动态话题追踪中的时序权重

吴树芳[1] 徐建民[2]
(河北大学管理学院 保定 071002)[1] (河北大学数学与计算机学院 保定 071002)[2]

摘 要 在贝叶斯信念网络的基础上,给出了一个新的动态话题追踪模型作为文章的表示模型。依据时间距离量化动态话题追踪中的时序信息,并将其应用于特征权重的动态调整。考虑到较长时间没有再现的特征权重应该衰减,给出了权重衰减函数,若衰减后的特征权重低于一定的阈值,则将其视为冗余信息。实验采用 TDT4 测试集合和 DET 曲线进行评测,通过反复实验获得基于 TDT 语料的最优时间距离阈值 α 和决定是否为冗余特征的阈值 β。实验证明,使用时序权重后可有效提高动态话题追踪模型的追踪性能。

关键词 话题追踪,时序权重,衰减,贝叶斯信念网络
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.2.048

Temporal Weight in Dynamic Topic Tracking

WU Shu-fang[1] XU Jian-min[2]
(College of Management, Hebei University, Baoding 071002, China)[1]
(College of Mathematics and Computer, Hebei University, Baoding 071002, China)[2]

Abstract A new dynamic topic tracking model was proposed based on Bayesian belief network, which is used as the representation model in this paper. We used time distance to quantify temporal information which is then used to dynamically adjust feature weight. A weight decay function was given to deal with the long-time disappearing features. If the weight of a feature is lower than the given threshold after decaying, the feature will be viewed as redundant information. TDT4 corpora and DET curves were used to run experiments. We firstly obtained the optimal time distance threshold α and the threshold β to determine whether a feature is redundant information. Experimental results show that the tracking performance of dynamic topic models can be effectively improved by using temporal weight.

Keywords Topic tracking, Temporal weight, Decay, Bayesian belief network

1 引言

话题追踪的任务是监控新闻报道信息流以发现与某一已知话题有关的新报道。这类技术是现实中亟需的,比如,自动监控各种信息源,并从中获得关于已知事件的新信息,它被广泛应用于舆情分析、信息安全、证券分析等领域。

话题追踪系统主要包括 3 部分:话题建模、相关性判定和阈值估计。话题建模是对话题核心内容进行模型化描述,话题模型的前瞻性研究源自 J. Allan[1],他将信息检索领域的向量空间模型用于话题建模;L. S. Larkey 等人[2]提出了语言模型;R. Nallapati[3]利用话题先验样本中的各类语义关系,建立了基于语义的语言模型;洪宇等人[4]利用篇章结构和依存关系建立了语义域语言模型;骆卫华[5]提出了层次话题模型;赵华[6]对结构化话题模型进行了相关研究。这些话题模型可分为静态话题模型和动态话题模型。静态话题模型从已知新闻报道中挖掘出核心事件,将其作为话题的恒定核心贯穿话题追踪的始终,有悖于话题的演化规律。动态话题模型中话题的核心则随着相关事件的出现而产生演化现象。直觉上,动态话题模型应该是话题建模的首选样板[7]。

在动态话题追踪中,时序信息是一个重要属性[8,9],一个话题常常存在于一定的时间段内,将时序应用于话题识别与追踪是 TDT 领域一个重要的研究方向。这方面的先驱性研究来自于 Y. Yang[10],他提出了基于时间和空间排序的假设;贾自艳[11]建立了统一的时间表述机制;赵华[12]的研究工作类似于贾自艳的研究。以上研究或者基于经验假设,或者基于训练数据,不能保证追踪系统性能的稳定性。洪宇[13]和保丽丽[14]考虑了时间粒度对系统性能的影响;Masahiro[15]在分析舆情形成的过程中,对时序信息进行了量化;廖俊华[16]运用时间信息体现话题的演化,有效地提高了系统对热点话题的发现能力。但是,他们的研究或者采用线性的方法调整时序和其它因素的贡献度,或者将时序作为附加因素,实际上,时序信息直接影响了特征权重的计算,应视为直接性因素。如果一个特征在短时间内反复出现,则其权重应该提高,如果一个特征在长时间内没有再现,则其权重应该衰减。针对这

到稿日期:2013-04-08 返修日期:2014-06-10 本文受中国博士后科学基金资助项目(20070420700),河北省自然科学基金资助项目(F2011201146),河北省科技计划项目(13450337)资助。
吴树芳(1979—),女,博士生,讲师,主要研究领域为信息检索、不确定信息处理,E-mail:shufang_44@126.com;徐建民(1966—),男,博士,教授,主要研究领域为信息检索、不确定信息处理,E-mail:hbuxjm@hbu.cn(通信作者)。

种现象，本文提出了基于时间距离的时序权重和时间衰减函数，用于优化动态话题追踪模型的性能，实现其应用领域对话题追踪高准确率的要求。实验结果显示，应用时序权重可有效提高动态话题追踪模型的追踪性能。

2 相关工作

2.1 相关定义

为区别语言学中的概念，下文给出话题和其它常用的概念[17]。

定义 1(话题，Topic) 一个种子事件或活动，以及所有与之直接相关的事件或活动。

定义 2(事件，Event) 由某些原因、条件引起，发生在特定时间、地点，并可能伴随某些必然结果的一个特例。

定义 3(报道，Story) 与话题紧密相关，包含两个或多个独立陈述某个事件的子句的新闻片段。

定义 4(话题追踪，Topic Tracking) 目标话题由 1—4 个属于该话题的报道定义，追踪任务是将后续报道正确地划分为属于或是不属于该目标话题。

除以上定义外，给出时间距离的定义：

定义 5(时间距离，Time Distance) 特征 k_i 出现的最近时间和最早时间之差，即 $\Delta T_i = T_{ie} - T_{ib}$，其中 T_{ie} 表示特征 k_i 的最近出现时间，T_{ib} 表示特征 k_i 的最早出现时间。

2.2 评价指标

话题识别与追踪(TDT)的所有子任务都可以转化为检测任务，检测性能采用漏报率(Miss Probability)和误报率(False Alarm Probability)进行评价，将这两个错误概率归一化，可得到检测指标 C_{det}[18]：

$$C_{det} = C_{miss} \times P_{miss} \times P_{target} + C_{fa} \times P_{fa} \times P_{non\text{-}target} \quad (1)$$

其中，C_{miss}，C_{fa} 分别表示漏报代价和误报代价，P_{target} 表示发现一个新报道的概率，这 3 个值均为预设值。表 1 列出了 TDT 评测会议给出的不同子任务下 3 个因子的预设值[19]。

表 1 不同子任务下的评估代价参数预设值

子任务	Cmiss	Cfa	Ptarget
报道切分	1.0	0.3	0.3
话题追踪	1.0	0.1	0.02
话题识别	1.0	0.1	0.02
首报道检测	1.0	0.1	0.02
关联检测	1.0	0.1	0.02

观察上表可以发现，在话题追踪子任务中，C_{miss}，C_{fa}，P_{target} 的预设值分别为 1.0，0.1，0.02。$P_{non\text{-}target} = 1 - P_{target}$ 表示发现一个旧报道的概率。式(2)、式(3)给出了漏报率 P_{miss} 和误报率 P_{fa} 的计算方法：

$$P_{miss} = \frac{c}{a+c} \quad (2)$$

$$P_{fa} = \frac{b}{b+d} \quad (3)$$

其中，a 为“本应为是判断也为是”的判断个数，b 为“本应为是判断为否”的判断个数，c 为“本应为否判断为是”的判断个数，d 为“本应为否判断也为否”的判断个数。

通常，式(1)给出的代价函数规范化为：

$$(C_{det})_{norm} = \frac{C_{det}}{\mathrm{Min}(C_{miss} \times P_{target}, C_{fa} \times P_{non\text{-}target})} \quad (4)$$

$(C_{det})_{norm}$ 为检测因子 C_{det} 的规范化表示，$\mathrm{Min}(C_{miss} \times P_{target}, C_{fa} \times P_{non\text{-}target})$ 表示取 $C_{miss} \times P_{target}$ 和 $C_{fa} \times P_{non\text{-}target}$ 中的最小值。

3 表示模型

这部分介绍本文用到的话题模型。语言模型的计算涉及到条件概率，故也被解释为概率模型[20]。贝叶斯信念网络[21]作为信息检索领域概率模型的扩展，在过去几十年已经成功应用于信息检索领域，本文试图将该模型应用于话题追踪，寻找新的突破。

3.1 特征选取

要建立基于贝叶斯信念网络的动态话题追踪模型(BDTM)，应该首先进行特征提取[22]，式(5)是采用的权重计算公式[23]：

$$w(k_i) = \frac{freq(k_i) + 0.5N_{begin} + 0.5N_{end} + N_{title}}{\sum freq(k_i)} \quad (5)$$

考虑到新闻报道中重要词汇一般出现在报道的标题、开头和结尾，式(5)在计算术语 k_i 的权重时额外考虑了这 3 个位置的术语频度。其中 $w(k_i)$ 是特征 k_i 的权重，$freq(k_i)$ 表示特征 k_i 在报道 s_i 中出现的次数，N_{begin}、N_{end}、N_{title} 分别表示特征 k_i 在报道开始、结束及标题中出现的次数，$\sum freq(k_i)$ 表示所有特征在报道 s_i 中出现次数之和。通过权重计算，权重较大的前 i 个特征用于描述该报道 s_i。

因为在后期处理中需要计算时间距离，所以对报道的描述除了术语外，还应包括术语的最近发生时间和最早发生时间，即：

$$s_i = \{(k_{i1}, T_{i1b}, T_{i1e}), (k_{i2}, T_{i2b}, T_{i2e}), \cdots, (k_{in}, T_{inb}, T_{ine})\} \quad (6)$$

式中，k_{ij} 表示报道 s_i 中第 j 个术语，k_{ijb}、k_{ije} 分别表述术语 k_{ij} 的最早发生时间和最近发生时间。建模初期，特征 k_{ij} 的最早发生时间和最近发生时间均为该特征所属报道的发生时间，此时的时间距离为 0。

话题 t_j 由属于该话题的相关报道的并集描述：

$$t_j = \bigcup s_i \quad (7)$$

3.2 模型拓扑结构及概率推导

图 1 为基于贝叶斯信念网络的动态话题追踪模型(BDTM)拓扑结构，该模型包括 3 类节点：话题节点、新报道节点和特征节点，弧标明索引关系。在话题追踪过程中，左边内虚线框内的部分将会动态更新。当新报道 s_n 出现时，通过计算 s_n 和话题 t_j 的相似度 $sim(t_j, s_n)$，判断该报道是否属于话题 t_j，依据贝叶斯概率[24]：

$$sim(t_j, s_n) = P(t_j \mid s_n) = \frac{P(t_j, s_n)}{P(s_n)} \quad (8)$$

由于新报道的出现是随机的，属于等概率现象，因此认为 $P(s_n)$ 是一个常量，用 α 表示，即：

$$sim(t_j, s_n) = P(t_j \mid s_n) = \frac{P(t_j, s_n)}{P(s_n)} = \alpha P(t_j, s_n) \quad (9)$$

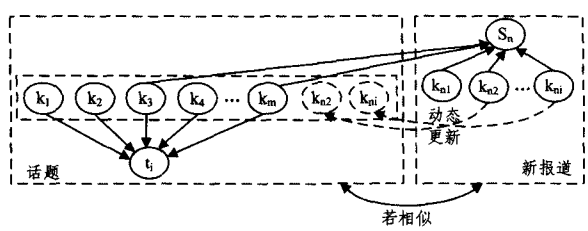


图 1 基于贝叶斯信念网络的话题追踪模型

观察图1，实际上是在已知术语集合 S 的条件下，计算联合概率 $P(t_j, s_n)$：

$$P(t_j, s_n) = P(t_j, s_n | S) \tag{10}$$

依据贝叶斯概率中的条件独立性假设，则上式可转化为：

$$P(t_j, s_n | S) = \sum_{\forall s \subset S} P(t_j, s_n | s) = \sum_{\forall s \subset S} P(t_j | s) \times P(s_n | s) \times P(s) \tag{11}$$

对条件概率 $P(t_j|s)$，$P(s_n|s)$ 的不同规定将会产生不同的排序策略，我们采用如下规定：

$$P(s_n | s) = \begin{cases} \dfrac{num(s_n \cap s)}{num(s)}, & num(s_n \cap s) > \gamma \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

$$P(t_j | s) = \frac{\sum w_{ji} \times w_{si}}{\sqrt{\sum w_{ji}^2} \times \sqrt{\sum w_{si}^2}} \tag{13}$$

其中，$num(s)$，$num(s_n \cap s)$ 分别表示集合 s 和集合 $s_n \cap s$ 中元素的个数。γ 是一个可调整因子，为了使得计算结果更加精确，只要新报道 s_n 和子集 s 的交集为1个元素，我们就将其列入计算范围，即预定义 $\gamma=1$。w_{ji} 表示话题 t_j 中术语 k_i 的权重，w_{si} 表示 s 中术语 k_i 的权重。术语集合 S 的任何子集 s 出现的概率是相等的，假设集合 S 中有 t 个术语，则其子集的个数为 2^t，故：

$$P(s) = \left(\frac{1}{2}\right)^t \tag{14}$$

4 时序权重及动态更新

如果新报道 s_n 和话题 t_j 相关，则话题 t_j 将会按照如下步骤动态更新：

步骤1　如果 $k_p \in s_n$ 且 $k_p \notin t_j$，直接将特征 k_p 插入话题 t_j，其权重仍然按照式(5)计算。特征 k_p 的最早发生时间 T_{pb} 等于其最近发生时间 T_{pe}，为新报道 s_n 的发生时间。

步骤2　如果 $k_p \in s_n$ 且 $k_p \in t_j$，特征 k_p 的最近发生时间修改为新报道 s_n 的发生时间，$\Delta T_p = T_{pe} - T_{pb}$ 为特征 k_p 的最新时间距离。若 ΔT_p 大于时间距离阈值 α，则 k_p 的权重按照步骤3进行衰减；若 ΔT_p 小于阈值 α，则特征 k_p 的权重依据式(17)修改为新的时序权重 $(w(k_p))_t$。ΔT_p 越小则意味着特征 k_p 出现得越频繁，其时序权重的提高越大，即：

$$(w(k_p))_t \propto \frac{1}{\Delta T} \tag{15}$$

式(16)是时序权重计算式：

$$(w(k_p))_t = w(k_p) + \frac{\mu}{\Delta T} \tag{16}$$

其中，μ 是一个调节因子，用以保证时序权重的值介于0到1之间。式(16)可以被规范化为：

$$(w(k_p))_t = \frac{w(k_p) + \frac{\mu}{\Delta T}}{\max((w(k_i)))_t}, 1 \leqslant i \leqslant n \tag{17}$$

$\max((w(k_i)))_t$ 表示话题 t_j 中最大的时序权重。

步骤3　当新报道 s_n 中的所有特征被处理完之后，计算话题 t_j 中剩余特征 k_r $(1 \leqslant r \leqslant l)$ 的最近发生时间和新报道 s_n 发生时间的差值 ΔT_r，如果 ΔT_r 大于阈值 α，则意味特征 k_r 在较长时间没有出现，其权重应该衰减。式(18)是指数衰减函数：

$$(w(k_r))_t = (w(k_r))_t \times \exp(-\lambda \Delta T_r) \tag{18}$$

其中，$\lambda \geqslant 0$ 是可调节参数，$(w(k_r))_t$ 表示特征 k_r 的时序权重。

步骤4　观察话题 t_j 中所有特征 k_i $(1 \leqslant i \leqslant m)$ 的权重，若权重小于阈值 β，则认为特征 k_i 为冗余信息，将其从话题 t_j 中删除。

5 实验

5.1 测试集合

美国语言数据联盟(Linguistic Data Consortium)提供了TDT4测试集支持话题识别与追踪领域的相关研究，该测试集合的数据来源于20个新闻源(APW、NYT、CNN、ABC等)，涉及3种语言：美国英语、汉语、阿拉伯语，时间上跨越4个月，共包括98245个报道，对40个话题进行了标注。本文采用TDT4中的汉语部分作为实验测试集。

5.2 实验结果

实验包括3部分：前两个实验用于获得基于TDT语料的时间距离阈值 α 和判断是否为冗余信息的阈值 β，最后一个实验比较了各种模型使用时序权重前后的性能。

5.2.1 时间距离阈值 α

TDT4语料的时间跨度为4个月，如果时间粒度过粗，例如“月”、“年”，则很难找到合适的时间距离；如果时间粒度过细，则会大幅度提高寻找最优时间距离的计算量，所以第一项实验采用的时间粒度为“天”。

若话题 t_j 中剩余特征 k_r 的最近发生时间与新报道 s_n 的发生时间之差大于阈值 α，则 k_r 的权重将会衰减。图2展示了当时间距离为1天，2天，…，120天时，话题模型的最优追踪性能曲线。观察图2可以发现，当时间距离为16天时，最优 $(C_{det})_{norm} = 0.1473$，值最小，即追踪性能最优，由此我们预定义阈值 $\alpha = 16$。

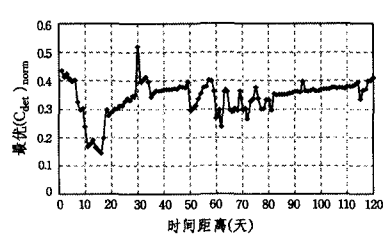


图2　不同时间距离下的最优 $(C_{det})_{norm}$

若时间距离小于16天，则会提高系统的漏报率。例如，TDT4语料中编号为41018(开罗的阿拉伯联盟首脑会议)的话题，若时间距离为14天，则特征“和平”的权重将会从0.4163衰减至0.3627，相应地，新的相关报道 docno = VOM20001022.1800.0067 fileid = 20001022_1800_1900_VOA_MAN.txt 和话题41018的相关度从0.5876降至0.5463，从而提高了将该相关报道定为不相关报道的可能性。相反地，如果时间距离大于16天，则会提高误报的风险。

此外，本实验中的 $\alpha = 16$ 仅适合于实验中的TDT4语料。实际上，不同题材的新闻也会影响 α 的取值，TDT4中的话题基本上都属于政治类话题，例如：塔利班袭击结束和平、阿拉伯财政部长在开罗会议、埃及和伊拉克恢复外交关系、象牙海岸的麻烦等，若是娱乐类话题则其最优时间距离可能会缩短。

5.2.2 权重阈值 β

当发现新的相关报道 s_n 时，话题模型将依据第 4 节给出的步骤进行动态更新，更新完成后观察该话题所有特征的权重，若特征 k_s 的权重小于阈值 β，则 k_s 对话题 t_j 具有较弱的描述能力，为提高系统性能，应将其作为冗余信息删除。通过上文的归一化处理，可知特征权重介于 0 到 1 之间，所以阈值 β 的取值范围介于 0 到 1。表 2 列出了反复实验中，不同阈值下[25]模型的最优性能，当 β=0.2 时，最优 $(C_{det})_{norm}$=0.1547，值最小。0.2 不一定是 β 的最优值，不过此项实验可以确定 β 的最优取值介于 0.2 和 0.3 之间，还可以进一步细分，通过相同的实验获得更为精确的 β，文章暂定阈值 β 为 0.2。

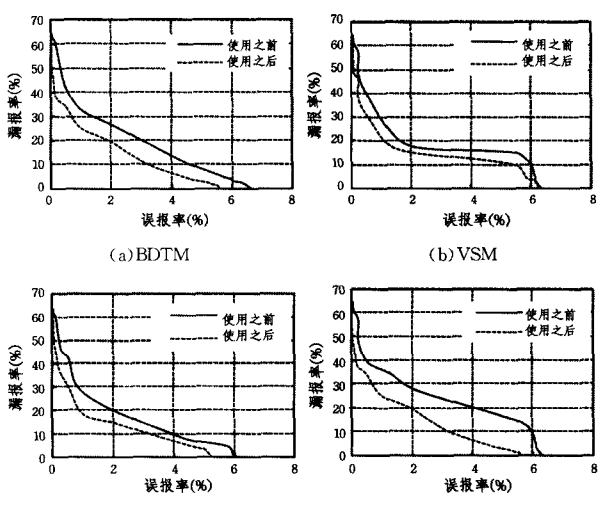
表 2 不同阈值 β 对应的最优性能

β	P_{miss}	P_{fa}	$(C_{det})_{norm}$
0.1	0.0517	0.0241	0.1698
0.2	0.0513	0.0209	0.1547
0.3	0.0499	0.0263	0.1788
0.4	0.0352	0.0255	0.1602
0.5	0.0547	0.0232	0.1684
0.6	0.0519	0.0243	0.1710
0.7	0.0521	0.0249	0.1741
0.8	0.0547	0.0239	0.1718
0.9	0.0549	0.0261	0.1828
1.0	0.0539	0.0259	0.1808

5.2.3 时序权重有效性验证

在不同阈值下，我们首先获得了所提出的 BDTM 使用时序权重前后的漏报率和误报率，并绘制了 DET 曲线[26]，图 3(a)示出该项实验结果。实线表示使用时序权重之前的性能，虚线表示使用时序权重之后的性能。由于 DET 曲线上的点越靠近坐标原点其性能越好，因此由图 3(a)可以看出，使用时序权重后提高了 BDTM 的追踪性能。

理论上，时序的确在话题追踪中提供了有用的信息。本文提出的时序权重是对动态话题追踪模型的进一步优化，可以应用于现有的所有动态话题追踪模型，图 3(b)－(d)验证了时序权重在向量空间模型(VSM)、语言模型(LM)和层次模型(HM)上的有效性。



(a)BDTM (b)VSM

(c)LM (d)HM

图 3 不同模型使用时序权重前后模型性能的比较

结束语 时序权重的应用，可有效提高系统的追踪性能。未来的工作主要包括以下几项：

(1)考虑到信息检索和话题识别与追踪的共性，将贝叶斯信念网络应用于话题建模具备可行性，在后期的工作中我们将进一步挖掘、验证该模型在话题识别与追踪领域的优势。

(2)实验中阈值 α 的值为 16 天，仅限于文中用到的 TDT 语料。实际上，不同题材的新闻，其时间距离应该有所不同。娱乐类新闻其持续时间一般比较短，故 α 值较小；政治类新闻其持续时间比较长，故 α 值应比较大。在后期的研究中我们将采用一定的措施，量化新闻题材和阈值 α 的关系。

(3)从话题追踪的应用领域来看，它需要更为准确的追踪结果。时序权重的应用可以提高追踪结果的准确性，属于优化手段，在未来的工作中，我们将继续寻找其它优化措施，进一步降低话题追踪中的误报率。

参 考 文 献

[1] Allan J, Papka J R, Lavrenko V. On-Line new event detection and tracking[C]// the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998. New York: ACM, 1998: 37-75

[2] Larkey L S, Feng F F, Connell M, et al. Language-specific models in multilingual topic tracking[C]// Proceedings of the 27th Annual Int'l Conference on Research and Development in Information Retrieval, 2004. New York: ACM, 2004: 402-409

[3] Nallpati R. Semantic language models for topic detection and tracking[C]// Proceedings of the HLT-NAACL 2003 Student Research Workshop, 2003. USA, 2003: 1-6

[4] 洪宇，张宇，范基礼，等. 基于语义域语言模型的中文话题关联检测[J]. 软件学报，2008，19(9)：2265-2275

[5] 于满泉，骆卫华，许洪波，等. 话题识别与跟踪中的层次化话题识别技术研究[J]. 计算机技术与发展，2006，43(3)：489-495

[6] 赵华，赵铁军，于浩，等. 面向动态演化的话题检测研究[J]. 高技术通讯，2006，12(16)：1230-1235

[7] 洪宇，仓玉，姚建民. 话题跟踪中静态和动态话题模型的核捕捉衰减[J]. 软件学报，2012，23(5)：1100-1119

[8] Thomas T, Stumpf M P. Inference of temporally varying Bayesian networks[J]. Systems Biology, 2012, 28(24): 3298-3305

[9] 邓冬梅，朱建，陈端兵，等. 时序阵发性对信息传播的影响[J]. 计算机科学，2013，40(11A)：26-28

[10] Yang Y, Pierce T, Carbonell J. A study on retrospective and On-Line Event detection[C]// the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998. New York: ACM, 1998: 28-36

[11] 贾自艳，何清，张俊海，等. 一种基于动态进化模型的事件探测和追踪算法[J]. 计算机研究与发展，2004，41(7)：1273-1280

[12] 赵华，赵铁军，张姝，等. 基于内容分析的话题检测研究[J]. 哈尔滨工业大学学报，2006，10(38)：1740-1743

[13] 洪宇. 基于语义结构和时序特征的话题检测与跟踪技术研究[D]. 哈尔滨：哈尔滨工业大学，2009

[14] Bao L L, Wen J L, Qin L. Enhancing topic tracking with temporal information[C]// the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006. New York: ACM, 2006: 667-668

(下转第 240 页)

之后，能够分类正确，因此召回率提高得很明显。由于“Origin”和“Destination”这两类角色数量不多，因此对整个系统性能（见表2）提高不明显。

表3 Origin&Destination类别未做实体选择操作

	论元识别(%)			论元角色分配(%)		
	P	R	F1	P	R	F1
未加触发词语义特征	44.2	59.2	50.6	40.4	56.3	47
新增触发词语义特征	—	—	—	46.1	57.7	51.3 (+4.3)

结束语 本文在预处理阶段根据实体语义进行实体挑选，降低候选实体正负比率不均衡性；在论元识别阶段将候选论元按照角色语义分为4类进行分类训练、分类抽取，使得论元识别和论元角色分配性能分别提高了3.7%和3.9%；最后针对“Origin”和“Destination”两类角色错误率偏高的情况，加入了触发词语义信息。结果表明，论元抽取性能在加入角色、实体和触发词的语义信息之后有较大的提高。下一步的工作将考虑如何利用一些全局的特征进行推理，以获得更好的性能。

参 考 文 献

[1] Grishman R, Westbrook D, Meyers A. NYU's English ACE 2005 System Description[C]// Proceedings of ACE 2005 Evaluation Workshop. Gaithersburg, MD, 2005

[2] Chun Hong-woo, Hwang Young-sook, Rim Hae-chang. Unsupervised Event Extraction from Biomedical Text Based on Event and Pattern Information [C]// Proceedings of Computational Linguistics and Intelligent Text Processing. Volume 2945, 2004: 533-536

[3] Chun Hong-woo, Hwang Young-sook, Rim Hae-chang. Unsupervised Event Extraction from Biomedical Literature Using Co-occurrence Information and Basic Patterns [C]// Proceedings of UCNLP 2004. Springer Berlin Heidelberg, 2005: 777-786

[4] Chambers N, Jurafsky D. Template-based Information Extraction without the Templates[C]// Proceedings of ACL. 2011: 976-986

[5] Ahn D. The Stages of Event Extraction [C]// Proceedings of the Workshop on Annotations and Reasoning about Time and Events. 2006: 1-8

[6] Hardy H, Kanchakouskaya V, Stzalkowski T. Automatic Event Classification Using Surface Text Features[C]// Proceedings of AAAI06 Workshop on Event Extraction and Synthesis. Boston, MA, 2006

[7] Hong Yu, Zhang Jian-feng, Ma Bin, et al. Using Cross-Entity Inference to Improve Event Extraction[C]// Proceedings of ACL. 2011: 1127-1136

[8] Liao Shas-ha, Grishman R. Using Document Level Cross-Event Inference to Improve Event Extraction [C] // Proceedings of ACL. 2010: 789-797

[9] Huang Rui-hong, Riloff E. Peeling Back the Layers: Detecting Event Role Fillers in Secondary Contexts[C]// Proceedings of ACL. 2011: 1137-1147

[10] 赵妍妍，秦兵，车万翔，等. 中文事件抽取技术研究[J]. 中文信息学报，2008，22(1)：3-8

[11] 谭红叶. 中文事件抽取关键技术研究[D]. 哈尔滨：哈尔滨工业大学，2008

[12] Li Pei-feng, Zhou Guo-dong, Zhu Qiao-ming, et al. Employing Compositional Semantics and Discourse Consistency in Chinese Event Extraction[C]// Proceedings of EMNLP. 2012: 1006-1016

[13] 侯立斌，李培峰，朱巧明. 基于CRFs和跨事件的事件识别研究[J]. 计算机工程，2012，38(24)：191-195

[14] Huang Yuan, Li Pei-feng, Zhu Qiao-ming . Chinese Argument Extraction Based on Trigger Mapping [C] // Proceedings of NLPCC. Springer Berlin Heidelberg, 2013: 41-49

[15] Zheng Chen, Heng Ji. Language Specific Issue and Feature Exploration in Chinese Event Extraction [C] // Proceedings of NAACL. 2009: 209-212

[16] Fu Jian-feng, Liu Zong-tian, Zhong Zhao-man. Chinese Event Extraction Based on Feature Weighting[C]// Proceedings of Information Technology Journal. 2010: 9: 184-187

[17] Li Pei-feng, Zhu Qiao-ming, Zhou Guo-dong. Argument Inference from Relevant Event Mentions in Chinese Argument Extraction[C]// Proceedings of ACL. 2013: 1477-1487

[18] 侯立斌. 中文事件抽取与缺失角色填充的研究[D]. 苏州：苏州大学，2012

（上接第236页）

[15] Kimura M, Saito K, Ohara K, et al. Opinion formation by voter model with temporal decay dynamics[J]. Machine Learning and Knowledge Discovery in Databases, 2012, 7524: 565-580

[16] 廖君华，孙克迎，钟丽霞. 一种基于时序主题模型的网络热点话题演化分析系统[J]. 图书情报工作，2013，57(9)：96-102

[17] Kaur K. Topic tracking techniques for natural language processing[C]// ACAI '11 Proceedings of the International Conference on Advances in Computing and Artificial Intelligence. 2011: 65-71

[18] 张晓艳，王挺，梁晓波. LDA模型在话题追踪中的应用[J]. 计算机科学，2011，38(10A)：136-152

[19] The National Institute of Standards and Technology (NIST). The 2005 Topic Detection and Tracking (TDT2005) Task Definition and Evaluation Plan [Z]. ftp://jaguar.ncsl.nist.gov//tdt/tdt2005/. Eval. Plan. Vllps

[20] 骆卫华，刘群，程学旗. 话题检测与跟踪技术的发展与研究[J]. 全国计算语言学联合学术会议，2003，10(38)：1740-1743

[21] de Cristo M A P, Calado P P, de Lourdes da Silveira M, et al. Bayesian belief networks for IR[J]. International Journal of Approximate Reasoning, 2003, 34: 163-179

[22] Namsrai E, Munkhdalai T, Li M J, et al. A feature selection-based ensemble method for arrhythmia classification[J]. Journal of Information Processing Systems, 2013, 9(1): 31-40

[23] 朱靖波，陈文亮. 基于FIFA的主题相似性计算模型[J]. 东北大学学报：自然科学版，2003，24(11)：1041-1044

[24] Dash S K, Reddy K S, Pujari A K. Adaptive Naive Bayes method for masquerade detection[J]. Journal of security and communication networks, 2011, 4(4): 410-417

[25] 张晓艳. 新闻话题表示模型和关联追踪技术研究[D]. 长沙：国防科技大学，2010

[26] Martin A, Doddington G, Kamm T, et al. The DET curve in assessment of detection task Performance[C]// Eurospeech, 1997. 1997: 37-41