文章编号: 1003-0077(2007)06-0071-17

话题检测与跟踪的评测及研究综述

洪 宇,张 宇,刘 挺,李 生

(哈尔滨工业大学 计算机科学与技术学院 信息检索研究室,黑龙江 哈尔滨 150001)

摘 要: 话题检测与跟踪是一项面向新闻媒体信息流进行未知话题识别和已知话题跟踪的信息处理技术。自从1996 年前瞻性的探索以来,该领域进行的多次大规模评测为信息识别、采集和组织等相关技术提供了新的测试平台。由于话题检测与跟踪相对于信息检索、信息挖掘和信息抽取等自然语言处理技术具备很多共性,并面向具备突发性和延续性规律的新闻语料,因此逐渐成为当前信息处理领域的研究热点。本文简要介绍了话题检测与跟踪的研究背景、任务定义、评测方法以及相关技术,并通过分析目前 TDT 领域的研究现状展望未来的发展趋势。

关键词: 计算机应用;中文信息处理;综述; 话题检测与跟踪;自然语言处理;事件;新闻报道

中图分类号: TP391 **文献标识码**: A

Topic Detection and Tracking Review

HONG Yu, ZHANG Yu, LIU Ting, LI Sheng

(Information Retrieval Lab, School of Computer Science and Technology,
Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: Topic detection and tracking, as one of natural language processing technologies, is to detect unknown topic and track known topic from the information of news medium. Since its pilot research in 1996, several large-scale evaluation conferences have provided a good environment for evaluating technologies of recognition, collection and organization. As topic detection and tracking shares similar challenges with information retrieval, data mining and information extraction in abrupt and successive data, it has become a hot research issue in the field of nature language processing. This paper introduced the background, definition, evaluation and methods in topic detection and tracking, and explored its future development trend through analyzing current research.

Keywords: computer application; Chinese information processing; overview; topic detection and tracking; natural language processing; event; news story

1 引言

话题检测与跟踪(Topic Detection and Tracking,简写为 TDT)起源于早期面向事件的检测与跟踪(Event Detection and Tracking,简写为 EDT)。TDT 面向多语言文本和语音形式的新闻报道,主要从事报道边界自动识别、锁定和收集突发性新闻话题、跟踪话题发展以及跨语言检测与跟踪等相关任务。与 EDT 不同,TDT 检测与跟踪的对象从特定时间和地点发生的事件扩展为具备更多相关性外延的话题,相应的理论与应用研究也同时从传统对于事件的识别跨越到包含突发事件及其后续相关报道的话题检测与跟踪。

TDT 的任务以及评测体系是由美国国防高级研究计划局(DARPA)、马萨诸塞大学(University of Massachusetts)、卡耐基—梅隆大学(Carnegie Mellon University)和 Dragon Systems 公司联合制定

收稿日期: 2007-03-13 **定稿日期**: 2007-07-04

基金项目: 国家自然科学基金资助项目(60435020,60575042,60503072)

作者简介: 洪宇(1978—),男,博士生,研究方向为话题检测与跟踪,个性化信息定制;张宇(1972—),男,副教授,主要研究方向为软件容错和基于相应系统的星载计算机的设计与实现;刘挺(1972—),男,教授,主要研究方向为信息检索和自然语言处理。

和设计完成的。来自这些单位的学者历经一年的时间对 TDT 进行了前瞻性的研究(1996～1997, Pilot study)[1],包括检验当前普遍应用于信息检索(Information Retrieval,简写为 IR)和信息抽取(Information Extraction,简写为 IE)等领域的技术是否能够有效解决 TDT 问题,以及鉴定和设计统一标准的评测规范。虽然大部分 IR 和 IE 技术都可以应用于早期的 EDT,但过高的误检率说明该领域仍然具备很大的探索空间,尤其对于拓展后的 TDT 则暴露了更多现有技术的缺陷。因此探索更适合于 TDT 任务的创新性研究对自然语言领域的发展具有重要意义。

TDT 涉及两类最主要的信息获取问题,即信息的检测与集成、信息的采集与跟踪。这两方面的研究课题分别与目前信息检索(IR)和信息过滤(Information Filtering,简写为 IF)对应的问题非常相似[2]。在 IR 系统中,用户通过动态地定义需求(Query),从海量信息中检索满足自己当前兴趣的信息,信息以相关度为尺度进行组织、集成与反馈;而在 IF 系统中,用户通过定义静态的用户需求(Profile),从动态变化的信息流中实时地获取相关知识,这种知识的获取方法侧重于跟踪信息的时空进程并将最新的相关信息反馈给用户。基于这些相似点,许多基于 IR 和 IF 的信息获取技术都相应地应用于 TDT 并获得了良好的效果,尤其近期逐渐发展起来的个性化信息检索技术和自适应信息过滤技术,都与 TDT 研究具有更深层次的共性。但是,TDT 在许多方面与 IR 和 IF 存在差异,比如对于 TDT 的新事件检测任务(New Event Detection,简称为 NED),系统欠缺任何话题的先验知识,TDT 系统必须在对话题毫不了解的情况下,自主地进行识别与检测,这一点与具备了背景知识或先验需求的 IR 系统截然不同。同时,话题检测系统通常需要维护固定的存储空间保存曾经发生过的话题线索,从而作为衡量新话题的背景信息。对于话题跟踪而言,话题对应的"Query"是隐含给定的,构成话题的是若干(1～4 篇)相关报道样本,这与具备明确需求(Profile)的 IF 问题也不相同。因此,面向 IR 和 IF 的相关方法更多地作为 TDT 的基础研究,而不能完全解决 TDT 的相关问题。

本文简要介绍 TDT 任务与评测的相关知识,重点论述和分析近期国内外在该领域的相关研究及其相互关系,并在篇尾展望 TDT 领域的未来发展趋势。本文组织结构如下,第二章和第三章分别介绍 TDT 使用的语料和评价体系;第四章简要介绍话题的含义及其与事件的区别,并概述 TDT 任务的定义与要求;第五章着重探讨 TDT 研究的层次关系及体系结构;第六章和第七章分别回顾 TDT 国内和国外的研究现状;第八章概述 TDT 领域的研究趋势;第九章结论。

2 TDT语料

LDC为 TDT 方向的研究提供了五期语料,分别是 TDT 预研语料、TDT2、TDT3、TDT4 和 TDT5。TDT 语料是选自大量新闻媒体的多语言新闻报道集合。其中,TDT5 只包含文本形式的新闻报道,而其他语料同时包含文本和广播两种形式的新闻报道。本章简要介绍各语料的组成、描述及其区别。

2.1 语料组成

TDT 评测最早使用的语料是 TDT 预研语料(TDT pilot corpus,简称 TDT-Pilot)。TDT-Pilot 收集了 1994 年 7 月 1 日到 1995 年 6 月 30 日之间约 16 000 篇新闻报道,主要来自路透社新闻专线和 CNN 新闻广播的翻录文本。TDT-Pilot 标注过程没有涉及话题的定义,而是由标注人员从所有语料中人工识别涉及各种领域的 25 个事件作为检测与跟踪对象。TDT2 收集了 1998 年前六个月的中英文两种语言形式的新闻报道。其中,LDC 人工标注了 200 个英文话题和 20 个中文话题。TDT3 收集了 1998 年 10 月到 12 月中文、英文和阿拉伯文三种语言的新闻报道。其中,LDC 对 120 个中文和英文话题进行了人工标注,并选择部分话题采用阿拉伯文进行标注。TDT4 收集了 2000 年 10 月到 2001 年 1 月英文、中文和阿拉伯文三种语言的新闻报道。其中,LDC 分别采用三种语言对 80 个话题进行人工标注。TDT5 收集了 2003 年 4 月到 9 月的英文、中文和阿拉伯文三种语言的新闻报道。LDC 对 250 个话题进行了人工标注,其中 25 %的话题同时具有三种语言的表示形式,其他话题则以相同的比例均匀地分配给三种语言分别进行标注。此外,TDT5 中每种语言的话题来自该语言当地媒体的报道。

LDC 根据报道与话题的相关性对所有语料进行标注。其区别在于 TDT2 与 TDT3 采用三类标注形式,而 TDT4 与 TDT5 采用两种标注形式。前者使用"YES"、"BRIEF"和"NO"作为报道与话题相关程度的标识。当报道论述的内容与话题绝对相关时标注为"YES",而报道与话题相关的内容低于本

身的 10 %则标注为“BRIEF”,否则标注为“NO”。TDT4 与 TDT5 只采用相关“YES”和不相关“NO”对报道与话题的相关性进行标注。其中,相关报道不仅需要相关于话题的核心内容,同时需要包含话题的部分信息。但是,报道与话题相关的内容并没有 TDT2 和 TDT3 中要求的长短之分,只要存在相关信息都被标注为“YES”。

2.2　语料描述方式

TDT 语料包含两种媒体形式的数据流: 文本和广播。区别于单一表示形式的文本类新闻报道,LDC 为广播类新闻语料提供了三种信息描述方式:

(1) 数据信号的音频采集;

(2) 对音频的人工识别与记录;

(3) 通过自动语音识别系统(Automatic Speech Recognition,简称为 ASR)识别和记录音频。

此外,广播类语料不仅包含新闻形式的报道,还包含部分非新闻类报道。其中关于商业贸易的报道以及目录形式的体育比分和财经数据都属于非新闻类语料。因此,LDC 为广播类语料额外提供了三种标注形式: 新闻报道(NEWS)、多元报道(MISCELLANEOUS)和未转录报道(UNTRANSCRIBED)。其中,没有经过识别与记录的广播报道被标注为 UNTRANSCRIBED。

如前文所述,TDT 语料主要包含三种语言形式: 中文、英文和阿拉伯文。对于中文和阿拉伯文,LDC 提供了两种不同的描述方式:

(1) 本地语言描述形式,即报道采用未经过翻译的本地语言。其中包括文本形式(如新闻专线)的描述,也包括采用人工或 ASR 对本地广播的识别与翻录;

(2) 采用机器翻译自动地将中文或阿拉伯文报道翻译成英文形式。

3　TDT 评测

NIST 为 TDT 建立了完整的评测体系①。由于各个研究方向针对的问题不同以及历届评测语料的标注方案存在差异,因此 TDT 不同任务之间的评测方法、参数以及步骤不尽相同。但总体而言,评测标准都是建立在检验系统漏检率和误检率的基础之上。TDT 评测公式定义如下:

$$C_{Det} = C_{Miss} P_{Miss} P_{target} + C_{FA} P_{FA} P_{non\text{-}target} \quad (1)$$

其中,C_{Miss} 和 C_{FA} 分别代表漏检率和错检率的代价系数; P_{Miss} 和 P_{FA} 分别是系统漏检和错检的条件概率; P_{target} 和 $P_{non\text{-}target}$ 是先验目标概率($P_{non\text{-}target} = 1 - P_{target}$); C_{Det} 是综合了系统漏检率与误检率得到的性能损耗代价。检验 TDT 系统性能时,评测体系可以根据阈值或平滑系数的变化绘制检测错误权衡图(Detection Error Tradeoff,简称 DET 曲线),如图 1 是关联性检测任务中在线概念模型(Online Conceptual Model,简称 OCM)与相关性模型(Relevance Model,简称 RM)对比实验得到的一组 DET 曲线图。其横轴表示系统误检率;纵轴代表漏检率。因此,根据评测公式的定义,越靠近 DET 坐标系左下角的曲线对应的系统性能越好,即漏检和错检的综合代价相对较小。

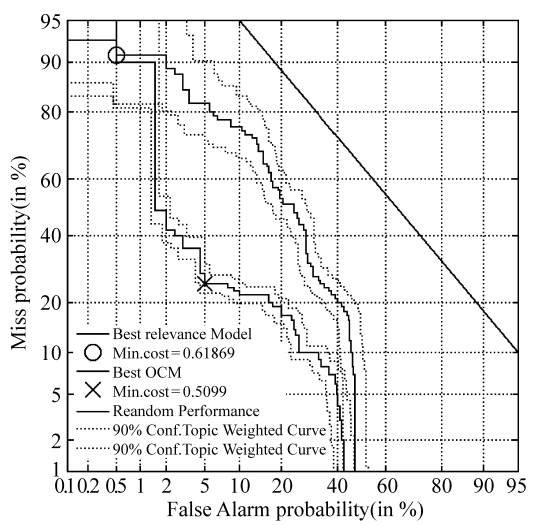

图 1　DET 曲线图样例(分别采用在线概念模型 OCM 和相关性模型 RM 实现关联性检测系统的性能对比)

评价 TDT 系统性能时常采用 C_{Det} 的规范化表示 $(C_{Det})_{Nom}$,其定义如下:

$$(C_{Det})_{Norm} = \frac{C_{Det}}{\min(C_{Miss} P_{target},\ C_{FA} P_{non\text{-}target})} \quad (2)$$

针对 TDT 涉及的语料及评测体系,本文提供了相应资源、指南及工具的获取方法和地址,其主要来源包括美国国家标准与技术研究院(简称 NIST)②和语言数据联盟(简称 LDC)③。其中 TDT 语料可通过光盘邮购和在线 LTP 下载两种方式获取,具体地址如表 1 所示。

① NIST 面向 TDT 领域的国际评测已因资金问题截止于 2004 年,但基于网络的技术性能评比仍在继续,详情可咨询 NIST 联系人 Jonathan,其联系方式为: jonathan.fiscus@nist.gov.

② http://www.nist.gov/speech/index.htm

③ http://www.ldc.upenn.edu/

表1 评测工具、指南及语料获取方式

名 称	用 途	URLP	联 系 人
DETware_v2.1.tar.gz	评测工具	http://www.nist.gov/speech/tools/index.htm	jonathan.fiscus@nist.gov
gnu_detware.tar.Z			
TDT3eval_v2.6	指南		
Dry Run Evaluation-2000	索引列表及正确答案	http://www.nist.gov/speech/tests/tdt/tdt2000/dryrun.htm	
Dry Run Evaluation-2001		http://www.nist.gov/speech/tests/tdt/tdt2001/dryrun.htm	
Dry Run Evaluation-2002		http://www.nist.gov/speech/tests/tdt/tdt2002/dryrun.htm	
Dry Run Evaluation-2003		http://www.nist.gov/speech/tests/tdt/tdt2004/dryrun.htm	
Dry Run Evaluation-2004			
LDC TDT2-TDT5	语料	http://www.ldc.upenn.edu/Obtaining/	ldc@ldc.upenn.edu

4 TDT话题定义及任务

4.1 话题定义

最初的TDT研究(TDT Pilot, 1996～1997)将话题定义为"事件"。事件是发生在特定时间和地点的事情。比如,"2001年9月11日针对纽约世贸大厦的恐怖袭击"是一个事件,而泛指的恐怖袭击则不是。此外,事件包括可预期事件(如"政府选举")和突发事件(如"飞机失事")。从TDT2开始,话题的定义有了更加广泛的含义,不仅包含了由最初事件引起或导致发生的后续事件,同时还包含了与其直接相关的其他事件或活动。直到TDT5,话题都一直沿用如下定义。

话题定义:一个话题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。

根据话题的定义,一篇报道只要论述的事件或活动与一个话题的种子事件有着直接的联系,那么这篇报道就与该话题相关,比如关于"飞机坠毁"与"坠毁殉难者葬礼"的报道都可以认为与坠毁事件直接相关,因此可以作为该话题的一个组成部分。但话题的外延并不是无限的,比如关于"联邦航空局通过调查飞机坠毁的原因修改航空条例"的报道与飞机坠毁的话题并不相关。

4.2 TDT任务

NIST为TDT研究设立了五项基础性的研究任务,包括面向新闻广播类报道的切分任务;面向已知话题的跟踪任务;面向未知话题的检测任务;对未知话题首次相关报道的检测任务和报道间相关性的检测任务。

4.2.1 报道切分任务

报道切分(Story Segmentation Task,简称SST)的主要任务是将原始数据流切分成具有完整结构和统一主题的报道。比如,一段新闻广播包括对股市行情、体育赛事和人物明星的分类报道,SST要求系统能够模拟人对新闻报道的识别,将这段新闻广播切分成不同话题的报道。SST面向的数据流主要是新闻广播,因此切分的方式可以分为两类:一类是直接针对音频信号进行切分;另一类则将音频信号翻录为文本形式的信息流进行切分。

4.2.2 话题跟踪任务

话题跟踪(Topic Tracking Task,简称TT)的主要任务是跟踪已知话题的后续报道。其中,已知话题没有明确的描述,而是通过若干篇先验的相关报道隐含地给定。通常话题跟踪开始之前,NIST为每一个待测话题提供1至4篇相关报道对其进行描述。同时NIST还为话题提供了相应的训练语料,从而辅助跟踪系统训练和更新话题模型。在此基础上,TT逐一判断后续数据流中每一篇报道与话题的相关性并收集相关报道,从而实现跟踪功能。

4.2.3 话题检测任务

话题检测(Topic Detection Task,简称TD)的主要任务是检测和组织系统预先未知的话题,TD的特点在于系统欠缺话题的先验知识。因此,TD系统必须在对所有话题毫不了解的情况下构造话题的检测模型,并且该模型不能独立于某一个话题特例。换言之,TD系统必须预先设计一个善于检测和识别所有话题的检测模型,并根据这一模型检测陆续到达的报道流,从中鉴别最新的话题;同时还需要根据已经识别到的话题,收集后续与其相关的报道。

4.2.4　首次报道检测任务

在话题检测任务中, 最新话题的识别都要从检测出该话题的第一篇报道开始,首次报道检测任务(First-Story Detection Task,简称 FSD)就是面向这种应用产生的。FSD 的主要任务是从具有时间顺序的报道流中自动锁定未知话题出现的第一篇相关报道。大体上,FSD 与 TD 面向的问题基本类似,但是 FSD 输出的是一篇报道,而 TD 输出的是一类相关于某一话题的报道集合,此外,FSD 与早期 TDT Pilot 中的在线检测任务(On-line Detection)也具备同样的共性。

4.2.5　关联检测任务

关联检测(Link Detection Task,简称 LDT)的主要任务是裁决两篇报道是否论述同一个话题。与 TD 类似,对于每一篇报道,不具备事先经过验证的话题作为参照,每对参加关联检测的报道都没有先验知识辅助系统进行评判。因此,LDT 系统必须预先设计不独立于特定报道对的检测模型,在没有明确话题作为参照的情况下,自主地分析报道论述的话题,并通过对比报道对的话题模型裁决其相关性。LDT 研究可以广泛地作为 TDT 中其他各项任务的辅助研究,比如 TD 与 TT 等等。

随着话题检测与跟踪研究的逐步深入与发展,历次 NIST 举行的 TDT 评测都对该领域内的各项子课题提出了新的设想与方向,因此相应的评测任务也随之有所更改。比如,TDT2004 撤销了报道切分任务(SST),其原因不仅在于评测语料 TDT5 中没有包含广播类新闻报道,同时也由于应用中的大部分实例片断本身具备了良好的可区分性。此外,TDT2004 将首次报道检测任务(FSD)转换成新事件检测任务。虽然 TDT2004 对 NED 与 FSD 给与了相同的定义,但本文将这两者定义为目的不同但相互依存的任务。FSD 与 NED 的区别在于前者注重鉴别事件初次报道的时空位置,后者除此之外还需要检测更多相关于事件的报道并进行汇总。此外,TDT2004 首次提出了有指导的自适应话题跟踪(Adaptive Topic Tracking,简称 ATT)和层次话题检测(Hierarchical Topic Detection,简称 HTD)概念。

5　TDT 研究体系

自 1996 年建立 TDT 研究雏形以来,历次评测都为 TDT 研究领域内出现的新问题设立相应的评测任务,截止到 TDT2004 为止,NIST 提供的所有评测任务基本上覆盖了 TDT 领域内大部分研究课题。

TDT 的研究方向主要分为五个组成部分,即报道切分、报道关联性检测、话题检测与跟踪以及针对各项任务的跨语言技术。其中每一项研究都不是孤立存在,而是与其他研究相互依存与辅助。比如,报道切分是一项基础性研究,实际应用中的 TDT 系统必须首先保证新闻报道流得到有效切分,才能进一步完成后续的检测与跟踪任务;报道关联性检测的目的在于检验两篇报道是否论述同一话题,而话题检测与跟踪的本原问题恰是检验话题与报道之间,或报道与报道之间的相关性,因此关联性检测是承载 TDT 其他各项任务的基本平台,也是性能保证的前提条件;话题跟踪系统的主要任务是跟踪特定话题后续的相关报道,而话题检测系统则在大规模新闻报道流中识别各种未知的话题,因此话题检测实质上为跟踪系统提供了先验的话题模型,而话题跟踪则辅助检测系统完善对话题整体轮廓的描述。此外,TDT 语料以及实际应用中的新闻资源都包含多种语言形式,因此各项 TDT 研究任务都需要涉及相应的跨语言技术。总而言之,TDT 研究框架下的各项任务互相关联并统一为有机整体。根据实际应用的需要,TDT 各项任务还可以进一步划分成面向不同问题的子课题,相对完整的 TDT 研究体系如图 2 所示。

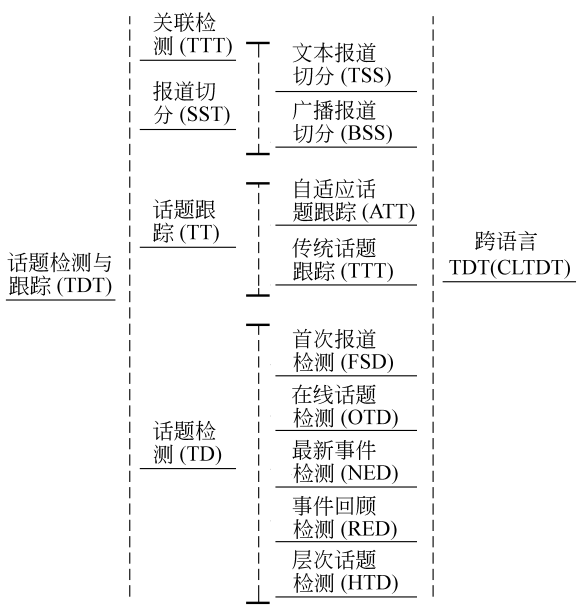

图 2　话题检测与跟踪研究体系

报道切分总体而言可以划分成两种研究子任务,一种是基于语音识别系统的报道切分,一种是基

于内容的报道边界识别。前者的识别对象是未经过翻录的广播,根据语音信号的分布规律划分报道边界;后者则将广播转录为文本形式,根据报道之间主题内容的差异估计报道边界。语音识别系统通常可以相对准确地识别边界,但是边界之间包含的信息却不一定准确地指向一个报道,往往其中包含多个报道。而基于内容的切分系统虽然可以根据话题的内涵识别出不同报道,但报道与报道之间边界的划分相对模糊。因此,如何既能公正地区分报道又能准确地定位边界是SST任务不容忽视的两个主要课题。

早期TDT中的话题检测任务(简写为TD)主要包含首次报道检测(简写为FSD)和在线话题检测(简写为OTD)两项子课题。FSD要求检测系统能够准确定位新话题出现的最初报道,OTD则不仅要求系统识别最新话题,同时需要收集该话题的所有相关报道。FSD可以看作OTD的前提:通常,新话题的首次报道构成该话题的最初描述,后续报道相关性的裁决都以该报道为对照标准,即使随着相关报道逐渐增多,话题模型的质心相应发生漂移,但是话题的主线并没有脱离首次报道描述的内涵。相反,OTD是对FSD的补充:新话题不仅包含对其进行报道的第一篇文本,同时也包含后续与之直接相关的外延,只有综合所有相关报道才能完整地勾勒出对应的话题。

近期,TD研究领域得到进一步拓展。其中,TDT2004设置了新事件检测(简称为NED)任务,NED要求检测系统能够针对具备时间顺序的新闻语料及时地检测出最新发生的事件。NED与FSD面向的问题非常类似,区别在于检测对象从话题具体化为事件,其原因是某些话题跳跃式出现的特性,即话题在消失一段时间后重现并起源于一个新的事件。比如关于"恐怖袭击"的话题包括2001年"9.11"自杀式炸弹袭击;2002年印度尼西亚的巴厘岛惨案和2004年马德里系列爆炸案等。其中,历次恐怖袭击都是一个种子事件并伴随大量相关报道,因此话题在不同时间由不同事件多次引发,从而成跳跃式地出现。话题的这一特性引起了关于TD研究的两种思考,即怎样区分不同事件引发的相同话题;是否当前被检测到的话题在历史上从未出现过。NED就是面向第一种思考提出的检测任务,区别于传统的FSD系统,NED更关注特定时间与地点发生的最新事件。此外,Yiming Yang[30]提出一种回顾式话题检测(简称为RED)的研究方向,目的在于回顾历史上所有报道,检测与话题相关的所有事件。由此,NED与RED补充了TD研究中出现的上述两项课题。

TDT2004设置的另外一项新任务是层次话题检测(简称为HTD),目的在于区分报道内容在层次上的差异,从而建立结构化的话题模型。总体而言,话题检测研究的发展逐步面向结构化和层次化,TD系统不仅需要善于识别话题和收集相关报道,同时需要有效分析话题内部的层次结构、区分不同组成部分并挖掘外界的相关历史信息。

区别于未知话题识别的TD系统,话题跟踪(简称为TT)的主要任务在于跟踪已知话题的后续报道。通常,突发事件的产生会引发大量相关报道,随着事件受关注程度的降低,相应报道逐渐衰减直至消失。在这个过程中,话题在不同历史阶段的论述重心将有所漂移。比如,2001年"9.11"事件发生的最初一段时间内,大量报道主要集中于事件本身,包括"客机撞击世贸"、"世贸大厦损毁"以及伤亡情况统计;随着事态的发展,相关报道的重心逐渐转移到"灾后处理"、"事件调查"和"美国民众的反应";最后话题集中于"恐怖主义"、"反恐战争"以及"世界范围内的反恐政策"等等。因此,一个完整的话题不仅包括最初事件的相关报道,还涉及后续相对拓展的外延,TT任务就是面向这一问题提出的。TDT2004设置了有指导的自适应话题跟踪任务(ATT),其与传统TT系统的区别在于嵌入了自学习机制,可以使跟踪系统实时地依据话题的发展自动更新话题模型,从而有效追踪话题的报道趋势。

6 TDT国外研究现状

6.1 关联检测(LDT)

LDT的主要任务是检测随机选择的两篇报道是否论述同一话题。与其他TDT任务不同的是LDT研究并没有直接对应的实际应用,但是它对其他TDT研究起到的辅助作用却是无法忽视的。比如,新事件检测任务(NED)中,NED系统可以通过LDT鉴定候选报道与每个先验报道之间的相关性,从而判断候选报道是否论述了一个新话题,或者相关于先验报道隶属的旧话题。就传统基于概率统计的TDT研究而言,报道与话题或者报道与报道之间的相关性,都是通过检验两者之间共有特征的覆盖比例进行评判。换言之,两者共有的特征越多,那么它们相关的可能性越大。因此,大部分针对LDT的研究都将问题的重心集中于文本描述以及特征选

择。James Allan[3]和 Schultz[4]采用向量空间模型(简称为 VSM)描述报道的特征空间,根据特征在文本中的概率分布估计权重,利用余弦夹角衡量报道之间的相似性。此外,Leek[5]和 Yamron[6]将参与检测的两篇报道分别看作一个话题和一篇报道,采用语言模型(简称为 LM)描述报道产生于话题的概率,并通过调换两篇报道的角色分别从两个方向估计它们的产生概率,最终的相关性则依据这两种概率分布,采用 Kullback-Leibler Divergence[7](简称为 KLD)算法综合得出。VSM 和 LM 存在的主要缺陷在于特征空间的数据稀疏性,通常解决这一问题的方法是数据平滑技术,但是平滑得到的特征权重往往被泛化,从而无法有效描述文本内容上的差异。另一种解决数据稀疏的方法是特征扩展技术。在信息检索中,特征扩展主要应用于 Query 扩展,其核心思想是将 Query 中的特征扩展为同义或直接相关的其他特征,从而降低稀疏性。Ponte 和 Croft[8]采用向量空间模型,并基于特征上下文的扩展技术执行 LDC 任务,其选择待测报道中权重较大的特征作为扩展对象,通过围绕特征经常出现的上下文信息[9]对其进行扩展,特征空间由原始和扩展的特征项共同组合而成。扩展技术不仅有助于解决数据稀疏问题,同时可以辅助 LDC 系统削弱特征的歧义性。

6.2　话题跟踪(TT)

6.2.1　传统话题跟踪(TTT)

传统话题跟踪(Traditional Topic Tracking,简写为 TTT)主要包括基于知识和基于统计的两种研究趋势。前者的核心问题是分析报道内容之间的关联与继承关系,通过特定的领域知识将相关报道串联成一体。后者则根据特征的概率分布,采用统计策略裁决报道与话题模型的相关性。

基于知识的 TTT 研究中,比较有代表性的方法是 Watanabe[10]面向日本语新闻广播开发的话题跟踪系统。Watanabe 通过形如"正如我所提到的……"、"正如我所报道的……"和"正如近期发生的……"等领域知识,检测论述同一话题的相关报道。该方法能够显著提高特定知识领域的话题跟踪性能。

基于统计策略的 TTT 研究则主要借鉴于基于内容的信息过滤(简称为 IF)。如前文所述,IF 面向静态需求从动态的信息流中识别和获取相关知识,TTT 则根据先验的话题模型追踪后续相关报道。虽然 TTT 更关注突发事件的识别与跟踪,但任务整体框架的相似性决定了 IF 中的许多相关技术都可以有效地应用于 TTT。其中最有代表性的方法是基于分类策略[11,12]的话题跟踪研究,比如 CMU[13,14]在 TTT 评测中采用了两种分类算法,分别是 k-最近邻(k-Nearest Neighbor,简写为 KNN)和决策树(Decision tree,简写为 D-tree)。其中,KNN 首先根据内容的相关性选择当前报道最相似的 k 个先验报道作为最近邻,然后根据最近邻所属话题类别综合判定当前报道论述的话题;D-tree 则根据训练语料预先构造话题的决策树,该树型结构中的每个中间节点代表一种决策属性,即报道相关于话题的条件,节点产生的分支则分别代表一种决策并指向下一层子节点,决策树的叶节点代表话题类别,输入决策树的待测报道经过逐层节点的判断,最终划分于特定话题类别。KNN 与 D-tree 面临的主要问题是先验相关报道的稀疏性,TTT 任务一般只给定少量相关报道作为训练(1～4 篇)。稀疏性造成 KNN 算法无法使待测报道的最近邻涵盖大量正确的相关报道,从而根据这些近邻得到的判断往往指向错误的话题模型;而 D-tree 则在训练过程中无法为每个属性节点嵌入准确的决策条件。总体而言,KNN 的性能优于 D-tree,其原因在于前者可以通过缩减最近邻的规模保证跟踪的正确率;而后者则受限于多层属性需要同时产生正确的决策,而相关报道稀疏的训练语料使多数属性本身不够准确(比如 Bigram 的概率统计),因此在没有改进漏检率的情况下加大了误检率。

UMass[15]采用二元分类方法跟踪话题的相关报道。UMass 借鉴了 ODT 的相关研究,即陆续到来的后续报道或者与已有话题相关,或者论述的是新话题。基于这种假设,二元分类将训练语料划分为相关和不相关两种报道类别,并根据两类报道与话题相关性的概率分布训练线性分类器[16,17],后续报道的相关性依据线性判别式进行裁决。二元分类方法的优点在于精确率很高,但必须依赖训练语料和分类器的选择,通常选择相关度指标较高的不相关报道构成反例类别,从而保证分类面的灵敏度;分类器的选择则必须确保线性判别式在训练过程中有解,而整体性能可以通过 Boosting 算法[17]进行提高。与 KNN 和 D-tree 类似的是,先验相关报道的稀疏性一定程度上影响了二元分类方法的召回率,UMass 相应地采用 Query 扩展技术完善了这一缺陷。

James Allan[2]和 Michael[18]采用 Rocchio 算法

实施跟踪。Rocchio[2]的核心思想是话题模型经验性的构造策略,即假设相关报道中的特征有助于话题的正确描述,因此这些特征在话题模型中的权重被加强,而不相关报道中的特征则趋向于错误地引导话题描述,因此权重被削弱。Rocchio算法的最大优点是:TTT系统可以利用跟踪到的后续报道不断改进和更新话题模型,从而跟踪话题的后续发展。缺陷在于Rocchio算法对阈值的依赖程度很高:如果初始阈值设置过高,则后续相关报道的漏检率加大;如果阈值设置过低,将引入大量噪声。其中,后者对TTT性能造成的损失最大,因为大量噪声直接误导话题模型的更新,从而导致跟踪方向的偏差。

其他面向TTT的研究工作还包括话题与报道的相似度匹配算法,比如Dragon[25]分别通过基于一元语言模型的文本相似度匹配[19]和基于二项式的相似度匹配[20]衡量话题与报道的相关性。而Franz和Carley[21]则尝试采用聚类方法将话题检测系统转化成跟踪系统。近期,Yiming Yang[22]和Larkey[23]分别采用小规模的先验报道翻译模型和源语言模型进行跨语言TTT研究。上述方法对于传统的话题跟踪任务能够发挥较好的作用,但由于构造话题模型的初始信息相对稀疏,因此无法有效跟踪一段时期以后话题的发展。

6.2.2 自适应话题跟踪(ATT)

如前文所述,NIST为话题跟踪任务仅提供1～4篇相关报道用于构造话题模型。类同的是,实际应用中的用户对突发性新闻具备的先验知识通常也很少,这就造成初始训练得到的话题模型不够充分和准确。因此,一种具备自学习能力的无指导自适应话题跟踪(Adaptive Topic Tracking,简写为ATT)逐渐成为TT领域新的研究趋势。总体而言,ATT的相关研究主要包括两个方面,即基于内容和基于统计的方法。

在基于内容的ATT相关研究中,GE R&D[24]尝试采用文摘技术跟踪话题的发展趋势。其核心思想是分别提取话题与报道的文摘代替全文描述,话题与报道之间的相关性通过文摘之间的相似度进行计算。通常,话题的相关报道在不同历史时期的侧重点不尽相同,因此话题的发展以初始事件为主线,并以后续直接相关的其他事件和活动为延续。基于这一特点,GE R&D将先验相关报道中的事件主体和相关外延以文摘的形式进行提取与组合,根据这种方法构造的话题模型除了涵盖主题信息以外,更注重话题发展的层次结构,从而使跟踪系统更善于检测话题的后续进展。其缺陷在于,GE R&D的跟踪系统没有嵌入自学习机制,话题模型没有利用检测到的后续相关报道自适应地进行更新。因此,当跟踪进行到一定阶段后,系统无法识别最新的相关报道。

基于统计策略的ATT研究主要借鉴于自适应信息过滤。核心思想是ATT系统可以根据伪相关反馈对话题模型进行自学习,不仅为话题嵌入新的特征,同时动态调整特征权重。其优点在于削弱先验知识稀疏造成的话题模型不完备性,并通过不断自学习提高ATT系统跟踪话题发展的能力。Dragon[25]和UMass[26]是最早尝试无指导ATT研究的单位之一。其跟踪系统每次检测到相关报道,都将它嵌入话题模型并改进特征的权重分布,后续报道的相关性则以新生成的话题模型为评估对象,从而实现跟踪系统的自学习功能。Dragon[25]与UMass[26]的区别在于,前者把系统认为相关的报道嵌入训练语料,并基于语言模型构造新的话题模型;后者则将所有先验报道的质心作为话题模型,并将先验报道与话题模型相关度的平均值作为阈值,后续跟踪过程中每次检测到相关报道,都将其嵌入训练语料,并根据上述方法重新估计话题模型和阈值。总体而言,这两种方法并没有很大程度地提高话题跟踪系统的性能。其主要原因在于自学习模块对于跟踪反馈不施加任何鉴别地全部用于话题模型的更新,而系统反馈本质上是一种伪反馈[27],即同时包含相关报道和不相关报道,因此学习过程将大量不相关信息也嵌入话题模型,从而导致话题漂移[14,28]。基于这一现象,LIMSI[29]在原有自学习过程中嵌入二次阈值截取功能,通过设置一个比阈值更高的过滤指标,截取伪反馈中相关度较高的报道嵌入话题更新模块,从而削弱了话题漂移。通常,ATT自学习过程中的核心问题是特征权重的更新策略,LIMSI[29]比较了基于静态和动态两种方式的权重更新策略:前者对权重的更新指标乘以经过训练的固定参数;后者将报道与话题的相关度映射为线性函数,特征权重根据线性函数动态确定。

该方法的特点在于话题每次更新后,特征权重基于话题模型的条件概率都相应得到改进。此外,动态更新机制优于静态更新的另一个原因在于,前者的特征调整融和了报道与话题模型的相似度,并且所有伪反馈都可以参与更新;而后者则独立地根据概率分布估计权重,并且必须依靠经验性的阈值,截取最相关的报道参与更新,因此在没有明显提高精确率的同时,大量损失召回率。

目前,话题跟踪的相应研究已经取得很好的效果,但如何更有效地追踪话题的后续发展仍然是该领域有待深入研究的课题。近期更多的研究集中于相关报道的概率分布和话题随时间衰减趋势的估计。未来的研究重心在于如何有效利用新闻语料的时间特征,并分析话题发展在时间轴上的分布。

6.3　话题检测(TD)

6.3.1　在线话题检测(OTD)

在线话题检测(On-line Topic Detection,简写为OTD)的主要任务是检测新话题并收集后续相关报道。通常,OTD系统的检测原理集中于相关报道的聚类算法,即在线监视后续的报道数据流,如果截获与之前聚类得到的话题不相关的报道,则检测到一个新话题,否则将该报道融合于相关聚类。对于OTD的早期研究主要集中在聚类方法的选择与融合上。比如,参加在线话题检测任务的所有单位都尝试使用单路径聚类算法对新话题进行检测。此外,CMU同时尝试采用凝聚层次聚类算法进行检测[30],但是获得的效果略差于单路径聚类。而Papka[31]则对比了不同聚类算法在OTD中的效果并尝试融合各自的优点解决OTD问题。

6.3.2　新事件检测(NED)

正如TDT研究体系中所提到的,FSD任务忽视了话题出现的跳跃性,从而使检测到的新话题经常是某些已知话题在不同时期出现的相关事件。因此,新事件检测(New Event Detection,简写为NED)逐渐成为辅助话题检测(TD)的重要组成部分。NED与首次报道检测(First Topic Detection)任务很相似,唯一的区别在于前者提交的最新事件可能相关于历史上的某一话题;后者必须输出话题最早的相关报道。NED中的主流方法来自于James Allan[32]和Yiming Yang[33],他们通过建立一个在线识别系统(OL-SYS)检验报道流中新出现的事件。其中,陆续进入OL-SYS系统的报道需要与每个已知的事件模型计算相关度,并根据先验阈值裁决报道是否为新事件的首次报道,如果条件成立则根据该报道建立新的事件模型,否则将其嵌入已知事件模型。后期NED的相关研究以这种统计方法为框架,涉及两个方面的改进,即建立更好的文本表示形式和充分利用新闻语料的时间特征。

传统的NED研究采用基于统计原理的文本表示形式,其中最常用的表示方法是向量空间模型(VSM),事件模型与报道的相似度计算则相应地采用余弦夹角和Hellinger距离公式[34]。统计模型的缺陷之一在于事件空间中的噪声信息对新事件检测造成的负面影响。基于这一问题,Yiming Yang[33]采用分类技术将先验的报道划分为不同类别,区别于将类别中的所有相关报道作为事件描述,Yiming Yang只选择每个类别中最优的相关报道描述事件模型,基于这种方法的NED系统在性能上获得了显著的提高。

统计模型的最大缺陷在于无法有效区分同一话题下的不同事件。前文曾经提到,话题经常被不同事件触发而重复出现,因此话题描述的是所有相似事件具备的共性,而事件之间的区别则集中于时间、地点和人物等实体之间的异同。仍然以"恐怖袭击"话题为例,其包括2001年"911"自杀式炸弹袭击事件;2002年印度尼西亚的巴厘岛惨案和2004年马德里系列爆炸案等。从内容上分析,这些事件的相关报道中都会频繁出现"恐怖分子"、"自杀式"、"袭击"、"损毁"和"死亡"等特征,并且这些特征在报道中出现的频率相对最频繁。因此,根据传统基于统计的策略,这些特征往往构成事件模型的主体,从而无法有效区分同一话题框架下的不同事件。与此不同的是,以名实体为主的特征集合,如"9.11"、"美国"、"巴厘岛"和"马德里"等,对于不同事件的区分贡献度更高。由此,Kumaran[35]、James Allan[36]、Yiming Yang[37]和Lam[38]等学者使用自然语言处理(NLP)技术辅助统计策略解决NED问题。其中最常用的NLP技术是命名实体(Named Entities,简称为NE)识别。比如Kumaran[35]以Yiming Yang的分类方法为统计框架,将报道描述成三种向量空间,分别为全集特征向量、仅包含NE的特征向量和排除NE的特征向量。最终Kumaran对比了三种向量空间模型对新事件检测的影响,并验证NE极大地促进了事件之间的区分。

NED研究应用时间特征的方式有两种,一种是基于文档输入的时间顺序,采用KNN分类技术;另一种是采用时间为参数的衰减函数[30,34]改进基于内容的相关度计算方法。这些研究在一定程度上提高了NED系统的性能。因此,NED未来的研究趋势将以区分话题与事件在时间轴上的概率分布为主线,并辅以NLP与统计策略相结合的事件与报道描述方法。

6.3.3　事件回顾检测(RED)

事件回顾检测(Retrospective News Event De-

tection,简写为 RED)的主要任务是回顾过去所有发生过的新闻报道,并从中检测出未被识别到的相关新闻事件。对于 RED 研究方向的理解必须涉及到事件与话题的定义。前文曾经提到事件是发生在特定时间和地点的事情,而话题则不仅包含作为种子的事件或活动,同时也包含与其直接相关的事件与活动。因此,RED 的任务实际上是辅助话题检测系统回顾整个新闻语料,从中检测相关于某一话题却并未被识别到的一类新闻事件。RED 研究的必要性来源于话题波动出现的特性。比如,CNN 关于"圣诞前夜"的话题在每年的圣诞前夕都会成为新闻与广播最关心的事件,其相关报道的概率分布如图 3 所示。因此,同一话题跳跃式地出现于不同时间,并且每次出现都伴随着大量相关报道。基于新闻语料的这种特性,话题检测系统往往只能识别出局限于一个时期的事件,而构成话题的全部事件并没有有机地结合起来,而是独立地作为一个话题被误检。RED 研究就是面向话题检测系统的这种缺陷提出的。

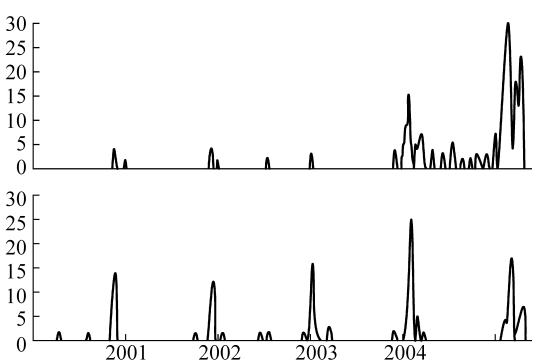

图 3 MSNBC(上)和 CNN(下)关于万圣节前夕的报道分布

首次提出 RED 研究并给予定义的学者是 Yiming Yang[30]。其采用凝聚式聚类算法与批平均聚类算法相结合的策略,将近似于同一话题模型的相关事件综合在一起作为话题检测的结果,从而使 TD 系统具备了回顾相关事件的能力。此外,Li[39]采用基于内容和时间的联合概率模型构造话题空间,从而有效识别话题在不同历史时期涉及的相关事件。虽然独立于 RED 方向的相关研究较少,但由于 RED 与 NED 中都涉及到未知事件的识别与发现,因此许多学者尝试使用 NED 中的相关研究同时处理 RED 问题。

6.3.4 层次话题检测(HTD)

TDT2004 定义了一项新的话题检测任务:层次话题检测(Hierarchical Topic Detection,简写为 HTD)。HTD 是面向话题检测中两种不恰当的假设提出的,其中一个假设是所有报道与相关话题的近似程度都在一个层次上,而另一个假设是每篇报道只可能相关于一个话题。实际上,报道的主题与话题的相关程度往往分布于不同层次,比如"人民币升值"和"建行、交行上市"两篇报道,虽然它们都相关于同一话题"2005 中国十大金融事件",但是主题侧重点的差异造成它们与话题的对应程度处于不同层次。此外这两篇报道都可以分别划分到"汇率制度改革"类和"商业银行改革"类的话题模型当中,因此报道不总是仅仅相关于一个话题,往往不同话题的相关报道存在交集。HTD 通常可以采用基于一个根节点的非循环有向图(Directed Acyclic Graph,简写为 DAG)描述话题包含的层次结构[40]。其中,根结点抽象地代表所有话题;沿有向图方向延伸的子节点则描述比父节点更具体的一类话题。因此,HTD 的主要任务是检测经过聚合得到的 DAG 体系中,每个话题的聚类效果,以及根节点与该话题之间路经的复杂度。映射为实际应用则是检验 HTD 系统是否能够辅助用户通过最便捷的查询获得最优的一类报道。

一种解决 HTD 的方法是凝聚层次聚类算法(Hierarchical Agglomerative Clustering,简写为 HAC)。其核心思想是计算当前聚类集合中每对聚类的相关度,将满足阈值条件的一对聚类融合成新的聚类,通过反复迭代这一过程,系统最终把话题模型构造成具有层次关系的 DAG。HAC 的一个重要的缺陷是时间和空间复杂度过高,比如 TDT5 总共包含 400 000 篇报道,直接采用 HAC 的时间和空间复杂度分别为 $O(n^2\log(n))$ 和 $O(n^2)$,仅存储空间就需要 80gigabytes,因此 HAC 不适合直接应用于 HTD。对 HAC 的一种改进方案是混合聚类算法,相关的研究来自 Cutting[41]。HAC 的另一种改进来自 TNO[42]的增量式层次聚类算法,其首先随机抽取小规模样本通过层次聚类构造初期的 DAG 体系,然后将不对称的聚类结构通过二次分支进行优化,最后将其余报道根据相关度大小融合于 DAG 体系,其中相关度大于特定阈值的报道被嵌入 DAG 中已有的话题,而相关度小于特定阈值的报道则确定一个新的话题结构。TNO 的增量式策略在不损失聚类性能的同时,降低了由根节点检测到话题的复杂度。

6.4 跨语言 TDT

TDT 研究面对的信息是包含多种语言的新闻

报道。无论是基于语料本身的语言多样性,还是面向实际应用的需要,TDT 的相关课题都需要涉及跨语言领域的相关研究。NIST 为 TDT 的评测提供了机器翻译(Machine Translation,简称为 MT)功能,基于不同语言的语料可以通过 MT 相互转化,从而由源语言和翻译语言共同组成形式统一的多源单一语言[23](Multiple Language-specific,简称为 MLS),比如英文语料以及翻译成英文形式的中文语料。因此大多数参加 TDT 评测的系统都是基于 MLS 的语言环境,对话题与报道模型进行描述。随着跨语言技术的发展,包括 James Allan[43]、Leek[6]和 Levow[44]在内的一些学者尝试采用不同的翻译策略解决 TDT 研究中的跨语言问题,并比较了机器翻译和其他翻译技术在 TDT 中的效果。这些研究的主要贡献在于规范化了基于翻译语言模型的相关度计算,从而削弱错译对系统整体性能的影响,但是这些工作仍然是一种面向单一语言符号的统计策略,而每种源语言本身具备的结构和上下文关系,以及特征的实际内涵都不能通过翻译的手段有效识别。

基于上述问题,目前跨语言 TDT 的核心问题是怎样在面向多语言信息时,使系统能够在不脱离任何一种语言的本原环境下运行。针对这一需要,UMASS 的 Larkey[23]尝试采用源语言模型解决跨语言问题。他首先建立了本地语言假设(Native Language Hypothesis,简称为 NLH),其核心内容是: 组成两篇报道内容的特征如果来自同一种源语言,那么针对这两篇报道之间的任何匹配算法,都只能在基于源语言的情况下才能获得最优的效果,而不是经过翻译的其他语言。TDT 中所有任务都涉及的一个基本问题是信息与信息之间相关性的衡量与评价。因此,NLH 可以广泛地运用于 TDT 中各项课题的跨语言研究。以话题跟踪(TTT)任务为例,话题只有很少的训练样本作为先验知识,并且这些训练样本都采用同一种语言进行描述,而后续报道流的描述语言则是多样的。这就给基于 NLH 的跨语言跟踪造成了困难,因为 NLH 要求参与匹配的报道对象,必须采用同一种源语言进行描述。Larkey 的解决办法是在系统运行初期采用机器翻译将报道转换成与话题模型相同的语言形式,如果检测到相关报道并且该报道的源语言与话题模型不相同,则将该报道作为话题模型新的训练样本并采用源语言进行描述。基于这种方法,话题模型的结构由不同语言形式的子结构共同组成,后续的报道流可以在满足 NLH 的假设下与话题模型进行匹配。这种方法的缺陷在于,源语言结构的性能对最初通过机器翻译得到的相关报道依赖性很强,如果机器翻译为源语言结构提供了错误的训练样本,那么即使后期的报道流可以在本源特征环境下进行匹配,也会因为话题模型的偏差被误导。

此外,Jin[45]采用统计策略解决跨语言问题。其核心思想是: 特征空间的上下文本身蕴含了源语言的语义信息,从而可以代替 MT 解决 TDT 的跨语言问题。该方法中没有涉及到文本的机器翻译,而是把文本描述成由独立特征组成的集合,而这些特征都在一种语言形式下进行表示。基于这种语言环境,Jin 采用 Bayesian 算法匹配话题与报道的相关度。Jin 的方法在性能上略优于采用 MT 的匹配算法。其原因在于语言的多义性往往使特征无法得到 MT 的正确翻译,从而误导文本匹配。但是,完全基于统计策略的跨语言方法仍然无法获得更大的提高,因为特征空间的上下文虽然蕴含了语义信息,但也给文本的描述引入了大量不相关的噪声。因此,Leek[46]采用自然语言信息与统计策略相结合的方式对其进行改进,其利用特征所在的上下文以及词典知识描述特征: 对非英文文本提取出现频率最高的若干特征,通过词典查找特征对应的英文含义,并在这个基础上通过英文语料背景获取特征的上下文及其权重。因此,每个非英文特征都是通过它在词典中对应的所有英文特征,以及这些英文特征在英文语料中的上下文统计而成。基于这种方法,TDT 系统的跨语言性能获得了明显的提高。

7　TDT 国内研究现状

TDT 作为信息处理领域新颖的研究分支逐步成为国内重要的研究热点。相比于国外以统计概率模型为主体的研究趋势,国内的相关研究更侧重基于 TDT 本身的特色进行探索。TDT 处理的信息是面向真实新闻事件的报道,其语义描述的精确性和可区分性更依赖于实体元素[47];此外,事件的产生和后续发展包含了报道之间的时序关系,其要求 TDT 系统不能单一基于内容建立相应的话题模型,而是融合时序特性参与检测报道间的关联性和跟踪话题的演化趋势[48]。在此基础上,国内的相关研究也面向建立结构化和层次化的话题模型进行了初步尝试。

名实体是描述话题或报道语义的一类特殊语言单位,其对于精确刻画核心内涵和区别不同主题具有重要意义。TDT 系统应用名实体改进性能的方

法主要包括如下两方面：⑴名实体特征权重的再分配，即希望区别名实体与其他特征对语义描述的能力；⑵名实体相关性与其他特征相关性的线性组合，即希望通过人工或自动的方式调整名实体在相关性匹配过程中发挥的作用。国内较早将名实体融入TDT系统的研究来自贾自艳[49]，其将文本内的特征标记为人名、地名和主题信息等类别，并预先指定每种特征类别的价值系数，特征的最终权重为词频和其所属类别价值系数的乘积。赵华[50]则通过分析英文写作的习惯，自动识别新闻报道中首字母大写和全部大写的特征，其认为该类特征不仅包含名实体，也包含报道需要重点强调的特征，并在此基础上采用相关度加权和的方式评估话题与报道的相关性。上述方法在一定程度上改进了TDT系统的性能，但由于是经验性地分配权重或调整相关性线性比例，因此无法保证系统性能的稳定性。张阔[51]基于χ^2分布统计TDT2中各名实体类别(人名类、地名类和机构名类等)与各话题类别(金融类、自然灾害类和科技类等)的关联性，并将这一关联性的量化指标融入特征权重的再分配，其在提高NED系统性能的同时确保了这一改进的稳定性。限制名实体在TDT领域性能的另一因素是义同形不同的实体无法匹配。针对这一问题，宋丹[52]面向地点类名实体建立地理树，匹配过程基于两名实体在地理树中路径的覆盖率进行计算，如“北京”在地理树中的路径“亚洲—中国—北京”与实体“北平”的路径基本一致，其高覆盖率可以有效匹配两实体之间的关联性，但该方法因无法处理诸如人名类等其他实体而存在局限性。在此基础上，骆卫华[53,54]基于概念一致性匹配同义的名实体，其通过建立别称表和后缀表识别不同形态的名实体是否隶属于同一概念，如通过别称表识别“李光耀”和“李资政”为同一概念；而基于后缀表识别“江苏省”和“江苏”为同义实体，该方法的缺陷在于依赖词典的规模和训练语料的新旧，对于报道流中最新出现的名实体依然无法匹配。

如前文所述，话题起始于种子事件并包含后续相关事件，而构成事件描述的一项重要特性是其产生的时间，因此话题模型内各相关报道之间往往具备时序关系。国内将时序融入TDT领域的主要策略是将其作为相关性评估的附加元素，通过线性加权的方式调整相关度指标。贾自艳[49]建立了统一时间表述方式的机制，在此基础上将当前报道与话题框架下新近事件的时间取差值，并利用该指标削弱基于内容匹配获得的相关度，其基本思想是：报道与事件时序关系越近，则它们相关的概率越大。该方法提高了TDT系统检测与跟踪话题演化趋势的性能，与其相似的工作是赵华[55]面向话题演化边界识别的研究，其训练一项表征话题演化周期的阈值，检测后续报道与话题模型内最新事件的时间差是否高于该阈值，将满足这一条件的报道作为话题演化的边界，该方法同样改进了TDT系统的性能。但由于上述方法或基于经验性的假设，或依赖于训练语料的规模，因此不能确保系统性能的稳定。有助于解决这一缺陷的研究来自宋丹[52]的时间“覆盖矩阵”，其将相关性匹配双方的时间信息统一为标准格式，并分别映射于横纵时间轴上的点，基于对角线检测所有同步点及其时间间隔，在此基础上以所有间隔的覆盖率描述匹配双方时序关系的相似性。该方法可获得相对稳定的性能提高。但如6.3.3节所述，话题的出现存在跳跃性，即在较长历史时间段内，同一话题在一定周期内规律性的出现。这一现象限制了上述假设，即时序关系较近或匹配双方包含较多近似时间信息则相关性较高。因此，国内在TDT领域应用时序关系的研究仍有较大可提升的空间。

话题模型层次化和结构化是目前TDT领域重要的研究方向。其中，层次化面向将同一话题下的相关报道组织为宏观到具体的层次体系；结构化则侧重挖掘和表征同一话题的不同侧面。国内尝试建立层次化话题模型的研究来自骆卫华[53]和张阔[51]，前者首先基于时序关系对报道分组，然后进行组内自底向上的层次聚类，最后按时间顺序采用单路径聚类策略合并相关类；后者则面向报道全集建立层次化的索引树，树中第一层节点对应特定话题，而其子树则描述了该话题的层次体系，其建树过程基于输入的报道相对于树中各层次节点是否为新事件进行组织。上述两种策略在改进检测性能的同时也提高了系统效率，但在如何基于层次关系刻画话题语义及其演化趋势方面仍需要更深入的探索。针对结构化话题模型的研究来自赵华[55]和金珠[56]。前者分别尝试基于时序和特征分布密度识别话题演化的边界，在此基础上以演化边界为划分将话题描述为初始质心和当前质心两项子结构，后续报道与话题的相关性取自其与两项质心的相关性最大者。后者则对话题内的相关报道进行聚类，抽取聚类中的特征建立事件框架以描述话题的不同侧面，此外其通过HowNet建立事件内的情态关系和角色框架，扶助描述话题不同侧面的倾向性。该两种方法提高了话题跟踪系统的性能，尤其前者对话题演化趋势的

识别和描述提高了跟踪系统的实用性。

总体而言,国内相关研究侧重挖掘 TDT 领域的特性,在方法上注重统计策略和自然语言处理技术相结合,在研究趋势上逐步面向融入数据挖掘、事件抽取和篇章理解等相关技术。此外,国内相关研究也朝着更加细化(如话题演化趋势的识别与跟踪)和更加实用化(如复旦大学媒体计算与 WEB 智能实验室的"互联网舆情分析跟踪系统"①)的方向发展。

8　TDT 研究的发展趋势

基于概率模型以及自然语言处理技术的信息描述与匹配方法在 TDT 中得到广泛应用: 前者利用特征的概率分布以及特征之间的共现率等统计信息描述文本,后者则利用特征的语言学信息描述文本,比如词性、词义、命名实体和指代关系等。TDT 采用最多的概率模型包括向量空间模型(VSM)[23]、语言模型(LM)[57,58]和相关性模型(RM)[60,61]。概率模型通过分析特征在信息集中的概率分布建立话题与报道的描述,并采用机器学习(ML)的相应策略匹配特征空间的相关性。这种方法的缺陷在于忽视了特征自身携带的语言信息,同时也遗漏了短语级、句子级和篇章级的结构与层次。此外,概率模型只将特征出现的频率和特征之间的共现率作为评价权重大小的标准,但自然语言中的指代关系、一词多义和名词短语等现象却并不支持这一理论。随着 TDT 的发展,更加智能化的自适应学子机制成为领域内的研究热点,这就对 TDT 系统正确理解知识提出了更高的要求,而传统基于统计策略不能真实地描述其语义空间,因此基于 NLP 技术及其与统计学原理相融合的相应研究将逐步成为 TDT 领域中的重要方向。

James Allan[36]是最早使用 NLP 技术解决 TDT 问题的学者之一。其采用 VSM 描述话题和报道,并对模型中的命名实体赋予更高的权重,以此执行 TDT 中的新事件检测(NED)任务。但这种方法并没有获得性能上的提高,主要原因在于其采用的命名实体加权方法是一种经验性的策略,而没有遵循语言学的原理进行估计。对于这种方法的一种改进来自 Nallapati[59],其首先将特征划分到不同的语法类别,比如词性中的名词类和动词类,以及命名实体中的时间类、人名类和地点类。在这个基础上采用语言模型的概率统计方法,估计特征产生于不同语法类别的概率,并以此标记特征的权重。另一类应用于 TDT 中的自然语言处理技术是语义链(Lexical Semantic Chaining,简称为 LSC)。LSC 是基于文本结构的凝聚假设[62]提出的,即构成文本的特征、短语和句子不是孤立存在,而是趋向于围绕一个中心内涵进行组织与论述。LSC 的含义是一组语义上具有继承性的相关特征。通常,来自一篇文本中的语义链不仅能为特征塑造相关的上下文,同时可以更好地描述文本内涵的继承性。最初,Hasan[63]使用 LSC 描述词汇的凝聚性,并基于这种模型评价文本之间的相关程度。Morris 和 Hirst[62]随后设计了基于词汇资源自动构造 LSC 的算法。近期使用 LSC 解决 TDT 问题的研究主要来自 Stokes 和 Hatch[64,65],其结合使用词典信息(WordNet)和文本的上下文信息同时构造 LSC,并基于 LSC 的文本描述形式采用单路径聚类算法解决新事件检测(NED)问题。语义链的使用从语言学的另一种角度解决文本的描述问题,即语义。通常,LSC 有两个优点,一个是语义链具备的上下文信息和词典结构信息可以有效削弱特征的歧义性;另一个优点在于对特征的扩展作用,即使原始文本之间特征的词形迥异,但词典提供的扩展信息仍然可以有效地将其关联在一起。目前,NLP 技术在 TDT 领域的应用已经逐步开展,并在一定程度上弥补了统计学原理在知识理解问题上的不足。但对于该领域的某些研究课题,NLP 技术却无法取代概率统计策略发挥决定性的作用,比如新闻报道的时序性研究。

利用时序特征解决面向新闻报道的检测和跟踪任务也是 TDT 领域重要研究趋势。最早分析时间因素对话题检测影响的研究来自于 CMU 的 Yiming Yang[30]和 UMASS 的 James Allan[31],他们同时提出了一种基于时空顺序的假设,即相对于产生时间较远的报道,产生时间接近的报道论述同一个话题的可能性更大。其中,CMU 采用 SMART[30]系统对报道和话题进行描述,并通过聚类解决话题检测问题。与传统 TD 技术不同之处在于,经过改进的 SMART 系统融合了时间因素对聚类的影响,其聚类相似性是结合基于特征相似度和报道时空距离综合得到的。UMASS 则将时间因素应用于聚类阈值的估计,其中阈值被设计成以时间为参数的函数,阈值可以随时间的变化连续动态地调整,从而适应话题被报道的概率随时间逐渐衰减的趋势。此外,Papka[5,31]改进了 UMASS 的 OTD 算法,同时

① http://www.cs.fudan.edu.cn/mcwil/tdt2005/index.php

将时间因素嵌入话题跟踪任务,其在 TDT2 语料中进一步验证了时空顺序假设对 TDT 的影响。而 Paula Hatch[65] 则融合了 CMU 和 UMASS 的算法,其话题检测系统选择距离当前报道最近并且刚刚参与过更新的 n 个聚类进行比较。当报道与聚类的相关度满足阈值要求时,对该聚类进行更新。同时将当前报道与更新后的聚类质心进行相关度计算,并乘以衰减速度因子,作为该话题新的聚类阈值。总之,时间信息是新闻语料的特色,依靠时间信息追踪话题的发展趋势能够辅助 TDT 相关技术获得更好的效果。因此,未来 TDT 的研究方向中,一方面概率统计和自然语言的融合与相互辅助,对话题理解和报道内容分析将发挥更重要的作用,而另一方面,诸如基于概率统计的报道流时序分析等具备新闻语料特色的课题将成为该领域新的研究热点。

9 总结

综上所述,话题检测与跟踪为自然语言领域的各项研究提供了新的测试平台,由于其面向突发性和延续性很强的新闻语料,因此也对相应技术提出了更高的要求。但是,目前的研究现状仍然以传统基于统计策略的信息检索、信息过滤、分类和聚类等技术为主,忽视了新闻语料本身具备的特点,比如话题的突发性与跳跃性、相关报道的延续与继承性、新闻内容的层次性以及时序性等等。基于这一问题,当前的研究趋势是将多种方法进行融合,并嵌入新闻语料特性实现话题的识别与追踪,比如结合命名实体的话题模型描述、以时间为参数的权重与阈值估计等等。虽然这些方法能够在一定程度上提高 TDT 系统性能,但其只是对传统统计策略的一种补充与修正,并没有形成独立于话题检测与跟踪领域特有的研究框架与模型。因此,TDT 领域未来的研究方向将主要集中于如下几个方面:

(1) 建立具备新闻语料特性的描述模型;

(2) 针对时序性新闻报道的检测与跟踪策略;

(3) 机器学习与自然语言处理技术的有效融合;

(4) 跟踪与检测模型的自适应学习与更新策略;

(5) 新闻语料特有的特征提取与信息挖掘技术。

总而言之,目前的 TDT 研究在国内仍然处于起步阶段,除了非自适应的话题跟踪研究已经达到实用化水平,其他各项任务的系统性能仍然无法满足实际应用的需要。随着信息安全、电子商务以及个性化信息定制等相关应用领域的发展,话题检测与跟踪将成为自然语言和信息处理领域中的重要研究方向。

参考文献:

[1] J Allan, J Carbonell, G Doddington, J Yamron and Y Yang. Topic detection and tracking pilot study: Final report [A]. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop [C], Virginia: Lansdowne, February 1998, 194-218.

[2] James Allan, Ron Papka, Victor Lavrenko. On-line New Event Detection and Tracking [A]. In: the proceedings of SIGIR'98 [C]. University of Massachusetts: Amherst, 1998, 37-45.

[3] J Allan, V Lavrenko, and R Swan. Explorations within topic tracking and detection [A]. In: Topic Detection and Tracking: Event-based Information Organization [C]. Kluwer Academic: Massachusetts, 2002, 197-224.

[4] J M Schultz and M Y Liberman. Towards an universal dictionary for multi-language IR applications [A]. In: Topic Detection and Tracking: Event-based Information Organization [C]. Kluwer Academic: Massachusetts, 2002, 225-241

[5] J Yamron, L Gillick, P van Mulbregt, and S Knecht. Statistical models of topical content [A]. In: Topic Detection and Tracking: Event-based Information Organization [C]. Kluwer Academic: Massachusetts, 2002, 115-134.

[6] Leek T, Schwartz R M., and Sista S. Probabilistic approaches to topic detection and tracking [A]. In: Topic Detection and Tracking: Event-based Information Organization [C]. Kluwer Academic: Massachusetts, 2002, 67-83.

[7] Franck Thollard. Probabilistic DFA Inference Using Kullback-Leibler Divergence and Minimality [A]. In: Proc of the 17th Int'l Conf on Machine Learning [C]. San Francisco: Morgan Kaufmann, 2000, 975-982.

[8] J Ponte and W B Croft. Text segmentation by topic [A]. In: Proceedings of the European Conference on Research and Advanced Technology for Digital Libraries [C]. Europe: ECDL, 1997, pages 113-125.

[9] J Xu and W B Croft. Improving the effectiveness of information retrieval with local context analysis [J]. ACM Transactions on Information Systems (TOIS), 2000, 18(1):79-112.

[10] Y Watanabe, Y Okaxta, K Kaneji, and Y Sakamoto. Multiple Media Database System for TV Newscasts and Newspapers [A]. In: Technical Report of IEIGE [C]. Japan, 1998, 47-54..

[11] C Buckley and G Salton. Optimization of relevance feedback weights [A]. In: Proceedings of SIGIR'95 [C]. Washington, United States: Seattle, 1995, 351-357.

[12] B Masland, G Linoff, and D Waltz. Classifying news stories using memory based reasoning [A]. In: Proceedings of SIGIR'92 [C]. Denmark: Copenhagen, 1992, 59-65.

[13] Y. Zhang, J. G. Carbonell, J. Allan. Topic Detection and Tracking: Detection-Task [A]. In: Proceedings of the Workshop of Topic Detection and Tracking [C], 1997.

[14] J Carbonell, Y Yang, J Lafferty, R D. Brown, T. Pierce, and X. Liu. CMU Report on TDT-2: Segmentation, Detection and Tracking [A]. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop [C]. San Francisco: Morgan Kauffman, 1999, 117-120.

[15] J Kupiec and J Pedersen. A trainable document summarizer [A]. In: Proceedings of the 18th Annual Int'l ACM SIGIR Conf on Research and Development in Information Retrieval (SIGIR'95) [C]. Seattle, Washington, USA: ACM Press, 1995, 68-73.

[16] D D Lewis, R E Schapire, J P Callan, and R Papka. Training Algorithms for Linear Text Classifiers[A]. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Konstanz: Hartung-Gorre Verlag, 1996, 298-306.

[17] R E Schapire. BoosTexter: A Boosting-based System for Text Categorization [J]. Machine Learning, 1999, 39(2-3):135-168.

[18] J M Schultz and Mark Liberman. Topic detection and tracking using idf-weighted cosine coefficient [A]. In: Proceedings of the DARPA Broadcast News Workshop [C]. San Francisco: Morgan Kaufmann, 1999, 189-192.

[19] J P Yamron, I Carp, L Gillick, S Lowe and P V Mulbregt. Topic Tracking in a News Stream [A]. In: Proceedmgs of the DARPA Broadcast News Tracnscription and Understanding Workshop [C], San Francisco: Morgan Kaufmann, 1999.

[20] S A Lowe. The Beta～Binomial Mixture Model and its Application to TDT Tracking and Detection [A]. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop [C], San Francisco: Morgan Kaufmann, 1999.

[21] M Franz, JS Mc Carley. Unsupervised and supervised clustering for topic tracking[A]. In: Proceedings of the 24th annual international ACM SIGIR[C]. New Orleans, Louisiana, USA: ACM, 2001, 310 317.

[22] Nianli Ma, Yiming Yang, Monica Rogati. Applying CLIR Techniques to Event Tracking [A]. In: AIRS 2004[C]. Berlin Heidelberg: Springer-Verlag, 2005, 24 35.

[23] L S Larkey, F F Feng, M Connell, V Lavrenko. Language-specific Models in Multilingual Topic Tracking [A]. In: Proceedings of the 27th annual international conference on research and development in information retrieval [C]. Sheffield, UK, 2004, 402-409.

[24] T Strzalkowski, G C Stein and G B Wise. GE. Tracker: A Robust, Lightweight Topic Tracking System [A]. In: Proceedings of the DARPA Broadcast News Workshop [C]. San Francisco: Morgan Kaufmann, 1999,

[25] J P Yamron, S Knecht, and P V Mulbregt. Dragon's Tracking and Detection Systems for the TDT2000 Evaluation [A]. In: Topic Detection and Tracking Workshop [C]. USA: National Institute of Standard and Technology, 2000, 75 79.

[26] J Allan, V Lavrenko, D Frey, V Khandelwal. UMass at TDT 2000 [A]. In: Proceedings of Topic Detection and Tracking Workshop [C]. USA: National Institute of Standar and Technology, 2000, 109-115.

[27] N Lester, H E Williams. TDT2001 Topic Tracking at RMIT University[A]. In: The Topic Detection and Tracking (TDT) Workshop [C], 2001.

[28] W Lam, S Mukhopadhyay, J Mostafa, and M Palakal. Detection of Shifts in User Interests for Personalized Information Filtering [A]. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. Konstanz: Hartung-Gorre Verlag, 1996, 317-325.

[29] Y Lo, J L Gauvain. The LIMSI Topic Tracking System For TDT 2002 [A]. In: Topic Detection and Tracking Workshop [C]. Gaithersburg, USA, 2002.

[30] Y Yang, T Pierce, J Carbonell. A study on Retrospective and On-Line Event detection [A]. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval [C]. 1998, CMU, USA: ACM, 28-36.

[31] Ron Papka. On-line New Event Detection, Clustering and Tracking [D]. Amherst: Department of Computer Science, UMASS,1999.

[32] Allan J, Papka R, Lavrenko V. On-Line New Event Detection and Tracking [A]. In: Proceedings of SIGIR'98:21st Annual International ACM SIGIR Conference on Research and Development in Information

Retrieval [C]. New York: ACM Press, 1998, 37-45.

[33] Y Yang, T Pierce, J Carbonell. A study on Retrospective and On-Line Event detection [A]. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval [C]. 1998, CMU, USA: ACM, 28-36.

[34] T Brants, F Chen, and A Farahat. A system for new event detection [A]. In: Proceedings of the 26th SIGIR conference on Research and development in information retrieval [C], 2003.

[35] G Kumaran and J Allan. Text classification and named entities for new event detection [A]. In: Proceedings of the SIGIR Conference on Research and Development in Information Retrieval [C]. Sheffield, South Yorkshire: ACM, 2004, 297-304.

[36] J. Allan, H Jin, M Rajman, C Wayne, G D, L V, R Hoberman, and D Caputo. Topic-based novelty detection [A]. In: Proceedings of the Johns Hopkins Summer Workshop [C]. CLSP, Baltimore, 1999.

[37] Y Yang, J Carbonell, C Jin. Topic-conditioned novelty detection[A]. In: Hand D, etal. Proceedings ofthe 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [C]. New York: ACM Press, 2002, 688-693.

[38] W Lam, H Meng, K Wong, and J Yen. Using contextual analysis for news event detection [J]. International Journal on Intelligent Systems, 2001, 16(4): 525-546.

[39] Z Li, B Wang, M J Li, W Y Ma. A Probabilistic Model for Retrospective News Event Detection [A]. In: Proceedings of the 28th annual international ACM SIGIR[C]. Salvador, Brazil: ACM, 2005, 106-113.

[40] The 2004 Topic Detection and Tracking (TDT2004) Task Definition and Evaluation Plan [H]. version 1.2, http://www.nist.gov.

[41] D R Cutting, D R Karger, J O Pedersen, and J W Tukey. Scatter/gather: a cluster-based approach to browsing large document collections [A]. In: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval [C]. NY:ACM, 1992, 318 329.

[42] D Trieschnigg and W Kraaij. TNO hierarchical topic detection report at TDT 2004[A]. In: The 7th Topic Detection and Tracking Conf[C]. 2004.

[43] Allan J, Bolivar A, Connell M, Cronen-Townsend S, Feng A, Feng F, Kumaran G, Larkey L, Lavrenko V, Raghavan H. UMass TDT 2003 Research Summary[A]. In: Proceedings of TDT 2003 evaluation, unpublished[C], 2003.

[44] Levow G A and Oard D W. Signal boosting for translingual topic tracking: Document expansion and n-best translation [A]. In: Topic detection and tracking: Event-based information organization [C]. MA: Kluwer, 2002, 175-195.

[45] Jin H, Schwartz R, Sista S and Walls F. Topic Tracking for Radio, TV Broadcast and Newswire [A]. In: Proceedings of the DARPA Broadcast News Workshop[C]. San Francisco: Morgan Kaufmann, 1999, 199-204.

[46] Tim Leek, Hubert Jin, Sreenivasa Sista, Richard Schwartz. The BBN Crosslingual Topic Detection and Tracking System[A]. In: Working Notes of the Third Topic Detection and Tracking Workshop[C]. 2000.

[47] 骆卫华,刘群,程学旗. 话题检测与跟踪技术的发展与研究[A]. 全国计算语言学联合学术会议(JSCL-2003)论文集[C]. 北京:清华大学出版社,2003,560-566.

[48] 李保利,俞士汶. 话题识别与跟踪研究[J]. 计算机工程与应用,2003,39(17):6-10.

[49] 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.

[50] 赵华,赵铁军,张姝,王浩畅. 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.

[51] Zhang Kuo, Li Juan Zi, Wu Gang. New Event Detection Based on Indexing-tree and Named Entity[A]. In: Sigir2007[C]. ACM: Amsterdam, 2007.

[52] 宋丹,卫东,陈英. 基于改进向量空间模型的话题识别跟踪[J]. 计算机技术与发展,2006,9(16):62-67.

[53] 于满泉,骆卫华,许洪波,白硕. 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.

[54] 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[J]. 中文信息学报,2006,20(1):29-36.

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

[56] 金珠,林鸿飞,赵晶. 基于 HowNet 的话题跟踪及倾向性分类研究[J]. 情报学报,2005,5(24):555-561.

[57] Ponte, J M and Croft, W B. A Language Modeling Approach to Information Retrieval [A]. In: ACM SIGIR [C]. NY: ACM, 1998, 275-281.

[58] V Lavrenko, J Allan, E DeGuzman, D LaFlamme. Models for Topic Detection and Tracking [A]. In: Proceedings of HLT-2002 [C], 2002, 104-110.

[59] R Nallapati. Semantic Language Models for Topic Detection and Tracking [A]. In: Proceedings of HLT-NAACL2003 Student Research Workshop [C]. 2003, 1-6.

[60] V Lavrenko and W B Croft. Relevance-based lan-

guage models[A]. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval[C]. New Orleans, Louisiana, USA: ACM, 2001, 267-275.

[61] W B Croft, S Cronen-Townsend, and V Lavrenko. Relevance feedback and personalization: A language modeling perspective [A]. In: Proceedings of the DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries [C]. 2001, 49-54.

[62] Jane Morris, Graeme Hirst. Lexical Cohesion by Thesaural Relations as an Indicator of the Structure of Text [J], Computational Linguistics, 1991, 17(1): 21-48.

[63] HASAN R. Coherence and cohesive harmony [A]. In: Flood L, eds. Understanding Reading Comprehension [C]. Newark, Delaware: International Reading Association, 1984, 181-219.

[64] Nicola Stokes, Paula Hatch, Joe Carthy. Lexical Semantic Relatedness and Online New Event Detection [A]. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval [C]. Greece: ACM, 2000, 324-325.

[65] Hatch P, Stokes N, Carthy J. Topic detection, a new application for lexical chaining? [A]. In: British Computer Society IRSG 2000 [C]. Cambridge: British Computer Society, 2000, 94-103.

中科院软件所筹建国内首家软件博物馆

近日,中国科学院软件研究所发起建设我国首家以计算机软件为主题的软件博物馆。

软件博物馆旨在记录软件发展历程,展示软件发展成就,传播软件科技知识,宣传软件科学文化。届时软件博物馆将以丰富翔实的史料和珍贵的实物,将计算机软件从起步到现在的发展状况以及未来发展趋势生动、直观地展示给大众。通过各种展示手段,追溯软件的发展历程,发掘软件文化内涵,弘扬科学精神,普及科技知识。

软件博物馆计划于 2008 年中期向公众开放。目前,正面向社会各界广泛征集能反映国内外软件发展历程和软件发展成就的实物、照片、回忆文章、模型、成果展示材料等。有捐赠意向的单位及个人请与软件博物馆建设办公室联系。

地址: 北京中关村南四街 4 号中科院软件园区 5 号楼 202 室

邮编: 100080

电话: 86-10-62661035

传真: 86-10-62661035

Email: rjbwg @iscas.ac.cn

联系人: 李洁