软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(5):1100−1119 [doi: 10.3724/SP.J.1001.2012.04045]

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

话题跟踪中静态和动态话题模型的核捕捉衰减*

洪 宇+, 仓 玉, 姚建民, 周国栋, 朱巧明

(苏州大学 计算机科学与技术学院,江苏 苏州 215006)

Descending Kernel Track of Static and Dynamic Topic Models in Topic Tracking

HONG Yu+, CANG Yu, YAO Jian-Min, ZHOU Guo-Dong, ZHU Qiao-Ming

(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

+ Corresponding author: E-mail: hongy@suda.edu.cn

Hong Y, Cang Y, Yao JM, Zhou GD, Zhu QM. Descending kernel track of static and dynamic topic models in topic tracking. *Journal of Software*, 2012,23(5):1100−1119. http://www.jos.org.cn/1000-9825/4045.htm

Abstract: Topic tracking is a task in research on identifying, mining and self-organizing relevant information to news topics. Its key issue is to establish statistical models that adapt the kind of news topic. This includes two aspects: one is topical structure; the other is topic evolution. This paper focuses on comparing and analyzing the features of three main kinds of topic models including words bag, hierarchical tree and chain. Different performances of static and dynamic topic models are deeply discussed, and a term overlapping rate based evaluation method, namely descending kernel track, is proposed to evaluate the abilities of static and dynamic topic models on tracking the trend of topic development. On this basis, this paper respectively proposes two methods of burst based incremental learning and temporal event chain to improve the performance of capturing topic kernels of dynamic topic models. Experiments adopt the international-standard corpus TDT4 and minimum detection error tradeoff evaluation method proposed by NIST (National Institute of Standards and Technology), along with descending kernel track method to evaluate the main topic models. The results show that structural dynamic models have the best tracking performance, and the burst based incremental learning algorithm and temporal event chain achieve 0.4% and 3.3% improvement respectively.

Key words: topic tracking; static topic model; dynamic topic model; descending kernel track; busty feature based incremental learning; temporal event chain

摘 要: 话题跟踪是一项针对新闻话题进行相关信息识别、挖掘和自组织的研究课题,其关键问题之一是如何建立符合话题形态的统计模型.话题形态的研究涉及两个问题,其一是话题的结构特性,其二是话题变形.对比分析了现有词包式、层次树式和链式这 3 类主流话题模型的形态特征,尤其深入探讨了静态和动态话题模型拟合话题脉络的优势和劣势,并提出一种基于特征重叠比的核捕捉衰减评价策略,专门用于衡量静态和动态话题模型追踪话题发展趋势的能力.在此基础上,分别给出突发式增量式学习方法和时序事件链的更新算法,借以提高动态话题模型的核捕捉性能.实验基于国际标准评测语料 TDT4,采用 NIST(National Institute of Standards and Technology)提出的最小

* 基金项目: 国家自然科学基金(61003152, 60970057, 60873105, 90920004, 60970056); 国家高技术研究发展计划(863)(2012AA011102); 国家教育部博士点基金(200932011 10006); 苏州市应用基础研究计划基金(SYG201030)

收稿时间: 2010-04-26; 修改时间: 2010-12-15; 定稿时间: 2011-04-28

检测错误权衡系数评测法,并结合所提出的核捕捉衰减评价方法,对各类主要话题模型进行测试.实验结果显示,结构化的动态话题模型具有最佳的跟踪性能,且突发式增量式学习和时序事件链的更新算法分别给予动态话题模型0.4%和3.3%的性能改进.

关键词: 话题跟踪;静态话题模型;动态话题模型;核捕捉衰减;突发式增量式学习;时序事件链

中图法分类号: TP391 **文献标识码:** A

话题跟踪的核心任务是从时序排列的新闻报道流中,实时识别和挖掘相关于特定新闻话题的系列相关报道.一般而言,话题跟踪系统需要包含3项基本组成:话题建模、相关性判定机制和阈值预估.其中,话题建模的目标是建立一种描述新闻话题核心内容的模型,即话题模型;相关性判定机制侧重形成一种话题模型和新闻报道相关性的度量方法;阈值预估则是借助训练样本获取最佳的相关性划分边界.由于话题模型直接决定了话题内容和语义组成的基本架构,使得话题模型一经确定,就基本限定了相关性判定和阈值预估方法的选择范畴,从而话题模型往往成为决定话题跟踪系统性能的关键性因素.

按照话题模型在整个跟踪过程中保持的状态进行划分,现有话题模型主要包含两种:一种是相悖于话题演化规律的静态话题模型(static topic model,简称STM);一种是依附话题演化趋势的动态话题模型(dynamic topic model,简称DTM).静态话题模型强调话题初始核心的守恒性,动态话题模型则注重话题核心随着系列相关事件的出现而产生的演化现象.根据话题的定义,特定话题是由一个种子事件以及后续系列相关事件组成的整体.新闻报道仅仅是1个或多个事件诉诸文字的表现形式.由此,建立静态话题模型的意图在于利用已知的新闻报道样本,挖掘和描述其中蕴含的种子事件,并将其作为恒定的话题核心贯穿整个后续相关报道的识别过程.相对而言,动态话题模型则淡化种子事件的绝对作用,仅将其作为话题结构的子成分之一;话题核心可以借助对后续相关事件的自适应学习,动态地变换至不同子成分,借以追踪话题焦点的演化;尤其是动态话题模型的形态本身也会不断调整和变化.

直觉上,动态话题模型更适应话题发展与演化的规律,应作为话题跟踪系统中话题建模的首选样板.但不容忽视的问题是,静态话题模型的核心守恒性能够保证跟踪过程的收敛性.换言之,因其专注于话题的关键焦点,即种子事件,能够有效避免跟踪过程偏离主线,从而确保较高的跟踪精确率.相比而言,动态话题模型则能融入新的焦点,即后续新颖事件,使得跟踪目标有所扩展,从而确保较高的跟踪召回率.由此,静态与动态话题模型在跟踪过程中孰优孰劣仍是一项值得深入验证的问题.

此外,话题跟踪的主要评价标准是误检率和漏检率,最小检测错误权衡系数则是误检率和漏检率的归一化最小折中值,简称CDet[1].这类评价标准适用于特定跟踪系统整体性能的评价.换言之,是一种针对“黑箱”的评价,跟踪系统内各个模块对评价过程并不透明,即话题建模、相关性匹配、阈值估价和自适应学习模块各自的性能无从评判,仅有“黑箱”输出端的二元分类值(即相关报道或不相关报道)单一地参与评价过程.然而,跟踪系统中的每个模块都有着重要的评估价值.尤其是,话题建模的质量往往直接决定各类新闻信息处理系统的性能.由此,针对话题模型提出一种专有的评价办法,对于现有的话题跟踪研究,以及后续开展的话题演化学习、话题变种检测和预报等研究都具有重要意义.

针对这一问题,本文提出一种核捕捉衰减(capturing kernel attenuation,简称CKA)比对方法,旨在建立一种横向比较各种话题模型捕获后续相关报道能力的数学模型.核捕捉衰减的基本数值计算是话题模型与新闻报道的交叉比,但不局限于某一时段交叉比的纵向指标,而是侧重检验交叉比在整个话题发展过程中的增益或衰减趋势.核捕捉衰减比对则针对各个话题模型的核捕捉趋势进行近似计算,借以用量化的指标区分不同趋势.在此基础上,本文重现了话题跟踪领域中的主要话题建模方法,借助核捕捉衰减比对法,并结合话题跟踪评测体系中通用的检测错误权衡系数(detection error tradeoff,简称DET),对静态和动态话题模型的跟踪性能给予综合评测和比较.此外,本文分别提出基于突发事件的增量式学习和时序事件链的动态自学习方法,两者是根据话题变异属性提出的动态话题模型自学习改进策略.实验验证,两者可分别减少检测错误代价0.4和3.3个百分点.

本文第1节回顾话题跟踪领域的相关研究,重点介绍其中话题建模的主要方法.第2节详细介绍一种基本

的静态话题模型,即词包式无结构话题模型,并列举基于这一模型的后续各类变体.第 3 节侧重介绍 3 种主要的动态话题模型,即嵌入增量式学习的词包、融入实时聚类的树体、基于时间索引的时序事件链.其中,时序事件链是本文面向树状话题模型提出的改进策略.此外,该节还简述了基于突发事件的增量式学习算法.第 4 节详细介绍核捕捉衰减(CKA)比对方法.第 5 节给出实验环境.第 6 节汇报评测结果并给予分析.第 7 节总结全文.

1 研究现状

话题跟踪的早期研究并未从真正意义上处理新闻话题的本源特性,从而关于话题形态及其建模方法的研究也并不深入.相对地,这一时期的研究往往致力于相关领域的技术移植,比如信息抽取、过滤和分类技术.Watanabe 等人[2]即把话题跟踪解释成一种信息抽取过程,借助诸如“正如我所提到的...”、“正如我所报道的...”和“正如近期发生的...”等规则标签,抽取话题内容进行相关性匹配.Zhang 等人[3]则利用基于内容的信息过滤技术识别和屏蔽报道流中的非相关报道.事实上,话题跟踪与信息过滤的信息处理过程确实存在一致性,即两者都需要从动态信息流中筛选相关信息和屏蔽不相关信息.因此,话题跟踪更像是信息过滤在特定知识领域(新闻信息)的分支研究.然而,话题跟踪具有诸多信息过滤涉及不到的特色问题,比如,话题跟踪需要借助时序事件(即时间顺序排列的真实新闻事件)学习话题演化脉络和检测变异锚点.由此,过滤技术并不能完全适用于跟踪问题,仅仅能够为跟踪过程中的文本相关性匹配和阈值估计方法提供参考.在更多情况下,这一时期的跟踪研究使用了分类技术,即相关和非相关报道的二元分类,例如 K 近邻(KNN)[4]、决策树(D-tree)[5]、线性分类器[6]等.然而,分类器往往在训练样本充分的情况下才能获得优越性能,而话题跟踪总是在预知少量相关报道(≤4)的情况下,实施后续相关与不相关报道的划分.因此,分类技术对话题跟踪问题的适应性不强.

针对话题模型的前瞻性研究来自 Allan 等人[7],其借用信息检索领域广泛采纳的向量空间模型(vector space model)描述话题的特征空间,从而建立了新闻话题的词包式描述雏形.Yang 等人[8]则对基于向量空间模型的话题描述提出扩展问题,并利用 Rocchio 算法从相关样本中挖掘话题特征实施扩展,且利用非相关样本屏蔽扩展中的噪声特征.此后,Allan 持续对基于 VSM 的话题模型提出各类改进,包括改进特征权重的估算函数[9]、利用名词、动词和名实体[10]改进话题模型的特征空间以及改进用于判定话题相关性的阈值估价函数[11]等.同一时期,以语言模型(language model)[12,13]为基础的话题建模方法相继出现.比如,Lavrenko 等人[14]的相关性模型(relevance model)从话题的先验相关样本中抽取关键特征形成查询描述,并将检索过程得到的伪相关反馈作为语言模型的训练样本,借以构造收敛于话题主旨的语言模型.此外,Nallapati[15]利用话题先验样本中的各类语义关系,建立了基于语义语言模型(semantic language model)的话题描述;洪宇等人[16]曾在前期工作中,利用篇章结构和依存关系建立了基于语义域语言模型(semantic domain based language model)的话题描述.虽然上述话题建模研究显著改进了话题跟踪系统的性能,但并没有对话题形态的机器学习理论带来跨越式的发展,甚至在一定程度上仍然沿用了相关领域的技术.比如,向量空间模型和语言模型都是通用于诸如检索、文摘和机器翻译等领域的统计模型;Rocchio 算法则一直用于信息过滤领域的用户模型(Profile)自学习.

相对来说,近期针对话题模型的研究则更多地融入话题结构特征[17,18]和演化特性[19],从而在真正意义上进入新闻话题形态学习的研究阶段.就话题结构的相关研究而言,多种聚类技术被引入话题内容的划分过程中,借以分治描述话题的不同子结构,比如表示种子事件的子结构和表示话题外延的子结构.聚类的最初动机是防止话题模型中的特征互为噪声.然而,特征划分和凝聚的优势不仅仅如此,比如,将内聚于同一内容的特征置入同一子结构,内容互斥的特征置入不同子结构,更有益于话题局部内容的语义描述[20].尤其是,凝聚式层次聚类既可以划分话题的不同子结构,同时又能建立子结构间的层次关系[21].其中最有代表性的工作来自 Zhang 等人[22]的层次树状结构,该树状结构的根节点能够描述话题的宏观概况,叶节点则可以具体描述话题的局部事件,根节点至叶节点的路径能够表述话题的发展脉络.然而,这类话题模型往往面临难以有效更新的困境,从而不适于话题演化的机器自适应学习.话题演化直接体现于话题发展过程中出现的新颖事件[23],以及话题在这类事件出现后作出的反应:保持话题主线或偏离话题主线.因此,针对事件的识别与描述成为学习话题演化的前提,He 等人[24,25]针对这一问题分别提出了基于毗邻特征的事件划分策略以及突发事件识别方法.但事实上,事件仅仅代

表话题的局部内容,而将事件贯穿为有机整体并形成话题发展脉络的特征是时序.然而,真正将时序和事件有效整合并形成话题模型的相关研究相对稀缺,相对地,时序往往作为特定话题的独有属性应用于增强相关话题的可匹配性[26−28].总之,话题结构特征和演化特性的发现开启了话题形态的研究,静态和动态话题模型是机器学习话题形态的两种重要路线,本文即重点针对这一问题展开探讨.

2 静态话题模型

静态话题模型构建过程可利用的资源仅仅是初期已知的有限相关报道,其性能优劣取决于模型本身是否充分描述了作为跟踪主线的种子事件.因此,针对初期相关报道的特征抽取及其权重估计,是决定话题模型性能优劣的关键,而这一点也恰是词包式文本描述关注的核心问题.本节首先给出一种基本的词包式话题模型,随后介绍以此为基础的各类变体.

2.1 基于词包的静态话题模型

向量空间模型(vector space model,简称 VSM)[1]是基于词包描述静态话题的基本模型,其他模型往往是对其特征抽取和权重估计方法改进后得到的变体.针对任意新闻话题,向量空间模型采用 N_t 篇(N_t=4)时序最早的已知相关报道作为话题样本,抽取 n_t 个(n_t=50)在话题样本中出现频率最高的词特征构造特征向量,每个词特征基于改进的 TFIDF 获取权重.给定词特征 i,其权重计算公式如下:

$$t_i = \frac{tf}{tf+2}\cdot(1-\log_N df_i) \tag{1}$$

其中,t_i 表示词特征 i 的权重,tf 是特征 i 在话题样本中的频度,df_i 是训练语料中出现特征 i 的新闻报道数,N 表示训练语料中新闻报道的总数.这一权重估计函数是检索模型 InQuery[1]中更为复杂的权重估计算法的简单形式,其假定语料中所有报道都粗略地具有相同长度,并且文档频率非零,即任意特征 i 的 $dfi\neq 0$.在此基础上,话题 T 与报道 D 的相关性估算函数如下:

$$r(T,D)=\frac{\sum_{i=1}^{n} t_i\cdot d_i}{\sqrt{\sum_{i=1}^{n_t} t_i^2}\cdot\sqrt{\sum_{i=1}^{n_t} d_i^2}} \tag{2}$$

其中,t_i 表示特征 i 在话题模型 T 中的权重,d_i 表示特征 i 在报道 D 中的权重.d_i 与 t_i 的估算方法相同,如公式(1),区别在于计算 d_i 时的 tf 为特征 i 在报道 D 中的频度.函数 $r(T,D)$事实上计算了两个高维向量夹角的余弦值.因此,两个向量包含权重相似的相同特征越多,两者相关性越高.在此基础上,跟踪系统只需预先训练恰当的阈值,即可实现相关报道的截取:相关性高于阈值则判定报道 D 为相关.

2.2 静态话题模型中的词包变体

以向量空间模型为基础,面向静态话题建模的词包描述具有多种变体.按其侧重点不同,可粗略分为两类:一类变体侧重挖掘最能表述种子事件内容的特征;另一类侧重划分特征权重,借以体现不同特征在表述种子事件时的价值.表 1 列举了目前绝大部分词包变体(表中在所有模型后加“-STM”后缀,以区分实验中的“-DTM”).

(1) 侧重特征抽取的词包变体

就特征抽取而言,基本的尝试是依据词性选择构建词包的特征,比如,表 1 中的静态话题模型 N-STM,V-STM 和 A-STM 分别抽取话题样本中的名词、动词和形容词形成向量空间模型.这类变体的建模依据是:不同词性的特征对不同类别的话题内容,具有不同的表述能力.比如,动词更善于表述行为,有益于诸如军事行动、自然灾害和气象变迁类话题的描述;形容词更善于表述状态,有益于诸如股市、金融和经济类话题的描述;而名词更善于表述实体,有益于诸如竞选、会晤和国际关系类话题的描述.

事实上,实体往往是新闻类信息的重要风向标,比如,时间、地点、人、物和机构等.为此,在名词中萃取实体特征,并专门以实体形成的词包也是静态话题建模中的重要一员,如表 1 中的 NE-STM.在此基础上,SR-STM 则进一步融入了实体的语义学属性,即由“施事”和“受事”标记的语义角色.由此,基于 SR-STM 的跟踪系统不仅需要匹配实体,还需检验实体的角色以否一致.

(2) 侧重权重估价的词包变体

词包描述中的重要权重估价方法是 Okapi BM25(对应表 1 中的 Okapi-STM).BM25 在衡量特征的重要程度上表现出的优势,能够有力地支持静态话题模型在初始样本稀疏的情况下准确地探测话题主线.相对于 BM25 复杂的数学模型,Rocchio 算法(对应表 1 中的 Rocchio-STM)[8]使用了一种简单的线性权衡方法度量特征与话题主线的依附程度.线性权衡的核心思想是:同时计算特征在已知相关样本和不相关样本中的频率,并使用后者削弱前者.显然,如果一类特征在某话题的相关样本中频繁出现,而极少出现于不相关样本,那么这类特征组成的词包更能凸显该话题的主旨,增强话题的排他能力.

Table 1 Words bag variant for static topic model

表 1 面向静态话题建模的词包变体

模型		特征	权重
侧重特征抽取	Basic-STM	任意特征	TFIDF
	N-STM	名词	TFIDF
	V-STM	动词	TFIDF
	A-STM	形容词	TFIDF
	NE-STM	名实体	TFIDF
	SR-STM	语义角色	TFIDF
侧重权重划分	OKAPI-STM	任意特征	BM25
	Rocchio-STM	任意特征	Rocchio
	LG-STM	任意特征	语言模型
	RM-STM	任意特征	相关性模型
	SM-STM	任意特征	语义模型

此外,语言模型(对应表 1 中的 LG-STM)在话题相关性判定中的成功应用,引领了一类独特的词包式静态话题模型.比如,相关性模型(对应表 1 中的 RM-STM)首先利用已知的话题样本建立查询表述;然后,依据这一查询在规模更大的报道集合中检索伪相关样本;最后,借助语言模型训练特征在这类样本中的生成概率.其中,伪相关样本的介入能够扩展用于平滑特征权重的上下文环境.在此基础上,语义语言模型(对应表 1 中的 SM-STM)更为直接地挖掘伪相关样本中的特征语义关系,特征权重的计算不仅依赖特征之间的共现率,还纳入了特征在句法结构中的依存关系以及这一关系的强度.

3 动态话题模型

静态话题模型与动态话题模型的基本区别是后者引入了自适应学习机制.自适应学习是一种机器自动学习新闻话题发展规律,并借助这一规律实时更新原有话题模型,使其有效检测和跟踪后续相关报道的方法.就基于词包的静态话题模型而言,引入增量式自适应学习[18]即可实现由静至动的转变.相对而言,具有层次结构的话题模型则需要借助更为复杂的自学习机制实现动态性.本节首先介绍基本的增量式学习方法;其次,给出一种基于突发新颖事件的改进策略;最后,介绍树状话题模型的自学习机制,并针对其不足给出基于时序事件链的动态话题建模和自学习方法.

3.1 基于词包的增量式学习

(1) 基本的增量式学习算法

基于增量式学习(incremental learning,简称位 IL)的跟踪系统往往将话题描述为向量空间模型,特征选择与权重估算与第 2.1 节给出基本静态话题模型类似.区别是,当跟踪过程开始后,系统从时序新闻流中每检测到一篇相关报道,增量式学习机制都将对话题模型进行更新.更新过程则根据特征在所有已检相关报道(包括最新检测到的相关报道)和初始话题样本(N_t=4)中的分布,重新估算特征的权重,如公式(1),并进行重排序;然后抽取 n_t(n_t=50)个权重最高的新特征重构话题模型.

增量式学习的理论依据是:随着学习过程不断利用新检测到的相关报道更新特征权重,后期相关报道中权重较高的特征将有机会融入话题模型.尤其是,当论述新颖事件的相关报道得以充分积累时,初始描述种子事件

的特征在所有已检相关报道中的分布将相对稀疏,而描述新颖事件的特征则密集出现,从而增量式学习借助新颖特征对固有特征的逐渐取代,渐进地引导话题模型的质心趋向话题漂移方向,从而提高系统召回后续相关报道的能力.

(2) 面向突发事件的增量式学习

然而,当话题因突发事件发生漂移时,上述渐进性学习方式难以及时识别漂移的拐点,从而产生迟滞性学习的现象.针对这一问题,本文提出一种突发式增量学习模型(burst incremental learning,简称 BIL),借以辅助增量式学习模型适应突发事件引发的话题漂移.BIL 侧重学习特征的爆发式频率,这一频率的计算方法如下:

$$BF(t)=\frac{tf}{l} \tag{3}$$

其中,*BF*(burst frequency)表示特征 t 的爆发式频率,tf 表示特征 t 在已检相关报道中出现的次数,l 表示特征 t 首次出现的时间点到跟踪系统当前检测点间的已检相关报道数量.基于这一方法,仅有较低频率的特征可能具有较高的爆发式频率.图 1 显示了一个估算爆发式频率的样例,其中,虽然特征 i 在所有已检相关报道中的频率高于特征 j,但它们爆发式频率的大小关系却恰恰相反.

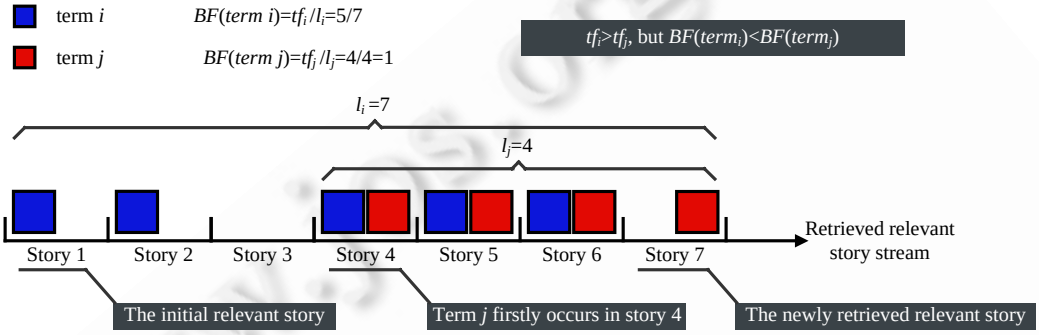


Fig.1 An example of burst frequency

图 1 爆发式频率估算样例

在此基础上,BIL 自学习为话题模型同时准备了两组特征序列 L_{inc} 和 L_{nvl}.其中,L_{inc} 中的特征基于频率 *TF* 计算权重并自高至低进行排序,而 L_{nvl} 中的特征则基于爆发式频率 *BF* 计算权重并进行排序.跟踪过程中,当系统检测到一个新的相关报道时,BIL 自学习机制将该报道中新颖的特征嵌入序列 L_{inc},利用公式(1)重新估算序列中所有特征的权重并进行重排序;与此同时,INL 模型则将新颖特征嵌入序列 L_{nvl},利用 *BF* 重新估算序列中所有特征权重并进行重排序.在此基础上,动态话题模型从序列 L_{inc} 中抽取 n_t 个权重最高(即频率最高)的特征构造特征向量 V_{inc},该向量侧重描述话题小幅度的漂移趋势;此外,动态话题模型从序列 L_{nvl} 中也抽取 n_t 个权重最高(即爆发式频率最高)的特征构造向量 V_{nvl},该向量侧重描述话题由突发事件引发的大幅漂移趋势.

事实上,上述两个特征向量不可避免地包含某些相同的特征.因此,BIL 自学习机制从向量 V_{nvl} 中删除已出现在 V_{inc} 中的特征,借以获得更纯粹影响话题大幅漂移的特征.跟踪系统的相关性判定机制则折中了向量 V_{inc} 和 V_{nvl} 对总体相关性的影响,估算函数如下:

$$R(T,D)=\alpha\cdot r(V_{inc},D)+\beta\cdot r(V_{nvl},D) \tag{4}$$

其中,T 表示某一新闻话题;D 表示报道的核,即报道中 n_t 个权重最高的特征组成的向量,权重估算方法如公式(1);$r(*,*)$表示两特征向量的相似度,计算方法如公式(2);参数 α 和 β 用于分配向量 V_{inc} 和 V_{nvl} 对相关度 $R(T,D)$ 的影响程度.本文经验性地设置 α 和 β 都为 0.5,其用意是假设向量 V_{inc} 和 V_{nvl} 对相关性判定有着相同的价值.换言之,长期的高频特征和短期的爆发式特征对识别相关报道有着相同的作用,前者可辅助话题模型适应相关报道流中长期探讨的事件,后者则可辅助话题模型快速适应突发事件.

3.2 结构化话题模型的动态变形

(1) 树状动态话题模型

结构化话题模型中的典型代表是一种具有层次的树状模型(hierarchical-tree,简称为 HT).区别于词包类话题模型中无组织的系列离散特征,树状话题模型中的所有特征都按照它们表述话题内容的层次进行了划分.基本的层次划分是宏观层次和具体层次,即善于表述话题宏观概念的特征集以及适于表述具体事件内容的特征集.在此基础上,利用层次聚类技术并结合特征在话题样本中的分布概率,可以将层次划分为更多种粒度.由此,树状话题模型可将不同特征散布于自根节点至叶节点的不同话题脉络上,每条脉络都具有自身内容的凝聚性,且自顶向下表述宏观至具体的话题属性.如图 2 中的树状话题模型样例,其中,根节点表示宏观的话题内容,即“9/11”,包含两条主干脉络,即“恐怖袭击”和“嫌犯调查”,每条脉络的叶节点部分可以具体到特定相关事件,比如“恐怖袭击”的叶节点包括“飞机劫持”、“世贸遇袭”和“五角大楼遇袭”事件.利用树状结构,跟踪过程在判定待测新闻报道是否相关时,可以按照深度遍历的路线,有针对性地匹配话题与报道的各个局部特性.

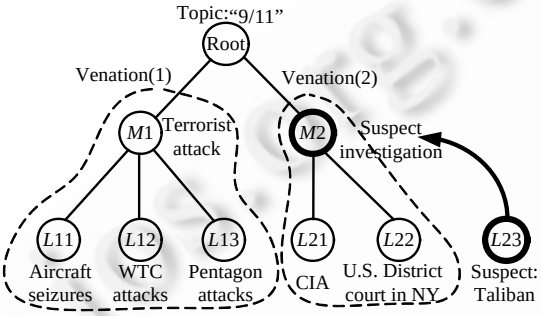


Fig.2 An example of hierarchical-tree topic model and its modification

图 2 层次树状话题模型及其更新样例

树状话题模型也可以借助自适应学习机制实现动态更新.区别于无结构的词包式话题模型,树状话题模型的自学习过程不仅对特征权重进行重新权衡,而且利用子结构的嵌入、剪枝与融合等步骤实现结构变形.基本的变形是:当跟踪系统识别出某一相关报道后,自学习机制将其作为叶节点嵌入树状结构,嵌入位置毗邻最相近的叶节点(即文本内容相似度最高的叶节点),同时融合邻居节点,抽取其中共性的特征形成新的父节点,原父节点从而升格为祖先节点.这一融合过程自底向上以此类推,使得与新叶节点最相关的话题脉络得以全面更新,如图 2 中,L23 嵌入话题模型后,对脉络 Venation(2)的直接影响,而不影响 Venation(1).

树状话题模型动态自学习的关键问题是如何选择恰当的嵌入点,比如,某些相关报道本身就包含多个事件的论述,或其主要内容是对话题的宏观概述,显然,这类报道不适合作为具体的叶节点嵌入树状结构,而应作为特定脉络的中间节点,甚至多个脉络的连接点.不恰当的嵌入将导致话题变形的误差,尤其是嵌入中间节点往往需要对嵌入点上下的毗邻结构进行相应更新,因此,错误的嵌入将误导一片区域内的大量节点描述.

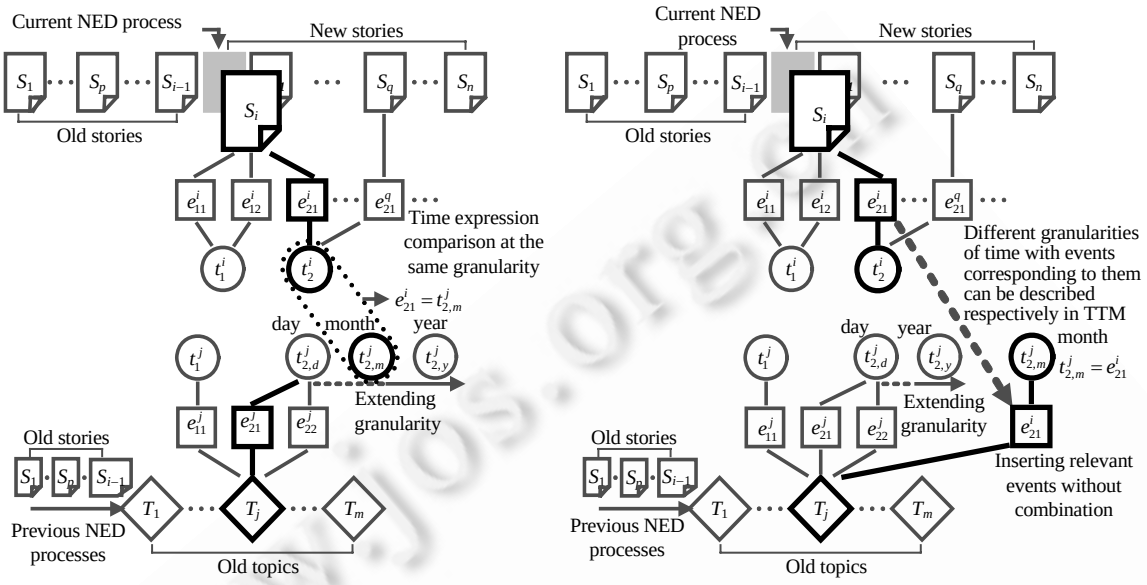
(2) 时序事件链式动态话题模型

针对树状动态话题模型的这一缺陷,本文提出一种时序事件链式的动态话题模型(temporal-event chain,简称为 TEC).时序事件链是一种以时间表达式为索引的事件集合,时间表达式指定特定事件发生的时间,比如图 3(a)中,话题 T_j 包含 3 个事件 e_{11}^j,e_{21}^j 和 e_{22}^j,其中,事件 e_{11}^j 对应时间表达式 t_1^j,事件 e_{21}^j 和 e_{22}^j 同时对应 t_{2d}^j.时间 t_{2d}^j 下标中的 d 表示 t_{2d}^j 的粒度为“day”.

时序表达式的基本粒度包括“year”,“month”和“day”,比如表达式“2001 年 9 月 11 日”的粒度为“day”,而“2001 年 9 月”的粒度为“month”.时间粒度的划分一方面是因为新闻报道流中本原地包含各种粒度的时间表达式,提高事件及其时间的召回率必然需要全面地挖掘各种粒度的时间信息;另一方面,不同时间粒度表述的事件内容往往具有不同层次,粒度粗糙的时间表达式对应的事件表述往往概括性强;相对来说,时间粒度精准时的事件表述往往具体细腻.因此,时间粒度的划分能够为链式结构的话题模型引入文本表述的层次信息.时间表达式可依

据 TERN(TERN 协会基于规则的标注规范覆盖了 TIMEX2 2001 指南中大部分时间类型(http://timex2.mitre.org/taggers/timex2_taggers.himl))协会基于规则的标注规范进行抽取.

此外,时序事件链中的事件是从新闻报道中抽取出的内聚文本块,抽取方法是基于时间表达式的 Textiling 改进算法[28].该算法不仅能够按照不同文本块的内聚相似性以及块与块之间的互斥性对不同事件内容进行划分,同时借助文本块中的时间表达式,能使事件对应于它发生的时间.由此,选取文本块中频率最高的词特征构成事件描述,并使这一描述对应文本块中的时间表达式,便形成了时序事件链中的基本结构:〈时间,事件〉对.



(a) 匹配过程 (b) 更新过程

Fig.3 Examples for comparison and modification of temporal-event chain

图 3 时序事件链匹配及更新样例

在此基础上,跟踪过程将话题和报道都描述为时序事件链,匹配两者相关性时,首先比对两者时序事件链中的时间索引,如果存在一致的时间,则进一步计算对应这一时间的事件相关度,如图 3(a),报道 S_i 包含的时间索引 t_2^i 和话题 T_j 的时间索引 t_{2m}^j 相同,则分别计算事件 e_{21}^i 与 e_{21}^j,e_{22}^j 的相关度.而话题与报道的相关性通过上述事件的相关度均值进行衡量,如图 3(a)中,话题 T_j 和报道 S_i 的相关度为 $Sim(e_{21}^i,e_{21}^j)$ 和 $Sim(e_{21}^i,e_{22}^j)$ 的加和取平均值.事实上,基于时序事件链的相关性衡量策略可以有多种方式,比如,取对应相同时间的最大事件相关度;借助时间粒度粗糙系数进行加权的事件相关度线性加和.此外,事件描述方式的变化也可以引入不同的相关性度量方法,比如,事件被描述为名实体和动词组合成的〈施事,受事,行为|状态〉语义关系,则匹配事件特征时,需要兼顾特征的语义属性.上述事件描述及其相关性度量方式将有效地改进跟踪的精确性.但本文着力单纯地评估不同结构的话题模型在静态和动态属性下的核捕捉性能,因此实验部分仅采用未加改进的简易时序事件链话题模型.

与树状动态话题模型相比,时序事件链的动态变形方式简明、直接,且对话题模型中的其他子结构并无影响.跟踪过程一旦检测到相关于特定话题的报道,即可启动自适应学习机制,对该话题的时序事件链实施动态变形.自学习过程如下:首先检测相关报道中的所有时间表达式是否都出现于话题事件链的时间索引中,对于未出现的时间表达式,自学习机制将其按时序嵌入索引,同时将其在相关报道中对应的事件描述嵌入事件链,并对应于这一索引;其次,对于已出现的时间表达式,自学习机制遍历对应的时间索引下所有的事件描述,并锁定最大匹配事件;同时,融合报道与话题模型中最匹配的事件描述,基于 TFIDF 重估特征权重并排序,选取权重最高的特征作为新的事件描述.

事实上,无论是匹配过程或是自学习过程,都对时间索引的粒度进行扩展,即某些事件在话题样本中只对应

一种时间粒度.为了提高召回率,则自发地赋予其他粒度.比如,图 3(a)的 e_{21}^j,e_{22}^j 在样本中只对应粒度为“day”的 t_{2d}^j,匹配时则扩展了另外两种粒度:“month”的 t_{2m}^j 和“year”的 t_{2y}^j,报道与话题的事件链实际可匹配的时间即是扩展后的 t_{2m}^j.相对来说,自学习过程虽然扩展了时间索引的粒度,但在查询最大匹配事件时,被扩展出的时间粒度并不参与其中,比如图 3(b)中,扩展粒度后的附加索引 t_{2m}^j 和 t_{2y}^j 不能用于最大匹配事件的查询与更新.但是,如果扩展出的时间索引与相关报道中的某些时间存在一致性,则自学习机制切断这类索引与原有事件的对应关系,然后将报道中一致时间对应的事件拉入事件链,同时将这类索引单独地指向这些事件.比如图 3(b)中,t_{2m}^j 与原有事件 e_{21}^j,e_{22}^j 的联系被切断,报道 S_i 中的事件 e_{21}^i 被嵌入事件链,同时,t_{2m}^j 单独指向 e_{21}^i.这一更新的目的是尽量保证真实的时序与事件对应关系,从而不同粒度的时间索引能够有效地反映事件描述的层次性(概括致具体).上述事件链的所有更新规则都参与了本文所涉的话题模型核捕捉实验.

4 核捕捉衰减比对策略

针对各类话题模型的性能,本文提出一种核捕捉衰减(capturing kernel attenuation,简称 CKA)评价标准.CKA 现象的直观体现是话题模型与相关报道的特征向量之间重叠的特征数呈现递减趋势.例如,图 4 显示的是 TDT 2002 评测中 40003 号话题(ID=40003)的 CKA 现象,该图横轴表示一系列按时间顺序排列的相关报道(相关于 40003 号话题),纵轴表示报道的核与静态话题模型重叠的特征比例.其中,话题模型的构造方法如第 2.1 节所述,即基本的静态词包向量 Basic-STM(N_t=50);报道的核由高频特征向量构成(n_t=50);特征重叠比为特征重叠数在报道核中所占的百分比.图 4 中,虚线上每个灰色点代表一个相关报道与静态话题模型的特征重叠比,整条虚线则描述了话题模型捕获相关报道核的 CKA 趋势,下文将其简称为 CKA 曲线.此外,为了直观表现 CKA 趋势,图 4 采用 5 阶多项式平滑策略近似地绘制了 CKA 曲线的平滑形式,如图中加粗的黑色实线.图中显示,虽然 40003 号话题的 CKA 曲线呈锯齿状分布,即重叠比例的增益与衰减交替出现,但是 CKA 的整体趋势呈现衰减.这一例证说明,话题后期的相关报道的核渐渐偏离话题初始的质心,但静态词包并没有适应这一偏离趋势.

核捕捉衰减比对指的是:给定同一话题的情况下,不同话题模型对这一话题的 CKA 趋势近似度.例如,图 5 显示,第 2.2 节的静态话题模型 Basic-STM 与 N-STM(N_t=50)在 40025 号话题上获得的 CKA 曲线有极为相似的衰减趋势.其中,尽管 N-STM 的 CKA 曲线在所有相关报道上取得了更高的特征重叠比例,但两者在后期相关报道流上取得的重叠比都呈现衰减趋势.

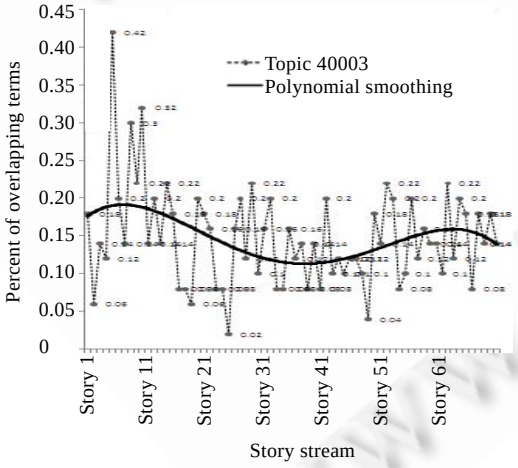


Fig.4 CKA curve of topic 40003 in TDT 2002

图 4 TDT 2002 中 40003 号话题的 CKA 曲线

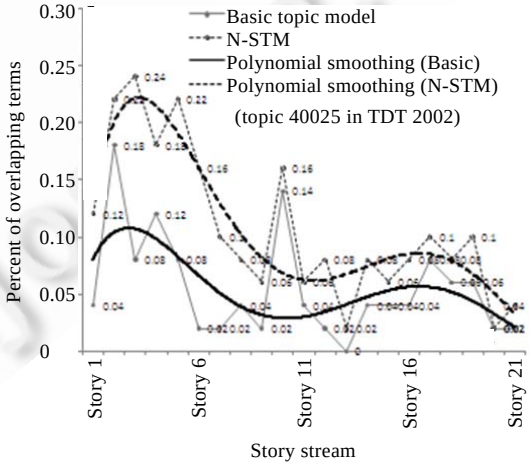


Fig.5 CKA trends comparison on Topic 20025 of TDT 2002

图 5 TDT 2002 中 20025 号话题的 CKA 趋势对

诸如图 5 所示的 CKA 曲线能够直观地反映不同话题模型是否存在近似的核捕捉衰减趋势.然而,针对大量新闻话题生成不同话题模型的 CKA 曲线,并借助人工观测获取比对结果,显然过于繁复.为此,本节提出一种 CKA 趋势的数值化对比策略,称为衰减趋近分析(attenuation approximation analysis,简称 A^3).针对某一新闻话题,两个话题模型取得的 CKA 曲线通过 A^3 进行近似性评估的过程包含如下 3 个步骤:

(1) 对于 CKA 曲线上每个点 dot_i(每个点对应横轴上某一新闻报道)建立衰减向量 $v_i=\{a_{i,1},\ldots,a_{i,i-1}\}$;向量的每一维 $a_{i,j}$ 表示 dot_i 是否高于点 dot_j,相比于 dot_i 对应的新闻报道,dot_j 必须对应 CKA 曲线上更早发生的报道,即 $i>j$;如果 dot_i 高于 dot_j,则 $a_{i,j}$ 等于 1,否则为 0.例如图 6,CKA 曲线 X 的衰减向量 $v_3(X)$,即 $\{a_{32},a_{31}\}$,的取值为{0,0},原因是 $dot_3(X)$既低于 $dot_2(X)$又低于 $dot_1(X)$,由此 a_{32} 和 a_{31} 都为 0.

(2) 针对两个 CKA 曲线上对应同一新闻报道的两个测试点,估算衰减向量的相似性,例如估算图 6 中测试点 $dot_2(X)$和 $dot_2(Y)$的衰减向量相似性.相似性的计算方法为向量空间夹角的余弦值.

(3) 两条 CKA 曲线整体衰减趋势的相似性通过两曲线上所有对应点的衰减相似度平均值进行估算,比如图 6 中 CKA 曲线 X 和 Y 的衰减相似性计算公式如下(由于 CKA 曲线的首点不存在对应更早相关报道的点 dot_j,因此该点不参与 A^3 计算):

$$A^3(X,Y)=\frac{\sum_{i=2}^{3}sim(v_i(X),v_i(Y))}{2} \tag{5}$$

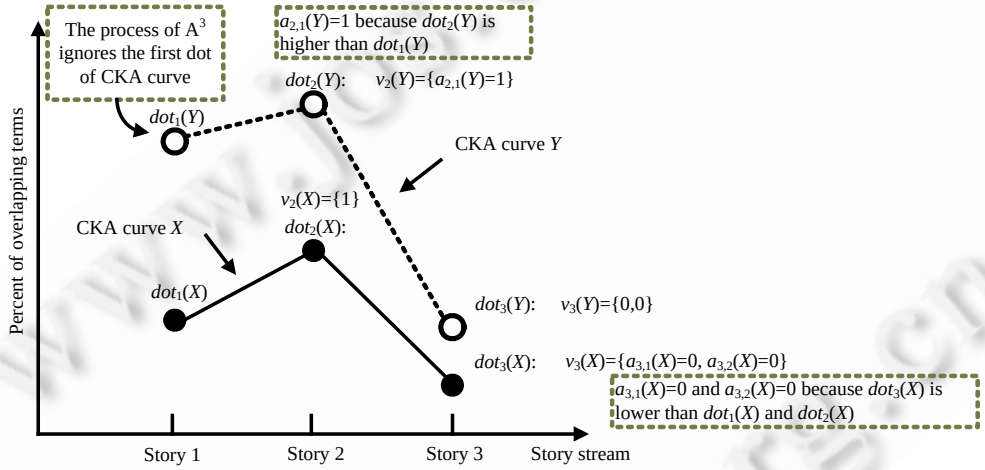


Fig.6 An example of step (1) of A^3

图 6 A^3 算法步骤(1)的样例

A^3 指标衡量的是两条 CKA 曲线衰减趋势的相似性.换言之,A^3 指标越高,则两条 CKA 曲线的衰减趋势越相近.由此,A^3 可以从如下 3 方面辅助话题模型的性能评价:

(1) A^3 可缩小观测规模.换言之,如果一类话题模型之间的 A^3 指标较高,则只需观测其中一种话题模型的 CKA 趋势,即可估测出其他同类话题模型的 CKA.

(2) 如果不同话题模型的 A^3 较高,但是各自隶属的话题跟踪性能(即跟踪系统输出端的最小 CDet 指标)差异较大,则可以判定劣势的跟踪系统并非在话题模型设计上相对较差,而是面向话题构建的特征抽取或权重估算存在缺陷.例如,图 5 中的话题模型 Basic-STM 和 N-STM 有着一致的衰减趋势,然而 N-STM 的 CKA 曲线整体低于 Basic-STM,即 Basic-STM 与相关报道流的重叠比始终较高.显然,Basic-STM 隶属的跟踪系统将更善于捕捉相关报道,降低漏检率,从而获得更优的最小检测错误权衡系数 CDet,而这一优势来自 Basic-STM 使用了全部词特征形式参与话题模型构建,而 N-STM 仅仅抽取名词.

(3) 相对来说,如果不同话题模型的 A^3 较低(即 CKA 趋势异同),且各自隶属的跟踪系统性能差异较大,则可断定劣势跟踪系统的缺陷很大程度上来自于话题模型设计的不合理.

对 A^3 而言,上述评价的最理想环境是不同话题模型所隶属的跟踪系统有着近似一致的相关度和阈值估算模块.然而如前文所述,话题模型的形制基本决定了相关度和阈值估算方法的选择,因此跟踪系统的性能差异主要源自话题模型设计本身.由此,上述基于 A^3 的断言具有一般可信性,且不仅可以用于本文所涉的话题跟踪研究,也可用于诸如话题关联检测、新事件检测、话题演化跟踪和变种检测等绝大部分包括话题模型设计的研究领域.

5 实验设计

5.1 实验语料

本文实验采用 TDT4 进行评测.TDT4 包括 98 245 篇新闻报道,其中 31 726 篇用于 2002 年话题检测与跟踪国际评测(Topic Detection and Tracking 2002,简称 TDT 2002).美国国家语言数据联盟(LDC)对其中 40 个新闻话题进行人工标注,共锁定 3 085 篇不重复的相关新闻报道,其他标注为不相关.此外,2003 年的 TDT 评测(简称 TDT 2003)使用了 TDT4 中的另外 17 802 篇新闻报道,并对其中 40 个新闻话题进行标注,共锁定 3 083 篇相关报道.本文实验采用 TDT 2002 对应的语料作为训练语料,将 TDT 2003 的语料用作测试.

5.2 系统评价标准(最小C_{Det})

美国国家标准技术研究院(National Institute of Standards and Technology,简称 NIST)为话题检测与跟踪系统提供了标准的评价策略.该策略主要权衡系统的漏检率和误检率,如公式(6):

$$C_{Det}=C_{Miss}P_{Miss}P_{target}+C_{FA}P_{FA}P_{non\text{-}target} \tag{6}$$

其中,C_{Miss} 和 C_{FA} 分别表示系统漏检和误检的代价系数,C_{FA} 等于 1,C_{Miss} 等于 10;P_{Miss} 和 P_{FA} 分别是系统漏检和误检的条件概率;P_{target} 和 $P_{non\text{-}target}$ 是先验目标概率($P_{non\text{-}target}$=1−P_{target}),P_{target} 等于 0.02;检测错误权衡系数(C_{Det})是综合了系统漏检率与误检率而得到的总体性能损耗代价.评价 TDT 系统时常采用 C_{Det} 的规范化表示$(C_{Det})_{Nom}$,其定义如下:

$$(C_{Det})_{Norm}=\frac{C_{Det}}{\min(C_{Miss}P_{target},C_{FA}P_{non\text{-}target})} \tag{7}$$

显然,较低的误检率表示系统具有较高查准能力,较低的漏检率表示系统召回能力较强,而通过权衡误检率和漏检率而得到的 C_{Det} 能够评定系统的综合性能.同样地,C_{Det} 越低,系统性能越优.对特定跟踪系统自身而言,调整阈值时,C_{Det} 会随之发生变化.其中,最小 C_{Det},即 $\mathrm{Min}(C_{Det})_{Norm}$,代表跟踪系统所能达到的最佳性能,$\mathrm{Min}(C_{Det})_{Norm}$ 对应的阈值即为最佳阈值.由此,一种评价不同跟踪系统性能的方法是横向比较它们各自的 $\mathrm{Min}(C_{Det})_{Norm}$,指标最低的系统具有最佳跟踪性能.

5.3 实验安排

本文实验主要包含如下两个部分:第 1 部分,利用 CKA 曲线观测法和 A^3 分析法检验静态和动态话题模型的核捕捉能力;第 2 部分,检验融入不同话题模型的跟踪系统性能,并利用 A^3 和 $\mathrm{Min}(C_{Det})_{Norm}$ 对各话题模型的主要缺陷给予剖析.

参与实验的静态话题模型包括第 2 节罗列的词包式静态话题模型及其所有变体,如表 1;动态话题模型包括嵌入增量式学习的词包式模型(实验中用 IL-DTM 标识)、层次树状动态话题模型(标识为 HT-DTM)以及本文分别针对 IL-DTM 和 HT-STM 提出的改进方法,即嵌入突发式增量式学习的词包式模型(标识为 BIL-DTM)和时序事件链式动态话题模型(标识为 TEC-DTM).其中,DTM 为动态话题模型的英文简写(即 dynamic topic model).

此外,为了评测的公正性,所有参与实验的话题模型统一地只允许 50 个特征(n_t=50)参与 CKA 比对和相关性匹配.对于所有基于词包构建的话题模型(包括静态和动态)而言,设 n_t=50 为特征向量 VSM 的维度即可满足这一要求.但对于结构化的话题模型,即树状话题模型 HT-DTM 和链式话题模型 TEC-DT,n_t 一般是多个子结构

的特征向量维度之和,因此需要对子结构数量进行限制.就 HT-DTM 而言,随着自学习机制对层次树的不断更新,树的深度和广度将不断扩展,包含的特征数量往往远大于 n_t(n_t=50).但 HT-DTM 实际用于度量话题与报道相关度的结构是一条自根节点到叶节点的最优路径,因此实验设置 HT-DTM 用于 CKA 比对和相关性匹配的最优路径深度为 5,路径上每个节点的特征维度为 10,从而满足 n_t=50 的要求.对于链式话题模型 TEC-DTM,实验设置事件链上的每个事件可以使用 10 个特征进行描述,且只有时序最晚的 5 个事件可参与 CKA 比对和相关性匹配,从而也满足 n_t=50 的要求.

6 实验结果及分析

6.1 CKA观测及A^3比对结果

(1) Basic-STM 的 CKA 观测结果

实验首先利用 CKA 曲线描绘基本的词包式静态话题模型 Basic-STM(建模过程如第 3.1 节)的核捕捉趋势,如图 7 所示.该图包含了 TDT 2002 语料中所有相关报道数多于 4 的新闻话题,每个话题对应一幅 CKA 趋势图.相关报道数少于 4 的话题未参与此实验,原因是话题模型构建的初始样本(即已知的相关报道)数为 4(N_t=4).

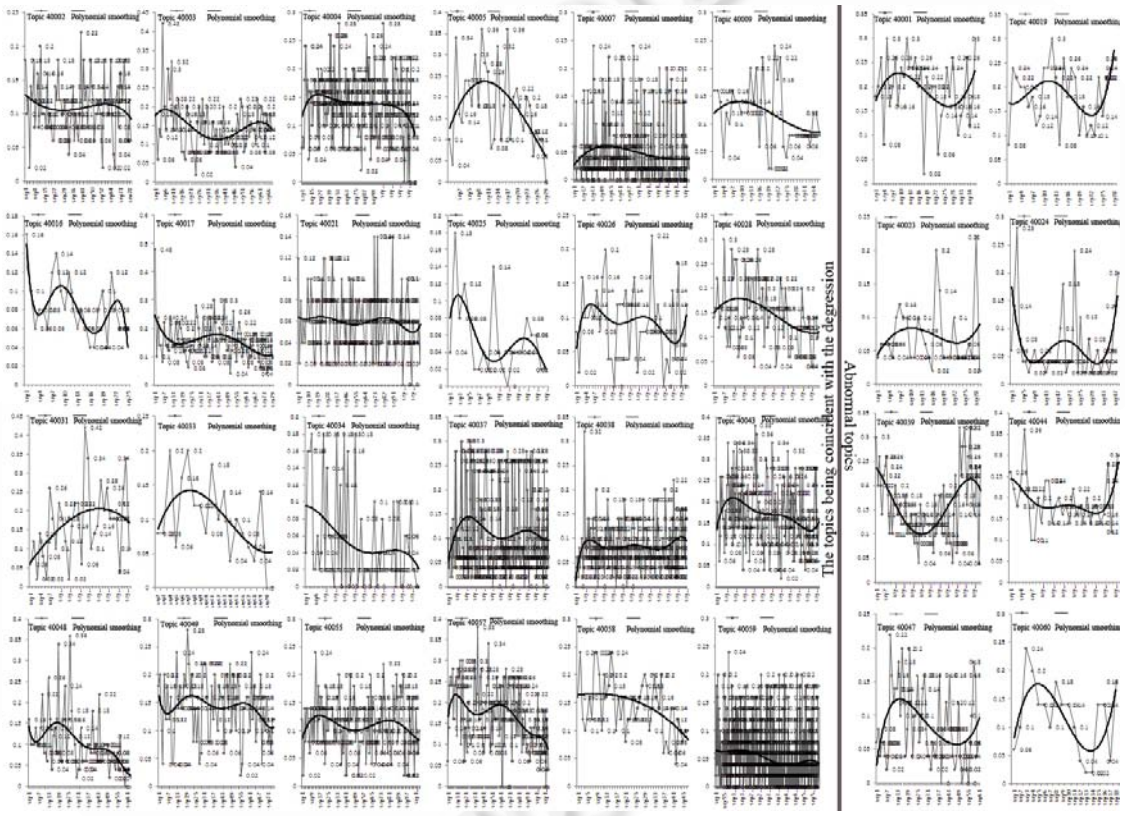


Fig.7 CKA curves of all the topics that have more than 4 relevant stories in TDT 2002 corpus (using Basic-STM)

图 7 TDT 2002 语料中所有相关报道数多于 4 的话题的 CKA 曲线(使用 Basic-STM)

实验结果显示,Basic-STM 对 TDT 2002 大部分新闻话题的核捕捉过程存在衰减现象,如图 7 中排列于灰色竖线左侧的多组子图.相对地,图 7 中灰色竖线右侧的新闻话题并未出现极为明显的衰减现象,它们对应的 CKA 曲线的平滑形式在尾部存在一定的上扬趋势.其原因是 TDT 2002 语料仅包含较短时间段内新闻事件的报道

(2000 年 10 月~2001 年 1 月),因此该语料并未囊括上述话题的所有后续相关报道,尤其长期话题,如话题“911 恐怖袭击”活跃于媒体的时间甚至波及次年的二次美伊战争.对于这类话题,CKA 检验不能获得完整的话题演化趋势.换言之,上述未出现衰减趋势的话题很大程度上是由于其 CKA 曲线仅仅记录了局部的话题发展过程.这一点可由这些话题相对稀疏的相关报道数量得以证明(图 7 灰色竖线右侧话题的相关报道数量明显少于左侧的话题).

值得关注的是,话题的 CKA 曲线普遍存在波状的衰减趋势,如图 7 中 CKA 曲线的平滑形式在衰减后存在小幅上扬,然后再衰减.这一波状衰减中的波峰表示静态话题模型与相关报道的核具有相对较高的特征重叠比例,而出现在 CKA 曲线中后部的波峰说明:虽然静态话题模型适应话题演化的能力随时间递减,但往往在特定时期会有所反弹.那么,既然话题后期的相关报道更倾向于讨论新颖事件,而静态话题模型更专注于初始的种子事件,为什么会存在这一反弹现象?其原因可归咎于新闻话题存在的回顾式叙述特点.回顾式叙述指的是当一篇新闻报道侧重论述某一话题的新颖事件时,其往往附带地回顾该事件的背景,即已经发生过的相关事件,尤其是激发这一话题的种子事件.比如,当一篇报道着力论述话题“911 恐怖袭击”的新颖事件“恐怖分子嫌疑人调查”时,往往附带地回顾种子事件“恐怖分子袭击世贸大厦及五角大楼”.话题回顾是新闻报道重要的叙述手段,其有助于读者了解新颖事件的相关背景.但话题回顾并非必不可少,这一点可以由 CKA 曲线的波谷得以验证.原因在于,如果背景事件发生的时间相对较近,则新闻报道往往不增添冗余的篇幅对其额外进行论述,恰似新闻媒体预先认定读者已经了解刚刚发生的事件一样.

总之,CKA 曲线验证了静态话题模型 Basic-STM 在适应话题演化过程中的显著不足,即无法有效地捕捉后期相关报道的核.但是,上述 CKA 现象是否在基于词包的其他静态话题模型中普遍存在?下面利用 A^3 方法对 Basic-STM 及其所有变体(变体的介绍如第 2.2 节)进行比对分析.

(2) 静态话题模型(STM)的 A^3 比对分析

基于词包的静态话题模型变体与 Basic-STM 的比对结果见表 2,比对过程同样使用了上述 TDT 2002 中相关报道数多于 4 的所有新闻话题(共 32 个话题).比对过程如下:给定 Basic-STM 的某一变体,获取该变体与 Basic-STM 在每个话题上的 CKA 曲线,并计算 CKA 曲线之间的 A^3 指标,记录分布于不同 A^3 指标范围内的话题数量.比如,给定话题模型 N-STM,表 2 的第 2 行第 5 列记录了共有 25 个话题的 A^3 指标位于 0.7~0.8 范畴之内.换言之,N-STM 与 Basic-STM 在这 25 个话题上取得的 CKA 曲线具有约为 0.7~0.8 的衰减相似性(即 $0.7<A^3<0.8$).为评估 CKA 曲线是否具有相似衰减趋势,有必要对 A^3 指标设置阈值δ加以裁决,即如果两条 CKA 曲线的 A^3 指标高于这一阈值,则它们具有相似的衰减趋势.通过对大量 CKA 曲线的观测,当阈值δ大于 0.6 时,CKA 曲线具有近似的衰减趋势.作为比照,图 5 中两条衰减趋势极为相似的 CKA 曲线具有约为 0.69 的 A^3 指标.表 2 中的符号(+)表示变体的 CKA 曲线整体位于 Basic-STM 的上方,比如,图 5 中 N-STM 对应的 CKA 曲线(虚线)位于 Basic-STM 的 CKA 曲线(实线)上方,则 A^3 将被标记(+);同理,符号(–)表示变体的 CKA 曲线整体位于 Basic-STM 的下方;符号(±)则表示 CKA 曲线相互交织.

如表 2 所示,侧重改进特征抽取的词包变体 N-STM,NE-STM,SR-STM 以及侧重改进特征权重的词包变体 OKAPI-STM,Rocchio-STM,LG-STM,RM-STM 和 SM-STM 在绝大多数 TDT 2002 的新闻话题上取得了较高(>0.6)的 A^3 指标,如表 2 中值域大于 0.6 的系列 A^3 指标.这一结果说明上述变体与 Basic-STM 往往具有相似的 CKA 趋势,从而验证它们也存在难以适应话题漂移趋势和捕捉后续相关报道核的现象.相对来说,变体 V-STM 和 A-STM 则在绝大多数新闻话题上取得较低(<0.6)A^3 指标,这一结果说明它们与 Basic-STM 往往具有异同的 CKA 趋势.然而实验发现,V-STM 和 A-STM 的所有 CKA 曲线都极为逼近坐标系的横轴,且曲线上绝大部分点都对应极低的特征重叠比例.其说明话题模型 V-STM 和 A-STM 几乎无法捕捉所有相关报道的核,包括早期的相关报道.造成这一现象的原因是动词和形容词描述话题时的稀疏性和多样性.总之,上述 A^3 评测结果验证了静态话题模型难以适应话题漂移的现象具有一定的普遍性.

(3) 增量式学习的 CKA 观测结果

实验同样利用 TDT 2002 对动态话题模型 IL-DTM 和 BIL-STM 进行 CKA 观测,IL-DTM 和 BIL-STM 是在

词包式静态话题模型 Basic-STM 基础上嵌入增量式学习后形成的动态话题模型,其中,BIL-STM 是本文针对 IL-DTM 提出的改进方法(如第 3.2 节).这一观测实验同时显示了 Basic-STM 的 CKA 曲线,借以直观地比较静态与动态话题模型的核捕捉能力,如图 8 所示.其中,每个子图对应 TDT 2002 中的一个话题,所有子图中的灰色虚线标识 Basic-STM 的 CKA 曲线,黑色点划线标识 IL-DTM 的 CKA 曲线,实线标识 BIL-DTM.

相对于静态话题模型 Basic-STM,图 8 中两个动态话题模型 IL-DTM 和 BIL-DTM 都获得了特征重叠比例更高的 CKA 曲线.但是如表 2 所示,基于名词构造的静态话题模型 N-STM 也可获得更高的重叠比.因此,增量式学习是否改善话题模型的适应性,应该侧重检验它们是否阻止了 CKA 曲线后期分布趋势的衰减现象.图 8 显示,增量式学习使动态话题模型的 CKA 曲线尾部得以显著提升.此外,通过观测发现,两个动态话题模型获得的 CKA 曲线和 Basic-STM 的 CKA 曲线在尾部的间距普遍大于首端的间距.

Table 2 Results of the evaluation A^3 among static topic models

表 2 静态话题模型的 A^3 评测结果

模型	A^3 值域					
	$A^3<0.5$	$0.5<A^3<0.6$	$0.6<A^3<0.7$	$0.7<A^3<0.8$	$0.8<A^3<0.9$	$0.9<A^3$
N-STM	0	0	1(+)	25(+)	3(+)	3(+)
NE-STM	0	1(−)	2(−)	22(−)	4(−)	3(−)
SR-STM	0	1(−)	5(−)	19(−)	4(−)	3(−)
V-STM	3(−)	13(−)	9(−)	2(−)	2(−)	3(−)
A-STM	3(−)	15(−)	5(−)	0	3(−)	3(−)
OKAPI-STM	0	0	1(±)	1(±)	20(±)	16(±)
Rocchio-STM	0	0	1(±)	0	15(±)	22(±)
LG-STM	0	0	1(±)	9(±)	16(±)	6(±)
RM-STM	0	1(±)	0	11(±)	17(±)	3(±)
SM-STM	0	2(±)	6(±)	18(±)	5(±)	1(±)

(+):表示评测中当前模型获得的 CKA 曲线上所有点都比 Baseline 模型的高;
(−):表示评测中当前模型获得的 CKA 曲线上所有点都比 Baseline 模型的低;
(±):表示评测中当前模型获得的 CKA 曲线上的点与 Baseline 模型 CKA 性能曲线上的点高低交错.

(4) 动态话题模型的 A^3 比对分析

这一现象(如图 8 中 CKA)验证增量式学习削弱了话题模型适应性衰减的幅度,换言之,经过学习得以更新的动态话题模型更善于识别话题后期的相关报道.此外,BIL-DTM 与 IL-DTM 具有非常类似的 CKA 趋势,且除了重合的 CKA 片断以外,其他 BIL-DTM 的 CKA 片断几乎都位于 IL-DTM 之上.由此可以预见,相比于 IL-DTM,基于 BIL-DTM 的跟踪系统将获得更优的性能.尤其是,BIL-DTM 在部分话题上的 CKA 曲线具有更为上扬的尾部(如图 8 中 Topic ID: 40028),说明基于突发事件的增量学习可以更为有效地捕捉话题漂移的后续脉络.换言之,动态话题模型及时地融入短期内爆发式增益的特征,有助于其迅速捕获突发的新颖事件.

本文所有动态话题模型(IL-DTM,BIL-DTM,HT-DTM 和 TEC-DTM)与静态话题模型 Basic-STM 的 A^3 比对结果如表 3 所示.结果显示,动态话题模型在绝大部分话题上取得的 CKA 趋势都与 Basic-STM 不一致.结合表 2 进行分析,绝大部分基于词包的静态话题模型都与 Basic-STM 具有极为相似的 CKA 趋势.由此,以 Basic-STM 为媒介,A^3 的比对结果能够证明动态话题模型与静态话题模型在捕捉话题漂移时的显著差异.尤其当话题结构不同时,如树状话题模型 HT-DTM 和事件链式话题模型 TEC-DTM,这一差异将更为明显.表 3 显示,HT-DTM 和 TEC-DTM 与 IL-DTM 和 BIL-DTM 相比,具有更多话题的 A^3 低于 0.5.

Table 3 Results of the evaluation A^3 between dynamic topic models and Basic-STM

表 3 动态话题模型与 Basic-STM 的 A^3 比对结果

模型	A^3 值域					
	$A^3<0.5$	$0.5<A^3<0.6$	$0.6<A^3<0.7$	$0.7<A^3<0.8$	$0.8<A^3<0.9$	$0.9<A^3$
IL-DTM	22(+)	8(+)	2(+)	0	0	0
BIL-DTM	25(+)	6(+)	1(+)	0	0	0
HT-DTM	26(+)	6(+)	0	0	0	0
TEC-DTM	29(+)	3(+)	0	0	0	0

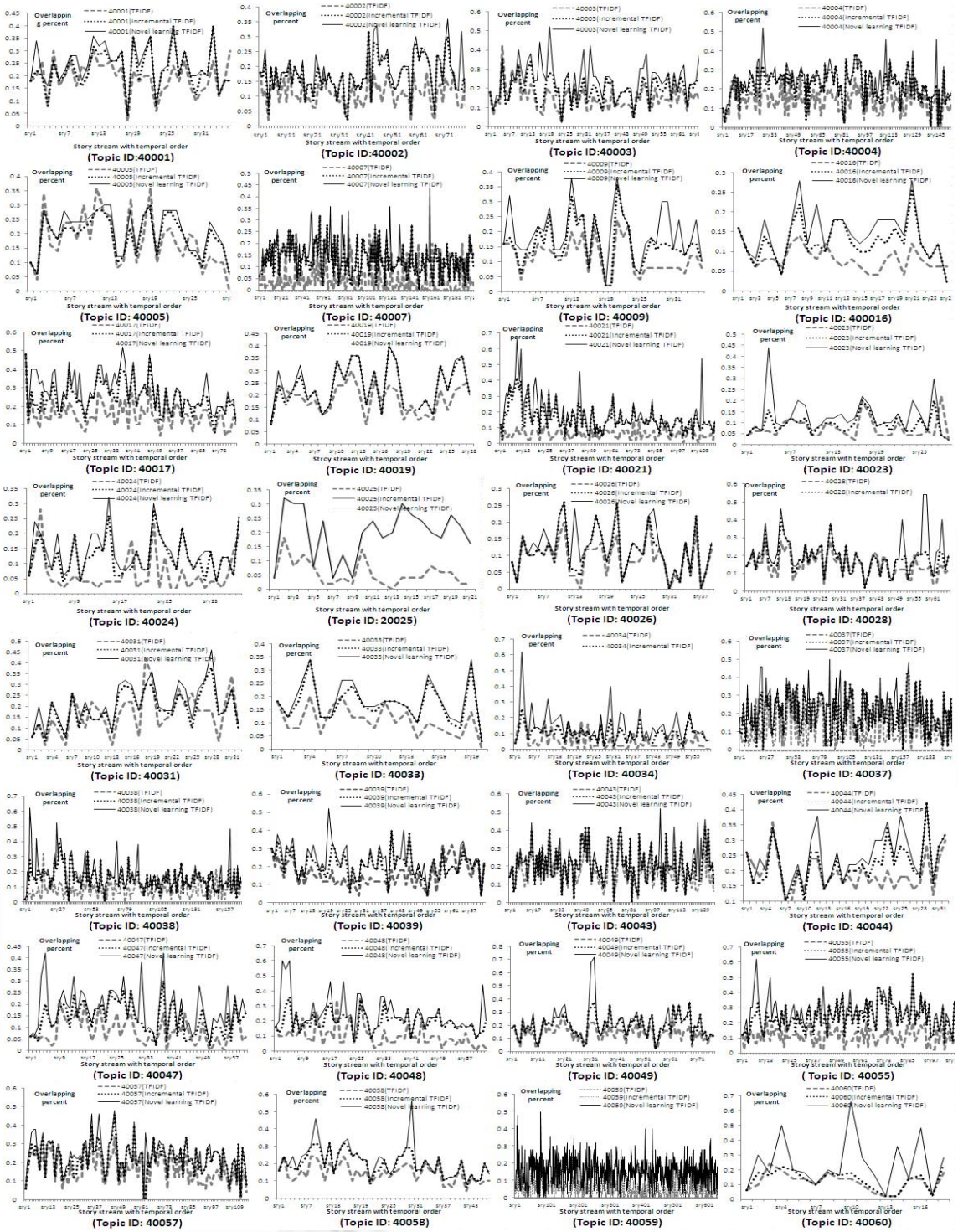


Fig.8 CKA results of dynamic topic models

图 8 动态话题模型 CKA 结果

此外,动态话题模型相互间的 A^3 比对结果见表 4.比对的基准是嵌入增量式学习的话题模型 IL-DTM.结果

显示,其他 3 种动态话题模型与 IL-DTM 在绝大部分话题上取得了高于阈值δ(δ=0.6)的 A^3 指标.尤其是,BIL-DTM 与 IL-DTM 有 26 项话题的 A^3 指标高于 0.9,占参评话题总数的 81%,表现出高度的 CKA 趋势一致性,如图 8 所示.其原因是,BIL-DTM 仅仅对 IL-DTM 的自学习机制进行了适量改进,即利用突发事件中具有促发概率的特征,补充 IL-DTM 实时修正的话题核心描述.相比之下,HT-DTM 和 TEC-DTM 分别只有 9 项和 6 项话题获得高于 0.9 的 A^3 指标,原因在于话题模型结构本身即具有差异.但总体而言,上述话题模型与 IL-DTM 都具有近似的 CKA 趋势($A^3>\delta$),换言之,动态话题模型普遍善于捕获话题的漂移趋势.

Table 4 Results of the A^3 comparison between dynamic topic models and Basic-STM

表 4 动态话题模型之间的 A^3 比对结果

模型	A^3 值域					
	$A^3<0.5$	$0.5<A^3<0.6$	$0.6<A^3<0.7$	$0.7<A^3<0.8$	$0.8<A^3<0.9$	$0.9<A^3$
BIL-DTM	0	0	1(±)	3(±)	2(±)	26(±)
HT-DTM	0	2(+)	2(+)	8(+)	11(+)	9(±)
TEC-DTM	0	3(+)	1(+)	6(+)	16(+)	6(±)

总之,CKA 趋势图和 A^3 指标的对比结果说明:话题模型的优劣更多地取决于话题结构的设计合理性,而不是特征选择.详细而言,表 2 中不同静态话题模型(结构不可变)的区别仅在于选择的特征形式(动词、名词与形容词等等)和权重度量方法,而获得的 CKA 趋势基本一致(A^3 指标较高),即捕捉大部分话题核的能力都存在近似的衰减趋势,特征选择及其权重度量方法的异同并未使这一趋势发生明显变化.相反,当融入动态学习机制后(结构可变),如 IL-DTM 和 BIL-DTM,CKA 趋势则出现了显著差异,两者与话题模型 Basic-STM 的 A^3 指标普遍很低.此外,根据图 8 的 CKA 观测数据,IL-DTM 和 BIL-DTM 具有更强的核捕捉能力.因此,话题结构的设计对话题模型的优劣具有决定性的作用.下文给出的跟踪系统性能将进一步验证,改善话题的结构设计,如层次树型和时序链式动态话题模型,能够进一步改进跟踪性能.

6.2 话题跟踪性能及分析

(1) 静态跟踪系统性能及分析

基于测试语料,即 TDT 2003,实验对嵌入上述各种话题模型的跟踪系统进行测试.测试中,所有跟踪系统的阈值皆基于训练语料(TDT 2002)所得.阈值训练过程是:首先采用 CKA 的 5 阶多项式平滑函数的最大取值作为采样上界,将 CKA 曲线上的最小值作为采样下界;然后,计算各跟踪系统的匹配算法在上述两界点可获得的相关度;最后,驱动阈值以特定粒度在两界点的相关度之间渐进取值,并检测对应最小检测错误权衡系数的阈值取值,这一取值即为训练最优解.与泛 0~1 之间的所有阈值取值相比,这一训练方法有效缩减了阈值的采样范畴.就阈值的渐变粒度而言,使用话题模型 LG-STM,RM-STM 和 SM-STM 的跟踪系统采用 0.001 为粒度,原因是上述话题模型的匹配需要利用 KL(kullback-leibler)距离进行计算,相关度粒度较小.除此之外,其他跟踪系统皆采用 0.01 为粒度.所有系统训练出的最优阈值θ见表 5 和表 6.

Table 5 Results of the A^3 comparison among dynamic topic models

表 5 基于静态话题模型的跟踪系统测试结果

模型	Basic-STM	N-STM	NE-STM	SR-STM	Okapi-STM	Rocchio-STM	LG-STM	RM-STM	SM-STM
$(C_{Det})_{norm}$	0.092 8	0.086 1	0.151 6	0.152 8	0.091 2	0.090 3	0.089 8	0.100 3	0.089 1
	θ=0.30	θ=0.33	θ=0.16	θ=0.11	θ=0.30	θ=0.29	θ=0.004	θ=0.003	θ=0.003

Table 6 Test results of tracking systems based on dynamic topic model

表 6 基于动态话题模型的跟踪系统测试结果

模型	IL-DTM	BIL-DTM	HT-DTM	TEC-DTM
$(C_{Det})_{norm}$	0.062 6	0.058 2	0.039 1	0.029 6
	θ=0.34	θ=0.35	θ=0.36	θ=0.36

测试结果中,基于静态话题模型的跟踪系统性能见表 5.其中,基于话题模型 NE-STM,SR-STM 和 RM-STM

的跟踪性能逊色于 Basic-STM(检测错误权衡系数 C_{Det} 越低说明系统性能越好).就 NE-STM 和 SR-STM 而言,跟踪性能较差的原因可以直观地从 A^3 比对结果中探寻.如表 2 所示,虽然 NE-STM 与 SR-STM 都有着近似于 Basic-STM 的 CKA 趋势(A^3>0.6 的话题占 98%),但两者在所有参训话题上的 CKA 曲线都低于 Basic-STM,见表 2 中标识为(–)的话题数.换言之,NE-STM 与 SR-STM 的特征重叠比全部低于 Basic-STM,即两者对测试报道的核捕捉能力自始至终劣于 Basic-STM.直觉上,NE-STM 借助名实体的匹配本应获得较高的准确率(即较低的误检率),从而具有更优的 C_{Det}.但实际上,由于同一实体可以采用多种命名方式,比如“乔治.布什”和“小布什”,且实体在行文中往往被指代,从而名实体在匹配过程中存在大量遗漏,也因此,NE-STM 较劣的 C_{Det} 源于较高的漏检率(即召回率低).SR-STM 重叠比较低的根本原因是语义匹配的限制过于严格,即不仅匹配特征的词形,还需匹配语义角色,同一词形作为不同角色时不能匹配.因此,SR-STM 较劣的 C_{Det} 也源于较高的漏检率.

相比之下,RM-STM 采用了同于 Basic-STM 的特征选择方式,即任意高权重词特征.同时,RM-STM 借助检索到的相关文本对描述话题核的有效特征实现了扩展.换言之,RM-STM 能够使用更丰富的相关词特征描述话题的核心内容,因此,RM-STM 应具有更高的重叠比.但是,表 2 显示的 A^3 指标却否定了这一直观判断,其大量的(±)标识说明 RM-STM 和 Basic-STM 的 CKA 曲线总是相互交织,即 RM-STM 重叠比无明显优势.原因在于,RM-STM 检索时召回的相关文本实际上是伪相关文本,蕴含部分噪声信息.因此,RM-STM 向话题核心描述融入相关特征的同时,也近似同比地引入了噪声特征,从而对重叠比改善不大.可是,OKAPI-STM,Rocchio-STM,LG-STM 和 SM-STM 也存在 CKA 曲线相互交织的情况,如表 2 的(±)标识,但跟踪性能却略优于 Basic-STM.通过 CKA 曲线观测,上述模型的交织情况不同于 RM-STM,其交织的 CKA 片断最多不超过 CKA 曲线全长的 25%,而 RM-STM 则远大于这一比例,近似为 50%.

值得注意的是,N-STM 的 CKA 曲线既不与 Basic-STM 相互交织,又都位于 Basic-STM 之上,如表 2 中的(+)标识,且 N-STM 的跟踪性能最佳,优于 Basic-STM 约 0.7 个百分点.相比之下,NE-STM 与 SR-STM 的 CKA 曲线也不与 Basic-STM 相互交织,但都位于 Basic-STM 之下,如表 2 中的(–)标识,且 NE-STM 与 SR-STM 跟踪性能都远逊于 Basic-STM.此外,其他静态话题模型的 CKA 曲线都与 Basic-STM 相互交织,且跟踪性能可优可劣,即使有所改进,幅度也不大(提高最多不超过 0.37%).由此可以产生 A^3 的一条评价规则:比对两话题模型的跟踪性能时,如果两者的 A^3 指标较高,即趋势相同,那么具有更高重叠比的 CKA 曲线对应的话题模型具有较好的跟踪性能.

(2) 动态跟踪系统性能及分析

测试结果中,基于动态话题模型的跟踪系统性能见表 6.结果显示,所有动态话题模型的检测错误权衡系数 $(C_{Det})_{nor}$ 皆低于任意静态话题模型.相比于最优的静态话题模型 N-STM,动态话题模型 IL-DTM,BIL-DTM,HT-DTM 和 TEC-DTM 分别提高 2.4%,2.8%,4.7%,5.7%.其中,IL-DTM 和 BIL-DTM 获得了相近的跟踪性能(相差 0.4%).事实上,依据图 8 的系列 CKA 曲线和表 4 的 A^3 比对结果,即可预测 IL-DTM 和 BIL-DTM 跟踪性能的近似性.表 4 显示,两者在训练语料中的 26 个话题上(占所有话题的 81%)具有高于 0.9 的 A^3 指标,即两者往往具有极为一致的趋势.同时,图 8 显示 BIL-DTM 仅在 CKA 的有限局部片断上高于 IL-DTM,即局部具有更高的特征重叠比.因此,BIL-DTM 对 IL-DTM 的改进比较有限.由此可以产生 A^3 的另一条评价规则:如果参评的两个话题模型,具有极高的 A^3 指标,即 CKA 趋势极为一致,且 CKA 曲线相互交织,如表 4 中 BIL-DTM 的(±)标识,则两话题模型将具有一致的跟踪性能.

BIL-DTM 对 IL-DTM 的改进(0.4 个百分点,如表 6)说明,快速地向话题模型融入突发事件的特征,有益于改善增量式学习的效率.但改进率低的现象也说明,IL-DTM 仍难以高效适应已然出现的话题漂移趋势.其原因应归咎于词包式话题模型更新策略固有的劣势,即所有相关特征参与更新过程的重排序,致使某些在话题模型中长期积累权重的特征始终占据排序前列.由此,即使这些特征对识别后续相关报道已无实质性作用,但更新过程仍会依据权重高低将其抽选作新的话题描述,使得话题模型中始终驻留大量冗余特征,成为一种空耗,且长期排挤新颖相关特征的注入.然而对于词包而言,自学习机制难以实时鉴别哪些特征在特定跟踪时段属于冗余信息,并对其进行屏蔽.尤其是某些描述种子事件的高权重特征,尽管在部分跟踪时段对捕获相关报道不发挥作用,但

因其对话题主线的收敛作用,不能随意武断地从话题模型中清除,借以避免话题描述的离散和偏差.由此,面向词包式话题模型的自学习机制往往陷入两难境地.

相对来说,结构化的话题模型 HT-DTM 和 TEC-DTM 则无须纠结于自学习过程的特征筛选问题.就 HT-DTM 而言,其相关性匹配过程通过深度遍历,查询最优的匹配路径,路径上的每个节点遍布于树状话题模型的各个层次,从而能够表述话题自宏观至具体的不同内容,因而话题不同层面的特征都能在匹配过程中发挥近似均衡的作用,避免了特征聚集于局部层次而导致的相关性匹配偏见性.尤其是,最优路径上每个节点的选择都以最大化相关性为目标,而不限制每个层次都有节点进入最优路径,从而实现了匹配过程的去冗余.因此,HT-DTM 的自学习机制可以根据话题发展脉络任意地调整话题形态,而不用顾及是否需要鉴别和屏蔽冗余信息.比如,种子事件的特征可以始终存在于 HT-DTM 的树状结构中,匹配过程可以根据相关性最大化原则,自动决定是否在特定匹配中使用这些特征.

就 TEC-DTM 而言,话题模型由不同事件的具体描述构成,每个事件都对应着特定的时序索引.相关性匹配过程依据时间的一致性决定是否匹配事件内容,且 TEC-DTM 也遵循最大相关性原则,即具有一致时间的最相关事件参与话题与报道的相关度计算.时间一致性及最大相关性原则能够保证 TEC-DTM 准确地匹配相关事件,而在特征时段扮演冗余角色的事件能够被自动屏蔽于相关性度量之外.因此,TEC-DTM 的自学习机制也可以根据话题发展趋势,自由地改变话题模型的形态,而不用顾及冗余特征的鉴别与屏蔽.从而,TEC-DTM 和 HT-DTM 的自学习机制都在保证准确匹配的前提下维护了话题结构的完整性,从而不会大量损失召回率.值得说明的是,TEC-DTM 使用了时间窗口(取值为 5),即仅仅选择话题模型中时间最晚的 5 个相关事件参与相关度计算.由此,种子事件及前期事件在跟踪后期几乎不参与相关度计算.虽然这一限制过于严格,但 TEC-DTM 获得了测试最优结果,且优于 HT-DTM 近 1 个百分点.这一结果从侧面说明:新闻话题的核往往随着新颖相关事件的出现产生明显的漂移.增强新颖事件在相关性匹配中的作用,能够有效提高话题模型的核捕捉能力.

最后,与沿用词包式描述的动态话题模型 IL-DTM 相比,融入层次树型结构的动态话题模型 HT-DTM[22]和本文提出的基于时序索引的链式动态模型 TEC-DTM,分别获得 2.3 和 3.3 个百分点的性能改进(见表 6).该结果进一步说明:话题的结构化特征(即话题内事件间的层次和时序关系)对于准确描述话题形态和演化趋势起着重要的辅助作用,有益于跟踪系统性能的改进.

7 结 论

本文重点分析了新闻话题的形态,包括结构特性和演变特性,并分类研究主流的静态和动态话题模型对新闻话题形态的拟合能力.尤其是,为了直观检验各类静态和动态话题模型跟踪话题发展脉络的性能,提出一种核捕捉衰减(CKA)观测及其数值比对分析(A^3)的评价方法.实验验证,CKA 能够直观体现话题模型的跟踪性能变化趋势,且 A^3 可通过 CKA 近似度的数值分布横向比较多种话题模型的跟踪性能.实验通过 CKA,A^3 和最小检测错误权衡系数,充分验证了动态话题模型在追踪话题演化过程中的优势.此外,本文分别提出突发式增量自学习机制(BIL)和时序事件链(TEC)动态更新方法,实验验证 BIL 和 TEC 分别获得 0.4%和 3.3%的跟踪性能改进.

目前,针对话题跟踪方向的研究已经取得显著成果,部分跟踪系统往往能够取得低于 10%的检测错误代价,基本具备实用化的条件.然而不容忽视的问题是,这类跟踪性能往往产生于不完备的新闻语料.比如,LDC 面向话题检测与跟踪任务提供的系列标准评测语料 TDT pilot~TDT5,每一期语料的采集周期最多不超过 12 个月(TDT4 由 2000 年 10 月~次年 1 月间的新闻报道构成).因此,语料中绝大部分话题的发展过程本身即不完整,从而上述低于 10%的错误代价仅仅是相对指标.尤其是,本文提供的 CKA 观测结果显示,在这种不完备的新闻语料上,现有话题模型也普遍难以拟合话题的后期发展趋势,即难以追踪话题演化.事实上,话题演化已成为现阶段话题跟踪研究的焦点问题,其对话题模型的可变形结构以及自适应学习能力提出了更高要求.然而,话题的发展趋势并非仅仅存在演化现象,往往还具有变异现象.变异现象是指,话题发展趋势彻底抛弃种子事件的主线,而以新的焦点事件为话题主线.本文取得的两项改进,即基于突发事件进行自适应学习的改进以及时序事件链在毗邻窗口内进行自学习的改进,初步验证了话题变异现象.然而,如何有针对性地检测变异锚点、如何针对变

异时的话题模型进行质心迁移、如何识别话题模型中的旧有主线特征并过滤等问题仍有待未来进一步探索.

致谢 感谢宗成庆教授对本文前期工作的支持.

References:

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

[2] Watanabe Y, Okaxta Y, Kaneji K, Sakamoto Y. Multiple media database system for TV newscasts and newspapers. Technical Report, IEIGE, 1998. 47−54.

[3] Zhang Y, Callan J. CMU DIR supervised tracking report. In: Proc. of the Workshop of Topic Detection and Tracking. 1997. 1−2.

[4] Carbonell J, Yang YM, Lafferty J, Brown RD, Pierce T, Liu X. CMU report on TDT-2: Segmentation, detection and tracking. In: Proc. of the DARPA Broadcast News Transcription and Understanding Workshop. San Francisco: Morgan Kauffman Publishers, 1999. 117−120.

[5] Masland B, Linoff G, Waltz D. Classifying news stories using memory based reasoning. In: Proc. of the SIGIR'92. Copenhagen, 1992. 59−65. [doi: 10.1145/133160.133177]

[6] Levow GA, Oard DW. Signal boosting for translingual topic tracking: Document expansion and *n*-best translation. In: Proc. of the Topic Detection and Tracking: Event-based Information Organization. Norwell: Kluwer Academic Publishers, 2002. 175−195. [doi: 10.1007/978-1-4615- 0933-2_9]

[7] Allan J, Papka R, Lavrenko V. On-Line new event detection and tracking. In: Proc. of the SIGIR'98. Amherst: University of Massachusetts at Amherst, 1998. 37−45. [doi: 10.1145/290941.290954]

[8] Yang YM, Ault T, Pierce T, Lattimer CW. Improving text categorization methods for event tracking. In: Proc. of the ACM SIGIR 2000. Athens: Association for Computing Machinery Press, 2000. 65−72. [doi: 10.1145/345508.345550]

[9] Allan J. Detection as multi-topic tracking. In: Proc. of the Topic Detection and Tracking: Event-based Information Organization. Norwell: Kluwer Academic Publishers, 2002. 139−157. [doi: 10.1023/A:1015793827697]

[10] Kumaran G, Allan J. Text classification and named entities for new event detection. In: Proc. of the SIGIR Conf. on Research and Development in Information Retrieval. Sheffield: ACM Press, 2004. 297−304. [doi: 10.1145/1008992.1009044]

[11] Nallapati R, Feng A, Peng FC, Allan J. Event threading within news topics. In: Proc. of the 13th ACM Int'l Conf. on Information and Knowledge Management. Washington, 2004. 446−453. [doi: 10.1145/1031171.1031258]

[12] Ma NL, Yang YM, Rogati M. Applying CLIR techniques to event tracking. In: Proc. of the AIRS 2004. Berlin, Heidelberg: Springer-Verlag, 2005. 24−35. [doi: 10.1007/978-3-540-31871-2_3]

[13] Larkey LS, Feng FF, Connell M, Lavrenko V. Language-Specific models in multilingual topic tracking. In: Proc. of the 27th Annual Int'l Conf. on Research and Development in Information Retrieval. Sheffield, 2004. 402−409. [doi: 10.1145/1008992.1009061]

[14] Lavrenko V, Croft WB. Relevance-Based language models. In: Proc. of the 24th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New Orleans: ACM Press, 2001. 267−275. [doi: 10.1145/383952.383972]

[15] Nallapati R. Semantic language models for topic detection and tracking. In: Proc. of the HLT-NAACL 2003 Student Research Workshop. 2003. 1−6. [doi: 10.3115/1073416.1073417]

[16] Hong Y, Zhang Y, Fan JL, Liu T, Li S. New event detection based on division comparison of subtopic. Chinese Journal of Computers, 2008,31(4):687−695 (in Chinese with English abstract).

[17] Shah C, Eguchi K. Improving document representation for story link detection by modeling term topicality. Information and Media Technologies, 2009,4(2):433−441. [doi: 10.2197/ipsjtrans.2.27]

[18] Feng A, Allan J. Finding and linking incidents in news. In: Proc. of the Conf. on Information and Knowledge Management. Lisbon, 2007. 821−830. [doi: 10.1145/1321440.1321554]

[19] Zhao H, Zhao TJ, Yu H, Zhang Z. Dynamic evolvement-oriented topic detection research. Journal of High Technique, 2006,16(12): 1230−1235 (in Chinese with English abstract).

[20] Lakshmi K, Mukherjee S. Using cohesion-model for story link detection system. In: Proc. of the IJCSNS. 2007. 59−66.

[21] Luo WH, Yu MQ, Xu HB, Wang B, Cheng XQ. The study of topic detection based on algorithm of division and multi-level clustering with multi-strategy optimization. Journal of Chinese Information Processing, 2006,20(1):29−36 (in Chinese with English abstract).

[22] Zhang K, Zi J, Wu LG. New event detection based on indexing-tree and named entity. In: Proc. of the 30th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2007). Amsterdam: ACM Press, 2007. 215−222. [doi: 10.1145/1277741.1277780]

[23] Zhang XY, Wang T, Chen HW. Story link detection based on event model with uneven SVM. In: Proc. of the 4th Asia Information Retrieval Conf. on Information Retrieval Technology. LNCS 4993, Berlin, 2008, 436−441. [doi: 10.1007/978-3-540-68636-1_44]

[24] He Q, Chang KY, Lim EP. Analyzing feature trajectories for event detection. In: Proc. of the 30th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Amsterdam, 2007. 207−214. [doi: 10.1145/1277741.1277779]

[25] He Q, Chang KY, Lim EP. Using burstiness to improve clustering of topics in news streams. In: Proc. of the 7th IEEE Int'l Conf. on Data Mining Workshops (ICDM). Omaha, 2007. 493−498. [doi: 10.1109/ICDM.2007.17]

[26] Li BL, Li WJ, Lu Q. Enhancing topic tracking with temporal information. In: Proc. of the 29th Annual Int'l ACM SIGIR. 2006. 667−668. [doi: 10.1145/1148170.1148308]

[27] Wang HZ, Zhu JB, Ji D, Ye N, Zhang B. Time adaptive boosting model for topic tracking. In: Proc. of the IEEE NLP-KE 2005. Wuhan, 2005. 488−492. [doi: 10.1109/NLPKE.2005.1598786]

[28] Hearst MA. Multi-Paragraph segmentation of expository text. In: Proc. of the 32nd Annual Meeting of the ACL. 1994. 9−16. [doi: 10.3115/981732.981734]

附中文参考文献:

[16] 洪宇,张宇,范基里,刘挺,李生.基于子话题分治匹配的新事件检测.计算机学报,2008,31(4):687−695.

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

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

洪宇(1978－),男,黑龙江哈尔滨人,博士,讲师,主要研究领域为话题检测与跟踪,个性化信息检索,舆情倾向性分析.

仓玉(1987－),男,主要研究领域为话题检测与跟踪.

姚建民(1971－),男,博士,教授,主要研究领域为机器翻译,数据挖掘.

周国栋(1967－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言理解,信息抽取,机器学习,机器翻译.

朱巧明(1964－),男,教授,博士生导师,CCF 高级会员,主要研究领域为中文信息处理,自然语言处理.