第 35 卷 第 3 期 计 算 机 工 程 2009 年 2 月
Vol.35 No.3 Computer Engineering February 2009
·软件技术与数据库· 文章编号：1000—3428(2009)03—0080—03 文献标识码：A 中图分类号：TP391

一种基于自适应重心向量的主题检测方法

潘 渊，李弼程，张先飞
(解放军信息工程大学信息工程学院，郑州 450002)

摘 要：针对影响主题检测性能的 2 个重要因素——相似主题的判定和主题漂移问题，提出一种基于自适应重心向量的主题检测方法。该方法将命名实体信息应用到特征表示上，将命名实体向量和关键词向量相结合表示主题的重心向量，以有效区分相似主题。采用增量聚类检测主题，在增量聚类过程中不断修正主题重心，以解决主题漂移的问题。实验结果与性能比较表明，该方法能有效提高主题检测的性能。
关键词：主题检测；主题漂移；命名实体；主题重心向量

Topic Detection Approach Based on Adaptive Center Vector

PAN Yuan, LI Bi-cheng, ZHANG Xian-fei
(Institute of Information Engineering, PLA Information Engineering University, Zhengzhou 450002)

【Abstract】Similar topic detection and topic excursion are two important factors which affect the performance of topic detection. For these two problems, this paper proposes a topic detection approach based on adaptive center vector. By using information of name-entity in feature representation, it combines name-entity vector and keyword vector to construct topic center vector, which can detect similar topic efficiently. Based on the idea of single-pass clustering, the algorithm modifies topic center dynamically. Experimental results show that the algorithm can improve the performance of topic detection effectively.
【Key words】topic detection; topic excursion; name-entity; topic center vector

1 概述

互联网的迅猛发展使得网上信息急剧增加，人们很难从海量信息中快捷准确地获取自己感兴趣的信息。主题检测与追踪(Topic Detection and Tracking, TDT)技术正是为了解决这个问题而产生的[1]，旨在依据事件对语言文本信息流进行组织利用的研究。主题检测是 TDT 的一项研究内容，是指将新闻报道、新闻专线等来源的数据流中的报道归入不同的主题并在必要时建立新主题的技术。

影响主题检测性能的一个重要因素是相似主题的判定，例如不同的飞机失事或不同的恐怖袭击。其主要原因是这些相似主题的报道中有大量的相同词汇，容易造成主题误判。文献[2]利用命名实体解决此问题。详细分析可以得知，利用命名实体虽然能在一定程度上区分相似主题，但新闻报道中的命名实体数目有限，仅仅依靠命名实体而放弃描述主题内容的大量其他关键词，必然造成对主题框架概括不全面，从而影响主题检测的性能。

影响主题检测性能的另一个重要因素是主题重心漂移问题。主题是动态演变的，比如主题 A 描述“9·11”事件。最初的报道都是关于飞机撞击世贸大厦的经过、损失和伤亡情况以及劫机恐怖分子的有关情况等。而后报道的重心转移到美国政府的营救情况、民众的反应以及政府打击这次恐怖活动的措施等。对于这样的主题重心漂移现象，如果不对其重心进行自适应修正，将可能无法检测到同一主题的后续相关报道。研究者们对主题动态演化进行了研究，如文献[3]给出了一种动态演化特性的主题定义；文献[4]刻画了主题的动态演化过程。目前，通常用动态调整主题模型的方法[5]解决这一问题。这种动态调整是根据检测系统判断出与该主题相关的文档来重新训练主题重心的，而反馈的文档都是基于初始主题检测模型的判定，一旦反馈的文档是错误的，主题就会产生错误的漂移，对后续报道给出错误的判断，造成主题重心的持续错误漂移。

针对以上 2 个问题，本文首先将命名实体用于主题检测的特征表示上，同时结合关键词向量，提出基于命名实体和关键词向量相结合的特征表示方法，解决相似主题的区分问题。检测方法则采用增量聚类，在聚类过程中，将每次更新后的主题重心与原主题重心进行线性组合，完成对主题重心的不断修正，以解决主题重心漂移的问题。本方法利用 TDT4 中文语料做测试，采用 TDT2004 评测标准，结果表明，新方法可以很好地解决相似主题误判问题以及主题漂移问题，有效提高了主题检测的性能。

2 主题重心向量

新闻文档和主题重心的表示采用向量空间模型，主题重心向量由命名实体向量 $\boldsymbol{V}_{\mathrm{NE}}$ 和关键词向量 $\boldsymbol{V}_{\mathrm{KW}}$ 组成。

2.1 文档向量的构造思想

MUC(Message Understanding Conferences)将命名实体定义为人名、地名、组织机构名、日期等类型。由于新闻报道必须具备 4 个要素：时间，地点，人物，事件，因此本文只提取人名、地名、组织机构名、日期 4 种命名实体。对新闻报道而言，这 4 种命名实体对表达新闻内容的贡献应大于其他特征词，因此，在构造新闻文档向量时，先提取文档中的命名实体，形成命名实体向量，再将抽取命名实体后的特征

基金项目：国家“863”计划基金资助项目(2007AA01Z439)
作者简介：潘 渊(1983－)，男，硕士研究生，主研方向：数据挖掘；李弼程，教授、博士生导师；张先飞，博士研究生
收稿日期：2008-07-04 **E-mail：**panyuan8348@163.com

词表作为关键词向量，并根据 2 个向量对文档内容的贡献大小赋予其不同的权重。理论上，这样构造的文档向量对文档内容的表示比单纯用特征词表及命名实体更准确。

2.2 主题重心向量形成算法

本文采取动态抽取的方式实时抽取主题特征项，即当有新的新闻报道加入主题时，便重新扫描该主题的所有文档，形成新的重心向量。这种动态抽取的方式能更好地体现主题的演变、表示主题的内容。具体步骤如下:

(1)预处理。

扫描一个主题内的所有文档，进行分词、去停用词处理。并将所有词放到一个词表中。本文称其为原始词表。

(2)计算特征评估值。

基于此主题用信息增益法对原始词表中的每个词评估，并根据评估值对词表进行排序。

(3)特征提取。

对排序后的词表采取截断处理，抽取特定数目的词作为特征词，形成最优词表。

(4)拆分特征词表。

从最优词表中提取命名实体形成命名实体词表，剩下的特征词形成关键词表。

(5)特征词权重计算。

权重计算借鉴“$TF\times IDF$”的思想采用“$TF\times INSF$”法。$INSF$ 法如下式所示:

$$INEF_w=\frac{\log_n\left(\frac{N_T+1}{freq_w}\right)}{\log_n(N_T+1)} \tag{1}$$

其中，N_T 为当前主题总数；$freq_w$ 是出现词 w 的主题个数。

(6)形成重心向量。

将主题中的新闻报道分别与命名实体词表和关键词表匹配生成各自的 $\boldsymbol{V}_{\mathrm{NE}}$ 和 $\boldsymbol{V}_{\mathrm{KW}}$。统计主题所有新闻报道中词的权重，根据累加平均算法生成主题的命名实体特征向量和关键词特征向量。累加平均算法如下:

$$weight(w,R_i)=tf_{R_i}\times INEF_w \tag{2}$$

$$Weight(w,t,T)=\frac{\sum_{R_i\in T} weight(w,R_i)}{N} \tag{3}$$

假设 R_i 是一篇已经处理的文档，式(2)中的 tf 表示特征项 w 在 R_i 中的词频；式(3)中的 $Weight(w,t,T)$ 表示特征项 w（$\boldsymbol{V}_{\mathrm{NE}}$ 特征项或 $\boldsymbol{V}_{\mathrm{KW}}$ 特征项）t 时刻在主题 T 中的权重；R_i（$1\leqslant i\leqslant N$）表示 t 时刻主题 T 的第 i 篇新闻报道；N 表示 t 时刻主题 T 中包含的新闻报道篇数；$weight(w)$ 表示特征项 w 在新闻报道 R_i 中的权重。

(7)有新的报道加入主题时，循环执行步骤(1)~步骤(6)。

3 主题重心的修正

新闻主题一般由时间、地点、人物、事件内容等要素构成，具有实时性、动态性等特点。随着主题检测的深入，主题数目会不断增加，涉及的新闻报道也会越来越多。同时关于某主题报道的侧重点会随着事件进程的变化出现越来越多的新词，所以，为了保持主题检测的准确性，应该不断修正主题重心。自适应技术就是从这个角度出发，在检测的过程中进行无监督的学习来修正主题模型。

主题重心修正就是选择能表达主题内容的新闻报道来更新主题重心，并把更新后的主题重心与初始主题重心线性相加。它是基于主题重心向量表示和主题重心更新的一种自适应模型表示，能够避免现有的动态调整主题模型方法对主题模型进行完全更改而造成主题的错误漂移。算法公式如下:

$$\begin{cases} Cg(D)=\sum_{i=1}^{n} t_i Cg(d_i) \\ \sum_{i=1}^{n} t_i<1 \end{cases} \tag{4}$$

其中，n 为主题中新闻报道的数目；$Cg(d_i)$ 为每次更新后的主题重心；t_i 为主题重心的加权系数；$Cg(D)$ 为所有更新后的主题重心线性组合形成的修正后的主题重心。这样，原有的主题并没有舍弃，即使有错误的更新出现，对整个主题检测的影响也比较小，由错误的更新产生的错误不会一直蔓延下去。

用于修正主题重心的新闻报道必须有条件选择。由于一篇文档与主题重心的相似度值越大就越能表示该主题的内容，因此系统需要灵活地调整检测的灵敏度，选择合理的文档来修正主题重心。

本文通过增加创新阈值 t_n 保证修正文档的正确性。通常 $t_c>t_n$（t_c 为聚类阈值），具体做法如下:

设 $s_{\max}$ 表示当前文档 d 与以前主题重心之间的最大相似度值:

(1)如果 $s_{\max}>t_c$，那么这篇文档被标示为“OLD”，将该文档归入最相似的主题中，并将该文档作为修正主题重心的文档。

(2)如果 $t_n\leqslant s_{\max}\leqslant t_c$，那么这篇文档被标识为“OLD”，不做进一步处理。

(3)如果 $s_{\max}<t_n$，那么这篇文档被标示为“NEW”，意味着该文档是表达一个新主题的第 1 篇文档。

4 基于自适应主题重心向量的主题检测算法

一篇新闻报道中命名实体的数量有限，而其他关键词的数目巨大。但命名实体对区分不同主题的贡献大于其他关键词，若不对命名实体进行加权，必将使其对主题的贡献被弱化甚至淹没。本文通过统计得知，一篇新闻报道中的命名实体与其他关键词的数目比大约介于 1:4~1:3，此时其贡献比应取其数目比的反比，并取其折中，实验也证实这样可以取得较好的检测效果。令 α 为命名实体的加权系数，β 为关键词的加权系数，所以，$\alpha:\beta=3.5:1$。

报道和主题的相似度是由报道分别与 $\boldsymbol{V}_{\mathrm{NE}}$ 的相似度和 $\boldsymbol{V}_{\mathrm{KW}}$ 的相似度的加权相加。算法如下式:

$$Sim(R,T)=\alpha Sim(R,\boldsymbol{V}_{\mathrm{NE}})+\beta Sim(R,\boldsymbol{V}_{\mathrm{KW}}) \tag{5}$$

其中，α 和 β 为加权系数；$Sim(R,\boldsymbol{V}_{\mathrm{NE}})$ 和 $Sim(R,\boldsymbol{V}_{\mathrm{KW}})$ 分别表示报道与 2 个重心向量的余弦相似度。

基于自适应主题重心向量的主题检测算法步骤如下:

输入 新闻文档流，创新阈值 t_n，聚类阈值 t_c

(1)算法顺序处理输入的每篇新闻文档，与以前生成的所有主题词表的总主题检测器进行相似性比较，得到最大相似度值 $s_{\max}$。如果 $s_{\max}>t_c$，跳至步骤(2)；如果 $t_n\leqslant s_{\max}\leqslant t_c$，跳至步骤(3)；如果 $s_{\max}<t_n$，跳至步骤(4)。

(2)将该文档归入到最相似的主题中，如果满足主题重心修正规则，根据式(4)修正主题重心，处理下一个新的文本。其中，重心修正规则如下: 1)当该主题中的新闻个数小于 5 时，立即修正主题重心；2)当主题中的新闻个数大于 5 时，

每增加 10 个新闻报道，修正一次主题重心；3)当主题检测算法每处理 200 个新闻报道后，每个主题的主题重心修正一次。

(3)将该文档归入最相似的主题中，处理下一个新的文本。

(4)判定该文档为新的主题，意味着该文档是表达一个新主题的第 1 篇文档，生成表示该主题的主题重心，处理下一个新的文本。

5 实验结果及性能分析

5.1 实验数据

本实验针对 TDT4 语料进行处理。TDT4 共有 98 245 篇报道，其中，中文语料 27 142 篇，分别来自于新华社、联合早报等新闻机构;英文语料有 28 390 篇，分别来自 NYT, CNN, VOA 等新闻机构；其余为非英文报道经机器翻译成英文后的结果。时间跨度为 2000 年 10 月~2001 年 1 月。语料形式为原始的新闻文本或语音转录得到的电视或广播新闻文本。

5.2 实验结果评测标准

在 TDT 的评价标准中，除了采用与传统的信息检索和文本分类类似的正确率 P 、召回率 R 、$F1$-*measure* 来评价结果外，还采用了系统错误率评价结果，主要包括漏报率 P_{miss} 和错报率 P_{FA} 。

此外，本文还使用了 TDT 在 2004 年的评测方法。将漏报率与错报率合并成一个检测开销 C_{Det} 。其计算公式为

$$C_{\text{Det}} = C_{\text{miss}} P_{\text{miss}} P_{\text{target}} + C_{\text{FA}} P_{\text{FA}} P_{\text{non-target}} \quad (6)$$

其中， P_{miss} 为系统的漏报率； P_{FA} 为系统的错报率； C_{miss} ， P_{target} ， C_{FA} ， $P_{\text{non-target}}$ 为事先定义好的值，TDT 在评测时这些参数设定如下： $C_{\text{miss}} = 1.0$, $P_{\text{target}} = 0.02$, $P_{\text{non-target}} = 1 - P_{\text{target}} = 0.98$, $C_{\text{FA}} = 0.1$ 。

为了使得到的性能指标落在更有意义的范围内，将 C_{Det} 规范化得到 $Norm(C_{\text{Det}})$ [6]:

$$Norm(C_{\text{Det}}) = C_{\text{Det}} / \min(C_{\text{miss}} P_{\text{target}}, C_{\text{FA}} P_{\text{non-target}}) \quad (7)$$

可以看出， $Norm(C_{\text{Det}})$ 越小，系统的性能越好。本文实验结果均采用 *Micro Average* 计算各项指标。

5.3 结果和性能分析

表 1 给出了基于增量聚类(SP_D)、基于命名实体识别(NE_D)[2]、基于主题模型动态调整(MDA_D)[5]以及本文基于自适应重心向量(ACV_D)4 种主题检测方法对于 TDT4 中文语料的实验结果。其中，本文方法的创新阈值 t_n 取经验值 0.35；聚类阈值 t_c 取 0.5。

表 1 主题检测实验结果

	漏报率	错报率	召回率	正确率	*F1-Measure*	$Norm(C_{\text{Det}})$
SP_D	0.252 4	0.011 2	0.747 6	0.735 5	0.741 5	0.307 3
NE_D	0.236 8	0.010 4	0.763 2	0.753 5	0.758 3	0.287 8
MDA_D	0.221 4	0.010 0	0.778 6	0.762 0	0.770 2	0.270 4
ACV_D	0.204 6	0.008 3	0.795 4	0.799 7	0.797 5	0.245 3

从表 1 的检测结果可以看出，本文方法 ACV_D 与基于命名实体法 NE_D 相比，准确率提高了 4.62%，召回率提高了 3.22%。其原因在于本方法将命名实体向量和关键词向量结合起来构成主题重心向量，并且针对命名实体与关键词对文本的贡献程度赋予不同的权重，使得对主题的表示更全面、更准确，因此，可以很好地区分相似主题。在解决主题重心的漂移上，本文方法通过对更新前后主题重心的重新线性组合来表示新的主题重心，避免了只采用更新后的主题重心而引起的重心漂移现象，使其准确率比 MDA_D 提高了 3.77%，召回率提高了 1.68%。从图 1 可以看出，ACV_D 的检测开销是 4 种方法中最低的。从图 2 可以看出，当创新阈值 t_n 取经验值 3.5 时，本文方法的准确率最高。

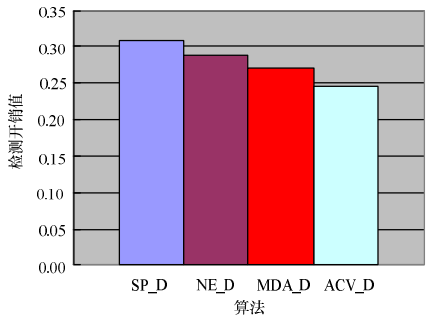

图 1 主题检测开销性能分析

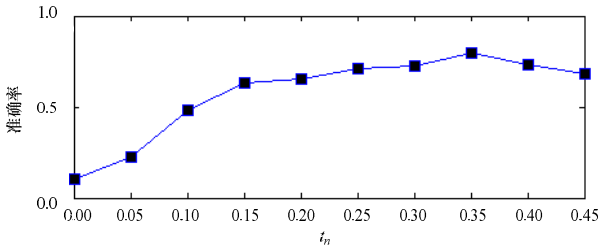

图 2 ACV_D 随 t_n 变化的性能曲线图

6 结束语

本文针对主题检测存在的两大问题，提出一种基于自适应重心向量的主题检测方法。本方法结合命名实体和关键词来表示文本特征，克服了仅仅依靠命名实体或者单纯的特征词来表示文本的局限性，解决了相似主题的区分问题。在检测方法上加入重心更新和线性组合策略对主题重心进行不断修正，解决了主题模型动态调整法容易引起重心漂移的问题。实验结果表明，本方法能有效提高主题检测的性能，降低运算开销。

参考文献

[1] Allan J. Topic Detection and Tracking: Event-based Information Organization[M]. Boston: Kluwer Academic Publishers, 2002: 1241-1253.

[2] Kumaran G, Allan J. Text Classification and Named Entities for New Event Detection[C]//Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. [S. l.]: ACM Press, 2004: 297-304.

[3] Makkonen J. Investigations on Event Evolution in TDT[C]// Proceedings of Student Workshop of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Edmonton, Canada: [s. n.], 2003: 43-48.

[4] Nallapati R, Feng Ao, Peng Fuchun. Event Threading Within News Topics[C]//Proceedings of International Conference on Information and Knowledge Management. Washington, USA: [s. n.], 2004: 446-453.

[5] 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.

[6] 骆卫华, 于满泉. 基于多策略优化的分治多层聚类算法的话题发现研究[J]. 中文信息学报, 2006, 20(1): 29-36.