 7 Running time of every method.

图 7 各个算法的运行时间

6 结束语

本文提出了一种基于时序的增量式子主题动态演化分析模型, 该模型针对新闻网站上专题栏目中的专题事件, 可以有效地对专题事件在不同发展阶段的侧重点进行动态演化分析, 使人们更直观地了解事件的进展. 该方法有效地结合了专题事件的时序特征, 通过阈值的动态设定, 验证了本文阈值选择方法的正确性; 通过应用相似度平滑模型, 解决了数据稀疏问题, 缩小了相似度之间的悬殊差异; 同时, 基于动态增量策略, 验证了时间信息对子主题动态演化的影响; 此外, 运用 χ^2 统计的思想来综合评价模型的性能, 大大减小了算法的时间复杂度.

本文后续工作将在以下 2 方面展开: 1) 从不同

的角度和不同的语料进一步验证方法的有效性, 以使该方法能适用更大规模的实验平台; 2) 结合主题事件中人物、地点等元素, 分析子主题信息, 以更好地理解事件.

参 考 文 献

- [1] Boschee E, Natarajan P, Weischedel R. Automatic extraction of events from open source text for predictive forecasting [G] // Handbook of Computational Approaches to Counterterrorism. Berlin: Springer, 2013: 51-67
- [2] Atkinson M, Du M, Piskorski J, et al. Techniques for multilingual security-related event extraction from online news [G] // Computational Linguistics. Berlin: Springer, 2013: 163-186
- [3] Wang Wei, Zhao Dongyan. Ontology-based event modeling for semantic understanding of Chinese news story [G] // Natural Language Processing and Chinese Computing. Berlin: Springer, 2012: 58-68
- [4] Linguistic Data Consortium. ACE (Automatic Content Extraction) Chinese annotation guidelines for events, Version 5.5.1 [S]. Philadelphia, PA: Linguistic Data Consortium, 2005
- [5] Allan J. Topic Detection and Tracking: Event-Based Information Organization [M]. Berlin: Springer, 2002
- [6] Chen Yan, Amiri H, Li Zhoujun, et al. Emerging topic detection for organizations from microblogs [C] // Proc of the 36th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2013: 43-52
- [7] Zhao Hua. Research on topic detection and tracking [D]. Harbin: Harbin Institute of Technology, 2008 (in Chinese) (赵华. 话题检测与跟踪关键技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2008)
- [8] Yin Hongzhi, Cui Bin, Lu Hua, et al. A unified model for stable and temporal topic detection from social media data [C] // Proc of the 29th IEEE Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2013: 661-672
- [9] Becker H, Naaman M, Gravano L. Beyond trending topics: Real-world event identification on Twitter [C] // Proc of the 5th Int AAAI Conf on Weblogs and Social Media. Menlo Park, CA: AAAI, 2011: 438-441
- [10] Takahashi T, Tomioka R, Yamanishi K. Discovering emerging topics in social streams via link-anomaly detection [J]. IEEE Trans on Knowledge and Data Engineering, 2014, 26(1): 120-130
- [11] Zhang Kuo, Li Juanzi, Wu Gang, et al. Term-committee-based event identification within topics [J]. Journal of Computer Research and Development, 2009, 46(2): 245-252 (in Chinese) (张阔, 李娟子, 吴刚, 等. 基于关键词元的话题内事件检测 [J]. 计算机研究与发展, 2009, 46(2): 245-252)

- [12] Mei Qiaozhu, Zhai Chengxiang. Discovering evolutionary theme patterns from text: An exploration of temporal text mining [C] //Proc of the 11th ACM SIGKDD Int Conf on Knowledge Discovery in Data Mining. New York: ACM, 2005: 198-207
- [13] Luo Kedong. Algorithm study on short time series mining [D]. Beijing: Tsinghua University, 2004 (in Chinese)
(骆科东. 短时间序列挖掘方法研究[D]. 北京: 清华大学, 2004)
- [14] Kleinberg J. Bursty and hierarchical structure in stream [C] //Proc of the 8th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2002: 91-101
- [15] Harpaz R, DuMouchel W, Shah N H, et al. Novel data-mining methodologies for adverse drug event discovery and analysis [J]. Clinical Pharmacology & Therapeutics, 2012, 91(6): 1010-1021
- [16] Lee C H, Yang H C, Chien T F, et al. A novel approach for event detection by mining spatio-temporal information on microblogs [C] //Proc of the 3rd Int Conf on Advances in Social Networks Analysis and Mining. Piscataway, NJ: IEEE, 2011: 254-259
- [17] Lee C H. Mining spatio-temporal information on microblogging streams using a density-based online clustering method [J]. Expert Systems with Applications, 2012, 39(10): 9623-9641
- [18] He D, Parker D S. Topic dynamics: An alternative model of 'bursts' in stream of topics [C] //Proc of the 16th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2010: 443-452
- [19] Leskovec J, Backstrom L, Kleinberg J. Meme-tracking and the dynamic of the news cycle [C] //Proc of the 15th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2009: 497-505
- [20] Chen Wei. Exploring temporal text mining for news content anatomy and recommendation [D]. Hangzhou: Zhejiang University, 2010 (in Chinese)
(陈伟. 基于时序文本挖掘的新闻内容理解与推荐技术研究[D]. 杭州: 浙江大学, 2010)
- [21] Xie Lei, Zeng Jia, Feng Wei. Multi-scale texttiling for automatic story segmentation in Chinese broadcast news [C] //Proc of the 4th Asia Information Retrieval Conf on Information Retrieval Technology. Berlin: Springer, 2008: 345-355
- [22] Wu Gang. Research and application on Chinese topic event extraction [D]. Suzhou: Soochow University, 2009 (in Chinese)
(吴刚. 基于主题的中文事件抽取技术研究及应用[D]. 苏州: 苏州大学, 2009)
- [23] Zhai Chengxiang, Lafferty J. A study of smoothing methods for language models applied to ad hoc information retrieval [C] //Proc of the 24th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2001: 334-342



Li Fenghuan, born in 1985. PhD candidate. Student member of China Computer Federation. Her current research interests include information extraction, event detection, anomaly detection, time series analysis, text mining, and natural language processing.



Zheng Dequan, born in 1968. PhD, associate professor of Harbin Institute of Technology. Member of China Computer Federation. His current research interests include information retrieval, knowledge engineering, natural language processing.



Zhao Tiejun, born in 1962. PhD, professor of Harbin Institute of Technology. Senior member of China Computer Federation. His current research interests include natural language processing, machine translation, content-based Web information processing, applied artificial intelligence(tjzhao@hit.edu.cn).