基于相关性模型的中文话题跟踪研究*

郑伟，张宇，邹博伟，洪宇，刘挺

哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001

E-mail: {zw, zhangyu, bwzou hy, tliu}@ir.hit.edu.cn

摘要：作为话题检测与跟踪的重要研究子课题，话题跟踪针对特定话题，识别后续信息流中的相关报道。针对话题本身的漂移现象，本文基于改进的相关性模型，对跟踪中伪相关反馈包含的新颖信息进行检测和建模，并在此基础上动态调整话题空间，跟踪话题漂移，降低漏检率。实验采用 TDT4 语料中文资源及 TDT2003 的评测标准，结果验证此方法可以有效地改进话题跟踪的效果。

关键词：话题跟踪；相关性模型；向量空间模型；新颖检测

Research of Chinese Topic Tracking Based on Relevance Model

Wei Zheng, Yu Zhang, Bowei Zou, Yu Hong, Ting Liu

Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001

E-mail: {zw, zhangyu, bwzou hy, tliu}@ir.hit.edu.cn

Abstract: As an important subtask of topic detection and tracking, topic tracking identifies and collects relevant stories on certain topics from information stream. To find and track topic shift in topic tracking task, this paper proposes the improved relevance model to detect the novelty information in topic tracking feedback and modifies topic model based on this novelty information. This method can track the topic shift and decrease high miss rate in topic tracking. This paper u the Chinese source in TDT4 and the TDT2003 evaluation criterion, the result proves this approach can improve the effect of topic tracking.

Keywords: Topic Tracking; Relevance Model; Vector Space Model; Novelty Detection

1 前言

随互联网信息的膨胀，用户难以从众多信息中快捷地获取自己需要的信息。话题跟踪（Topic Tracking Task，TTT）作为话题检测与跟踪（Topic Detection and Tracking，TDT）的重要子课题，任务是跟踪预先给定的话题相关的后续报道，从而帮助人们把分散的信息按照话题组织起来。

在TDT中，话题定义为由一个种子事件或活动以及与其直接相关的事件或活动组成的报道集合。因此，若报道论述的事件与话题的种子事件有直接联系，则其与话题相关。而随时间发展，媒体对某话题报道的侧重点会发生变化。例如：某话题描述2000年韩国总统金大中获得诺贝尔和平奖。颁奖前，报道侧重对诺贝尔奖得主的猜测上；颁奖时集中于颁奖情况的描述；颁奖后侧重揭示金大中获奖原因以及韩国和朝鲜对其获奖的反应上。这些报道都与话题相关，因此话题存在随时间逐渐漂移的现象。而原始话题模型无法跟踪到漂移现象，需要利用后续报道不断检测相关而新颖[1][2]（Novelty）的信息对话题模型进行调整，同时屏蔽话题模型更新过程中引入的噪声。

2 研究现状

在话题跟踪任务中，如何利用伪相关反馈报道对话题进行调整以跟踪话题漂移是一个重要的研究课题。UMass[3]的跟踪系统基于统计策略计算话题模型与报道流的相关度并据此判断话题的

*本文承国家自然科学基金项目（基金号：60435020，60575042，60503072）的资助

相关报道，然后将相关报道嵌入话题模型并改进其特征权重，从而实现自学习功能。但此方法对跟踪反馈不加任何鉴别地用于话题模型的更新，引入大量噪声，导致对话题的错误调整。LIMSI[4]通过在调整话题过程中设置阈值对伪相关反馈的报道进行二次截取的方法减少噪声的引入。

东北大学[5]利用初始追踪器与后续报道计算相似度，利用相关报道生成新的弱追踪器，并利用报道距离先验知识的时间差对新追踪器的权重进行衰减，再将所有追踪器融合成强追踪器。此方法通过追踪器的动态构建与合并追踪话题的漂移，并保证话题模型核心的恒定。但只利用时间信息对追踪器价值进行判定，而没有基于内容的相关性和新颖性衡量其价值。

基于上述方法的不足，本文将相关性模型[6]应用到话题跟踪中，此方法基于报道内容的相关性对其在话题调整时作用进行衡量。针对其缺点，本文提出基于向量空间的相关性模型。其次，本文采用话题核心与新颖部分分离策略，利用伪相关反馈中对话题漂移有益的新颖信息并屏蔽噪声。本文组织如下：第3节介绍相关性模型，并提出基于向量空间相关性模型概念；第4节论述基于话题核心与新颖部分划分的话题漂移跟踪策略；第5节介绍实验及结果分析；第6节总结与展望。

3 基于向量空间模型的相关性模型

针对话题漂移现象，本文采用相关性模型（Relevance Model，RM）基于伪相关反馈与话题内容相关性对话题进行自适应调整。针对RM存在的对词重要性评定不全面以及对话题调整力度不够的缺点，本文提出了基于向量空间改进的相关性模型（VSM based Relevance Model，VRM）。

3.1 相关性模型

RM 最初用于关联检测任务，对任意给定的两篇报道是否相关于同一话题进行判定。RM 利用待测报道的相关报道内容及相似度构建待测报道的话题模型，通过比较话题模型判定两报道是否相关。此方法通过伪相关反馈与话题内容的相似度决定其对话题建模的重要性，使得构建话题更加准确。由于话题跟踪本质也是将构建话题的报道与后续报道进行相关性判定，即关联检测任务是话题跟踪的本原问题，因此本文探讨将 RM 应用到话题跟踪中。RM 构建话题模型的公式为：

$$P(w \mid R_Q) = \sum_{D \in R_Q} P(w \mid D) P(D \mid Q) \qquad \text{（公式 1）}$$

其中，R_Q 为与报道 Q 相关的报道集合（包括 Q 本身）；D 为 R_Q 中的报道；$P(D|Q)$为 Q 产生 D 的概率，是采用贝叶斯变换，利用词概率的连乘$\Pi P(q_i|D)$（其中 q_i 是 Q 中包含的词）得到。

$$\begin{aligned} p(w \mid D) &= \theta P_{ml}(w \mid D) + (1-\theta) P_{bg}(w) \\ &= \theta \frac{tf_{w,D}}{|D|} + (1-\theta) \frac{cf_w}{coll.size} \end{aligned} \qquad \text{（公式 2）}$$

公式 2 计算报道产生词 w 的概率 $P(w|D)$，其中 $tf_{w,D}$ 是词 w 在报道 D 中出现的次数，cf_w 是词在背景语料集 $coll$ 中的出现次数，公式中的第二项作为对词出现概率的平滑，θ 是平滑系数。

相关性模型参考报道与话题的相关度描述其对话题调整过程产生的不同影响，但存在如下缺点：首先，公式 1 利用语言模型计算话题模型产生报道的概率 $P(D|Q)$，而 $P(D|Q)$通过词概率的连乘得到，其指标往往很小，因此利用相关报道对话题的调整力度有限，对话题的影响很小，无法达到有效调整话题以跟踪话题漂移的目的。此外，概率值 $P(D|Q)$随报道和话题长度的变化会发生显著变化，无法以统一标准描述不同话题和不同报道的相似度，因此需要对报道和话题长度进行统一。而对长度进行限制对于篇幅较长的报道往往遗漏重要特征，对于篇幅较短的报道则需要进行平滑，但平滑往往在泛化重要特征的同时引入噪声。同时，利用公式 2 计算 $P(w|D)$只利用 D 中包含特征 w 的频率信息并利用背景语料进行平滑得到，而没有有效利用词在语料中的 idf 值，因此词的价值评定不全面，往往给停用词赋予较大的权重作为噪声嵌入话题模型。

3.2 基于向量空间模型的改进相关性模型

针对相关性模型的上述缺陷，本文提出了基于向量空间模型的改进相关性模型（VSM based Relevance Model，VRM）代替 RM 模型。VRM 中利用伪相关反馈对话题调整的公式如下：

$$W_{i,T} = \sum_{Sj \in RT} W_{i,Sj} Sim(S_j, T) \qquad \text{（公式 3）}$$

其中 T 为话题，R_T 是与 T 相关的报道集合（包含 T 本身）；$Sim(S_j,T)$是对公式 1 中 $P(D|Q)$ 的替换，采用向量空间模型的余弦夹角公式计算得到的报道 S_j 与 T 的相关度，具体见公式 4；$W_{i,Sj}$ 为词 i 在报道 S_j 中的权重，通过计算词的 $tf*idf$ 得到，用来代替公式 1 中的 $P(w|D)$，可以更好地刻画词在报道中的重要度；$W_{i,T}$ 为词 i 在经过自适应调整后的话题中的新权重。向量空间模型中，两文档相似度计算公式为：

$$Sim(\alpha, \beta) = \frac{\sum_{i=1}^{t} W_{i,\alpha} \times W_{i,\beta}}{\sqrt{\sum_{i=1}^{t} W_{i,\alpha}^2 \times \sum_{j=1}^{t} W_{j,\beta}^2}} \qquad \text{（公式 4）}$$

其中 α 和 β 分别是两个待测的向量空间模型，$W_{i,\alpha}$ 和 $W_{j,\beta}$ 为两向量空间模型中词的权重，$Sim(\alpha, \beta)$ 为 α 与 β 的相似度。如果相似度大于阈值 λ 则认为 α 与 β 是相关的。

采用公式 3 代替公式 1 有如下优点：首先，$Sim(\alpha, \beta)$通过权重乘积的连加得到，其粒度相比于使用词概率连乘得到的 $P(D|Q)$值更适于更新话题空间，可以敏感地识别话题的漂移。其次，概率值 $P(D|Q)$随报道和话题长度的变化会发生显著变化，无法以统一标准描述不同话题和不同报道的相似度。而 $Sim(\alpha, \beta)$取值范围在 0 到 1 之间，不会随报道及话题长度的不同而急剧变化，不需要对话题与报道长度进行限制，可以准确刻画话题与报道之间的相似度，有利于利用统一标准判定报道与话题是否相关。

4 基于话题核心与新颖部分的话题跟踪

由于伪相关反馈中存在一些误判，在话题调整过程中会引入噪声信息，这些噪声会导致更多误判而形成错误的累积。因此在对话题进行调整的同时需要减小引入的噪声对话题的影响。

基于上述问题，本文提出基于话题核心与新颖部分的话题跟踪系统（Core and Novelty based Topic Tracking System，简称 CNTTS）。CNTTS 将话题模型划分为核心（Core）与新颖（Novelty）两部分。其中，Core 由训练集合中事先给定的与话题相关的报道训练得到，是对主题思想的核心描述，在整个话题跟踪过程中恒定不变。话题的 Novelty 由伪相关反馈中的相关报道构建，是对新出现与话题相关的事件描述，用于跟踪话题漂移现象，其在跟踪过程中由伪相关反馈动态调整。话题的 Core 与 Novelty 共同构成了话题模型。由于话题包含有动态调整的 Novelty，使话题跟踪过程中可以有效跟踪话题漂移的内容，而通过 Core 保持先验给定的主题思想不变，因此话题模型不仅具备可扩展性，并可以防止动态调整 Novelty 时引入的噪声干扰话题核心思想的正确描述。CNTTS 系统的框架图如下：

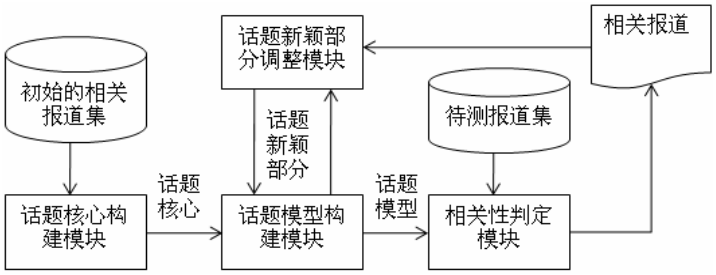

图 1 CNTTS 系统框架图

如图 1 所示 CNTTS 系统共分为 4 个模块。其中，话题核心构建模块利用事先给定的相关报道集构建话题的核心部分；话题新颖部分调整模块利用伪相关反馈的相关报道构建话题的新颖部分并在后续跟踪中不断对新颖部分进行动态调整；话题模型构建模块利用话题的核心与新颖部分对话题模型进行融合；相关性判定模块利用 VSM 相似度计算公式 4 计算话题与报道的相似度，并设置阈值 λ，如果相似度大于 λ 则认为此报道与话题相关。

4.1 话题核心的构建

TDT 的话题跟踪任务中，话题由给定的 N_t 个相关报道描述，因此首先需要通过训练报道构建话题模型，本文选取 $N_t=4$。由第一节分析了解到，TDT 给定的相关报道涉及话题若干侧重点，各侧重点可能各不相同，相似度不大，但其对话题的跟踪具有同样的重要性。所以直接将 4 篇相关报道的向量空间模型融合而构建话题，即把公式 3 中的 $Sim(S_j,T)$设为 1 从而生成话题模型。同时由于给定的相关报道对话题跟踪具有重要作用，因此本文把初始构建的向量空间模型作为话题的核心部分保留，后续对话题的调整是对话题新颖部分的调整，并不改变核心部分。保留话题核心的稳定可以防止伪相关反馈中引入噪声而对话题最重要内容的错误调整，避免错误累积现象。

4.2 利用改进相关性模型调整话题的新颖部分

保持话题核心稳定的前提下，需要对话题进行适当的调整以追踪话题的漂移。话题新颖部分是由伪相关反馈报道构建并在追踪过程中动态调整，功能就是对话题的漂移进行追踪。话题的新颖部分与核心部分共同组成话题的完整描述。

CNTTS 对话题模型进行调整的过程中，利用后续报道中与话题相关的报道对话题的新鲜部分进行动态调整。由于对与话题相关的某一事件或侧重点的报道往往集中于特定时间段内并频繁出现，而在该时间段之外的分布则很稀疏[7]，因此本文采用的话题调整策略是每判断一篇相关报道则立即用此报道对话题新颖部分进行调整，以便更加灵敏地跟踪话题漂移。对于跟踪系统反馈的每篇伪相关报道，CNTTS 利用 VRM 模型的公式通过伪相关反馈的内容及其与话题空间的相似度对新颖部分进行更新，而不是 VRM 中对话题直接调整，对新颖部分的更新公式如下：

$$W_{i,N'} = W_{i,N} + W_{i,S_j} Sim(S_j, T) \qquad \text{（公式 5）}$$

其中 T 是话题模型，由核心与新颖部分 N 组成。通过公式 5 直接把词 i 在报道 S 中的权重 $W_{i,Sj}$ 乘以报道与话题相似度，加上词 i 在原话题新颖部分 N 中的权重 $W_{i,N}$，结果作为词 i 在调整后的新颖部分 N'中的权重 $W_{i,N'}$。其中报道与话题的相似度 $Sim(S_j,T)$由下节的公式 6 计算得到。

4.3 话题模型的构建

本文将话题分为核心部分与新颖部分，其中核心部分由先验相关的训练语料组成，用于描述话题中的种子事件，在跟踪过程中不发生变化；新颖部分描述新出现的与种子事件直接相关的后续事件，根据上节所述方法利用后续报道对话题模型进行动态调整。当计算报道与话题的相关度时，CNTTS 将报道与话题核心及新颖部分分别计算相关度，利用相关度的线性加权和描述报道与整体话题模型的相关性。如果相关度大于阈值 λ 则认为报道与话题相关，此时利用上节所述方法对话题新颖部分进行调整。报道与话题的相关度计算公式如下：

$$Sim(S,T) = \alpha Sim(S,C) + (1-\alpha) Sim(S,N) \qquad \text{（公式 6）}$$

其中 $Sim(S,C)$是报道 S 与话题核心 C 的相似度，$Sim(S,N)$是 S 与新颖部分 N 的相似度，α 是刻画话题核心在话题模型中所占比例的系数，由第 5 节实验得知α 取 0.5。$Sim(S,T)$是 S 与话题 T 的相似度，如果大于阈值 λ 则认为报道与话题相关，则利用上节所述方法对新颖部分进行调整。

5 实验及结果分析

5.1 实验语料及评测机制

本文实验采用 TDT4 语料及 TDT2003 的评测方法[8]对话题跟踪进行评测。其中 TDT4 包含从 2000 年 10 月到 2001 年 1 月的报道，N_t=4 的中文语料评测给出 54 个待测话题，本实验选取其中的 TDT2002 的 10 个话题做训练，TDT2003 的 24 个话题做测试。TDT 评测公式定义如下：

$$(C_{Det})_{Norm} = (C_{Miss}P_{Miss}P_{target} + C_{FA}P_{FA}P_{non-target}) / \min(C_{Miss}P_{target}, C_{FA}P_{non-target})$$（公式 7）

其中，C_{Miss}、C_{FA}、；P_{target} 和 $P_{non\text{-}target}$ 是事先定义的值，分别取 1、0.1、0.02 和 0.98；P_{Miss} 和 P_{FA} 是系统漏检率和错检率。（C_{Det}）$_{Norm}$ 是系统性能损耗代价，此值越小则系统性能越好。

5.2 实验结果

本实验以基于相关性模型的话题跟踪系统作为 baseline，考察 VRM 模型和 CNTTS 策略对跟踪系统的影响。实验中以 RM、VRM 和 VRM-CNTTS 分别表示相关性模型系统、改进的相关性模型系统和基于 VRM 模型的 CNTTS 系统。

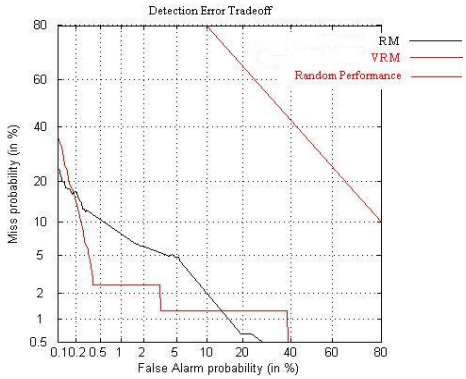

图 2 VRM 模型对话题跟踪的影响

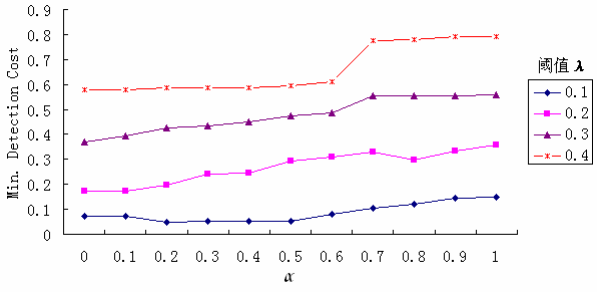

图 3 在不同阈值下 α 取值对实验效果的影响

图 2 是在训练集上考察 VRM 对话题追踪效果的影响。黑线和红线分别是 RM 和 VRM 构建的系统，DET 曲线越靠近左下角效果越好。从图中可以看出，VRM 效果较 RM 有了一定的提高。

图 3 是对 CNNTS 系统的参数 α 的训练，α 是刻画话题核心在话题模型中所占比例的系数。图中，横坐标为 α 的取值；纵坐标为$(C_{Det})_{Norm}$ 指标，曲线上的点越靠下说明效果越好。其中 $\alpha = 0$ 是直接采用改进的相关性模型的话题跟踪系统。每条曲线代表不同的相似度阈值 λ 。实验结果显示相关性阈值较大的情况下，系统的跟踪性能相对偏低。其原因在于阈值较大时，系统无法检测到主题相关但核心内容发生漂移的报道，无法针对漂移现象修正话题模型。此外，当相关性阈值较高时，随着 α 指标的提高，跟踪系统的性能逐渐衰减。其原因在于 α 指标越高，跟踪系统越趋近于将特征空间更近似于核心的报道进行反馈，在调整话题的过程中不断削弱新鲜信息的比重，从而逐步偏离对漂移现象的追踪。当相关性阈值 λ 降低为 0.1 时，跟踪系统的性能随着 α 的增加逐渐提高，当 α 为 0.5 时性能最佳，然后逐渐衰减。跟踪系统性能的这一变化过程近似地反映了报道流的话题漂移现象，即当相关性阈值较

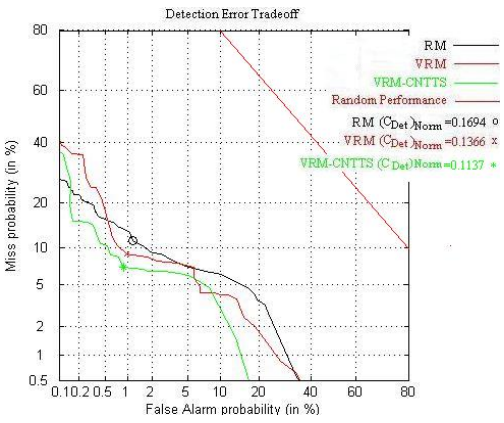

图 4 测试集上的 DET 曲线

低的情况下，系统反馈中包含主题相关但特征空间相似度较低的报道，将这种报道作为新鲜信息，并通过α 调整其在话题空间中的分布，有助于跟踪系统识别和追踪话题漂移现象。话题空间必须保证核心内容的适当分布，过分依赖新鲜信息将造成话题空间的偏差。图 3 显示当 λ 取 0.1 和 α 取 0.5 时系统损耗代价达到最小。因此，在 CNTTS 中选择$\alpha = 0.5$ 。

图 4 给出了在测试集上不同方法的 DET 曲线图。采用 VRM 模型使系统的效果有所提高，在测试集上将系统的性能最小开销从 RM 的 0.1694 降低到 0.1366。在 VRM 基础上使用 CNNTS 方法在测试集上对系统效果有了一定提高，将系统最小开销从采用 VRM 的 0.1366 降低到 0.1137。

5.3 结果分析

从实验图中可以看出，采用改进前的相关性模型性能较低，其原因在于相关性模型对话题调整的幅度较小而无法有效跟踪话题漂移，使得话题跟踪的漏检率较高。而 VRM 模型利用词的 tf*idf 值作为权重，利用报道与话题向量空间的相似度区分不同报道对话题调整的重要性，可以发现报道中的新颖信息并对话题进行适度的调整，有效跟踪了话题漂移。

使用 CNTTS 策略在测试集上使得系统的效果有了一定改善。主要原因是，VRM 利用伪相关报道对话题进行调整以跟踪话题漂移过程中，伪相关反馈中存在很多相关性判定错误的报道，使用此类报道对话题调整的过程中会在话题模型中嵌入噪声信息，会对话题进行错误调整而导致更多的误判。而 CNTTS 策略将话题分为核心与新颖部分，其中动态调整的新颖部分使话题跟踪过程中可以有效跟踪话题漂移的内容，而通过稳定不变的核心保持先验给定的主题思想不变，屏蔽噪声对话题核心内容的改动。因此话题模型不仅具备可扩展性，并可以防止动态调整 Novelty 时引入的噪声干扰话题核心思想的正确描述。

6 未来工作

针对话题跟踪中存在的话题漂移现象，本文探讨将相关性模型应用到话题跟踪任务中的方法，并针对其缺点提出基于向量空间的相关性模型，实验证明此模型可以提高话题跟踪效果。同时针为减少噪声影响，本文提出将话题分为核心与新鲜部分，屏蔽噪声对话题核心内容的影响。结果显示此方法使实验效果有了提高。但在训练集上效果有限，因此在未来的工作中可以尝试使用 TDT5 自适应话题跟踪任务的语料，研究在噪声较多的情况下此方法对话题跟踪效果的影响。

参考文献

[1] X Li, WB Croft. Novelty detection based on sentence level patterns. Proceedings of the 14th ACM international conference on Information and knowledge management. 2005. 744~751.

[2] GPC Fung, JX Yu and PS Yu. Parameter Free Bursty Events Detection in Text Streams. Proceedings of the 31st international conference on Very large data bases. 2005. 181~192.

[3J Allan, V Lavrenko, D Frey, V Khandelwal. UMass at TDT 2000. Proceedings of Topic Detection and Tracking Workshop, 2000.

[4] YY Lo, JL Gauvain. The LIMSI Topic Tracking System For TDT 2002. Proceedings of Topic Detection and Tracking Workshop, 2002.

[5] 王会珍,朱靖波,季铎.基于反馈学习自适应的中文话题追踪.中文信息学报,2006,03:94~100.

[6] V Lavrenko, J Allan, E DeGuzman. Relevance Models for Topic Detection and Tracking. Proceedings of the Human Language Technology Conference. 2002. 104~110.

[7] Yiming Yang, Tom Pierce, Jamie Carbonell, A study on Retrospective and On-Line Event detection, Canegie Mellon University, In the proceedings of SIGIR 1998. 28~36.

[8] http://www.nist.gov/speech/tests/tdt/tdt2003/evalplan.htm