话题检测与跟踪技术的发展与研究

骆卫华　刘群　程学旗

中国科学院计算技术研究所 软件研究室 北京 100080
{luoweihua, liuqun,cxq}@ict.ac.cn

摘　要：本文介绍了话题检测与跟踪技术的由来和发展历程，并展望其应用前景，同时比较系统地介绍了现有的话题检测与跟踪系统主要采用的方法，并对其效果进行了比较。

关键词：话题检测与跟踪，向量空间模型，语言模型

Development and Analysis of Technology of Topic Detection and Tracking

Luo Weihua　Liu Qun　Cheng Xueqi

Software Division, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080
{luoweihua,liuqun,cxq}@ict.ac.cn

Abstract: The paper introduces the origin and history of the development of technology of topic detection and tracking, and makes remarks on its prospect. It also describes systemically the methods adopted by the current systems of topic detection and tracking, and makes comparison among their performance.

Keywords: Topic Detection and Tracking, Vector Space Model, Language Model

1　应用背景

随着信息传播手段的进步，尤其是互联网这一新媒体的出现，我们已经摆脱了信息贫乏的桎梏。在目前信息爆炸的情况下，如何快捷准确地获取感兴趣的信息成为人们关注的主要问题。目前的各种信息检索、过滤、提取技术都是围绕这个目的展开的。由于网络信息数量太大，与一个话题相关的信息往往孤立地分散在很多不同的地方并且出现在不同的时间，仅仅通过这些孤立的信息，人们对某些事件难以做到全面的把握。而基于关键词的检索工具返回的信息冗余度过高，很多不相关的信息仅仅因为含有指定的关键词就被作为结果返回了，因此人们迫切地希望拥有一种工具，能够自动把相关话题的信息汇总供人查阅。话题检测与跟踪（Topic Detection and Tracking，以下简称 TDT）技术就是在这种情况下应运而生的，它可以帮助人们把分散的信息有效地汇集并组织起来，从整体上了解一个事件的全部细节以及与该事件与其它事件之间的关系。

TDT 技术可以用来监控各种语言信息源，在新话题出现时发出警告，在信息安全、金融证券、行业调研等领域都有广阔的应用前景。此外，它还可以用来跟踪某个话题的来龙去脉，进行历史性质的研究。

2 发展历程

TDT 的概念最早产生于 1996 年，当时美国国防高级研究计划署（DARPA）根据自己的需求，提出要开发一种新技术，能在没有人工干预的情况下自动判断新闻数据流的主题。1997 年，研究者开始对这项技术进行初步研究，并做了一些基础工作（包括建立一个针对 TDT 研究的预研语料库）。当时的研究内容包括寻找内在主题一致的片断，而且能自动判断新事件的出现以及旧事件的再现。从 1998 年开始，在 DARPA 支持下，美国国家标准技术研究所（NIST）每年都要举办话题检测与跟踪国际会议，并进行相应的系统评测。这个系列评测会议作为 DARPA 支持的 TIDES 项目下的两个系列会议之一，越来越受到人们的重视。参加该评测的机构包括著名的大学、公司和研究所，如 IBM Watson 研究中心、BBN 公司、卡耐基－梅隆大学、马萨诸塞大学、宾州大学等。国内这方面的研究开展得要晚一些，1999 年国立台湾大学参加了 TDT 话题检测任务的评测，香港中文大学参加了 TDT2000 某些子任务的评测。最近北京大学和中科院计算所的研究人员也开始进行这方面的跟踪和研究。总的来看，TDT 系列评测会议呈现两大趋势：一是努力提高信息来源的广泛性，其来源包括互联网上的文本数据，也包括来自广播、电视的语音数据；二是强调多语言的特性。从 1999 年开始，TDT 会议先后引入了汉语和阿拉伯语的测试集。

TDT 会议采用的语料是由语言数据联盟（Linguistic Data Consortium，简称 LDC）提供的 TDT 系列语料，目前已公开的训练和测试语料包括 TDT Pilot Corpus、TDT2 和 TDT3，这些语料都人工标注了若干话题作为标准答案。TDT2 和 TDT3 收录的报道总量多达 11 万 6 千篇，从而很大程度上降低了数据稀疏问题的影响，同时能较好地验证算法的有效性。

可以看到，话题检测与跟踪和信息抽取研究一样，其建立与发展是以评测驱动的方式进行的。这种评测研究的方法具有以下一些特点：明确的形式化的研究任务、公开的训练与测试数据、公开的评测比较。它将研究置于公共的研究平台上，使得研究之间的比较更加客观，从而让研究者认清各种技术的优劣，起到正确引导研究发展方向的目的。

3 研究内容与现状

与一般的信息检索或者过滤不同，TDT 所关心的话题不是一个大的领域（如美国的对华政策）或者某一类事件（如恐怖活动），而是一个很具体的“事件（Event）”，如 911 事件、江泽民访美等等。为了区别于语言学上的概念，TDT 评测会议对“话题”进行了定义：所谓话题（Topic），就是一个核心事件或活动以及与之直接相关的事件或活动。而一个事件（Event）通常由某些原因、条件引起，发生在特定时间、地点，涉及某些对象（人或物），并可能伴随某些必然结果，可以简单地认为话题就是若干对某事件相关报道的集合*。“话题检测与跟踪”则定义为“在新闻专线和广播新闻等来源的数据流中自动发现主题并把主题相关的内容联系

*显然，对这种相关性必须做一个界定，不能任由集合无限扩大。为此，LDC 在构造 TDT 语料时，对挑选出来的每个话题都定义了相关性判定规则。

在一起的技术”。例如，“俄克拉荷马城爆炸案”这个主题包括1995年美国联邦大楼被炸、悼念仪式、政府的一系列调查、对McVeigh的指控等等。这个定义和其它与话题有关的研究不同，那些研究主要处理信息分类问题，比如任何与爆炸有关的事件。处理分类问题需要专门的分类体系，注解起来效率低且主观色彩浓厚。

TDT是一项综合的技术，需要比较多的自然语言处理理论和技术作为支撑，因此这些测评对其进行了细化。根据不同的应用需求，TDT评测会议把话题检测和跟踪分成五个子任务。

任务	定义
报道切分（Story Segmentation）	找出所有的报道*边界，把输入的源数据流分割成各个独立的报道。
话题跟踪（Story Tracking）	给出某话题的一则或多则报道，把后输入进来的相关报道和该话题联系起来。它实际上包括两步，首先给出一组样本报道，训练得到话题模型，然后在后续报道中找出所有讨论目标话题的报道。
话题检测（Story Detection）	发现以前未知的新话题。
首次报道检测（First Story Detection，TDT2002改称New Event Detection）	在数据流中检测或发现首次，并且只能是首次讨论某个话题的报道。与话题检测本质相同，区别只在于结果输出的形式不同。
关联检测（Link Detection）	判断两则报道是否讨论的是同一个话题。

表1 TDT的技术任务

TDT会议对参评的TDT系统定下的目标是“实现一个功能强大、用途广泛的全自动算法用以判断自然语言数据的主题结构，同时要做到与来源、媒介、领域和语言无关”。目前的成果表明切分定界的性能已经和人工相差无几，话题跟踪技术也已基本实用，但话题检测技术还有待改进。尤其值得一提的是，单一语言的测试性能并不随语种的变化而发生很大变化，对汉语话题的跟踪和检测性能与英语十分接近。为了对不同的系统进行量化比较，TDT会议制订了一套评测规范。每一个参评系统的性能是由误报率和漏报率加权求和的结果进行衡量的，其计算公式是：

$$C_{Det} = C_{Miss} \cdot P_{Miss} \cdot P_{target} + C_{FA} \cdot P_{FA} \cdot P_{non-target}$$

其中C_{Det}是系统的性能评测指标，称为检测错误开销。C_{Miss}和C_{FA}分别是漏查和误报的开销；P_{Miss}和P_{FA}分别是漏查和误报的条件概率；P_{target}是目标话题的先验概率，$P_{non\text{-}target}=1-P_{target}$。$C_{Miss}$、$C_{FA}$和$P_{target}$都是预设值，作为调节漏报率和误报率在评测结果中所占比重的系数。检测开销通常被归一化为0和1之间的一个值：

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

一般直接用$(C_{Det})_{Norm}$作为评价系统性能的分数。

* 在TDT的评测中，“报道”定义为“论述某个话题的新闻片断，它包括两个以上独立表述该事件的说明语句”。

4 主要实现方法

构造一个实用化的 TDT 系统是进行 TDT 研究的主要目的之一，也是检验现有方法优劣的基础。从参评的数量来看，话题发现和话题跟踪两个子任务最受关注。因此我们介绍的实现方法也以这两个任务为主。总体而言，要实现话题发现与跟踪功能，需要解决以下问题：（1）话题/报道的模型化；（2）话题一报道相似度的计算；（3）聚类策略；（4）分类策略。

整个系统的流程大致是（以话题跟踪为例）：

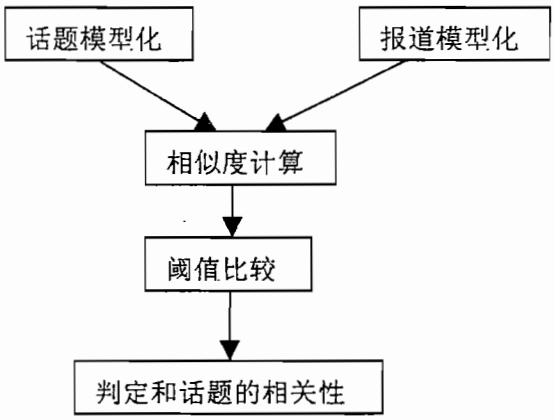

图 1 话题跟踪系统流程

针对以上问题，我们将逐一介绍一些已经被广泛采用并得到实际评测验证的方法。

4.1 话题/报道模型

要判断某个报道是否和话题相关，首先需要解决用什么模型来表示它们的问题。目前常用的模型有语言模型（Language Model，LM）和向量空间模型（Vector Space Model，VSM）。

（1）语言模型

语言模型是一种概率模型。假设报道中出现的词 δ_n 各不相关，则某则报道 S 和话题 C 相关的概率：

$$P(C|S) = \frac{P(C)\cdot P(S|C)}{P(S)} \approx P(C)\prod_n \frac{P(\delta_n|C)}{P(\delta_n)}$$

其中 p(C)是任何一则新报道和话题 C 相关的先验概率，$p(\delta_n|C)$是表示词 δ_n 在某话题 C 中的生成概率。$p(\delta_n|C)$可以表示成一个两态的混合模型，如图 2 所示：

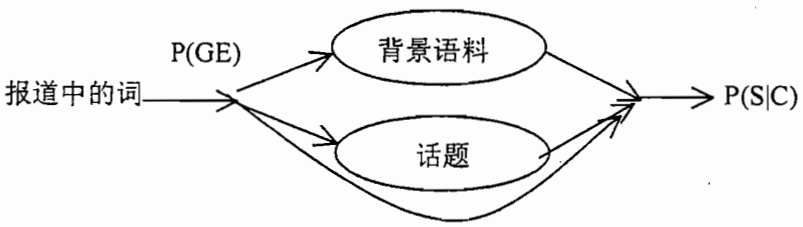

图 2 $p(\delta_n|C)$的两态模型

其中一个状态是词在该话题中所有报道的分布，另一个状态是词在整个语料中的分布。计算此模型中的两个状态采用的是最大似然估计算法。因为话题语言模型很稀疏，这里必须解决未见词的 0 概率问题，通常采用线性插值法把背景语言模型加入进去：

$p'(\delta_n|C) = \alpha \cdot p(\delta_n|C)+(1-\alpha) \cdot p(\delta_n)$

（2）向量空间模型

向量空间模型是目前最简便高效的文本表示模型之一。其基本思想是：给定一自然语言文档 $D=D(t_1, w_1; t_2, w_2; \cdots; t_N, w_N)$，其中 t_i 是从文档 D 中选出的特征项，w_i 是项的权重，$1 \leqslant i \leqslant N$，因而 $D(w_1, w_2, \cdots, w_N)$ 被看成是 N 维空间中的一个向量，而两个文档 D_1 和 D_2 之间的（内容）相关程度常常用它们之间的相似度 $Sim(D_1, D_2)$ 来度量。在实际的参评系统中，多数都以词作为文本特征项。特征（词）加权采用的是 tf*idf 加权策略。某些系统把词分成命名实体和内容词两类，视其对文档表达重要度的不同赋予不同权重。

（3）中心向量模型

中心向量模型实际是向量空间模型的一种变形。每个话题用一个中心向量表示，所谓中心向量就是在此类中所有报道的向量表示的平均值。输入的报道和每个话题的中心向量相比较，选择最相似的那个话题。

无论选择哪种模型，一般都需要进行初始化，即消去禁用词，对于英语而言，还需要做词根还原的工作。

4.2 相似度计算

对所有的话题 C_1、C_2、……C_n，要判断某一则报道 S 属于哪一个话题，就需要计算报道和各个话题之间的相似程度，最后把最高相似度和阈值进行比较，对于语言模型而言，就是寻找 k 满足：$k = \arg\max_i P(C_i \mid S)$

由前面的语言模型并取 log 值，相似度计算公式就表示为

$$D(S, C) = \log \prod_m \frac{P(\delta_m \mid C)}{P(\delta_m)}$$

通常用语言模型算出的话题与话题之间的相似度不可比较，因为单个语言模型都有各自不同的概率特征，比如，有的话题所用的词很特殊，像“霍根班德在 200 米自由泳中击败索普”，而有的话题用词就很普通，像“克林顿总统访问中国”。这样测试文档和不同话题之间算出的分数差异很大，不能用单一的阈值进行比较，必须进行归一化。考虑到用上面的 D(S,C) 算出的值基本上是一组独立的随机离散变量值，如果值足够多的话，由中心极限理论，其分布近似为高斯分布，假设τ为原来的概率，μ为所有报道对某话题概率的平均值，σ是这些概率的标准方差，则新的分值可以归一化为 $\tau' = (\tau-\mu)/\sigma$ 。

向量空间模型和中心向量模型通常采用 cosine 公式来计算报道－话题的相似度，即求两者的内积，相似度计算公式表示为

$$D(S,C) = \frac{\sum q_i d_i}{\sqrt{\left(\sum q_i^2\right)\left(\sum d_i^2\right)}}$$

其中 q_i、d_i 分别是报道和话题中的特征项的权值。cosine 相似度在比较两个长文档时比较有效，此时如果两个文档的向量维数不进行任何压缩，系统性能最佳。因为本身已进行了归一化，所以 cosine 相似度不依赖于特定的特征加权方法。

近来有些系统开始尝试用 OKAPI 公式来计算报道-话题相似度，其形式是：

$$Ok(d^1,d^2;cl) = \sum_{w\in d^1\cap d^2} t_w^1 t_w^2 (idf(w) + 2\lambda \frac{n_{w,d}}{n_w + n_{cl}})$$

所得结果表示文档和文档之间的距离，其中 d^1,d^2 是两个文档，cl 是 d^1,d^2 中较早出现的那个文档所属的话题。t_w^i 是词 w 在文档 i 中调整后的词频，对其进行归一化处理使得 $\sum_w t_w^i = 1$ 独立于 d^i 的长度，idf(w)是词 w 的倒文档频率，n_w 是包含词 w 的文档数目，n_{cl} 是话题 cl 中文档的数目，$n_{w,cl}$ 是话题 cl 中包含词 w 的文档的数目，λ 是控制词的权值中和话题相关的那部分“动态权值”的可调参数。文档和话题之间的分数是一个平均值：

$$Ok(d,cl) = |cl|^{-1} \sum_{d^f \in cl} Ok(d,d^f;cl)$$

在做跟踪训练时，把所有的训练报道分成一个或多个话题，然后对每一则测试报道计算它跟某个话题之间的分数。根据分数做两个阈值判断。如果分数超过高阈值 Θ_m，则把该报道并入话题（因而通过 n_{cl} 影响了将来的分数）。如果分数超过了低阈值 Θ_d，则表示此报道与话题相关，但不把它并入聚类。

4.3 聚类分类策略

判断某个新报道是属于已有话题还是一个新话题，往往是同时进行的。通常的做法是把新报道和已有话题进行比较，如果相似度高于某个阈值，则把新报道归入相似度最高的话题中，如果对所有话题的相似度都低于阈值，则创建一个新话题。但在具体实现中，还牵涉到选用哪些聚类、分类和根据反馈进行参数调整的策略。

最简单的方法称为增量聚类算法，它顺序处理报道，一次处理一则，对每一则报道它执行两个步骤：（1）选择：选出和报道最相似的聚类；（2）比较阈值：把报道和阈值相比较，决定是把报道分到聚类里还是创建一个新的聚类。这种算法非常直观，易于实现，但它的缺点也很明显：①对一则报道只能做一次决策，因此早期根据很少的信息所做的错误判断累计到后面可能相当可观；②随着报道的不断处理，计算开销会越来越大。

针对这些缺点稍加改进，就形成了增量 k-means 方法，它在当前报道窗口中进行迭代操作，每一次迭代都要做适当的改变。具体步骤是：

1. 使用增量聚类算法处理当前窗口中的全部报道。
2. 把窗口中的每一则报道和旧的聚类进行比较，判断每则报道是要合并到聚类中去还是用作新聚类的种子。
3. 根据计算结果立刻更新所有的聚类。
4. 重复步骤（2）—（3），直到所有的聚类不再变化
5. 查看下一批报道，转向（1）。

此外，常用的文本分类算法 KNN 算法，应用在话题跟踪上也有比较好的效果。

对于参数调整，各个系统也采用不同的策略。有些只根据正例（和话题相关）对话题模型进行调整，而有些则兼顾正例和反例。对以向量空间表示的话题而言，Rocchio 方法是一种较为有效的参数调整方法，其形式为：

$$\omega'_{jc} = \alpha\omega_{jc} + \beta \frac{\sum_{i\in C} x_{ij}}{n_C} - \gamma \frac{\sum_{i\notin C} x_{ij}}{n - n_C}$$

其中 ω'_{jc} 是调整之后的权值，ω_{jc} 是原来的权值，i 表示已处理的报道，C 表示某个话题，

x_{ij} 是 i 中的特征项，n 是已处理报道的总数，n_C 是正例的总数。

总的来看，目前对话题本身特征的研究还不深入，仅仅借用分类、过滤的一些方法还不能有效地解决话题发现、新事件检测等任务。此外，有些研究机构也在尝试新的算法，比如支持向量机（Support Vector Machine）、最大熵（Maximum Entropy）、文档扩展等，但都还需要在评测中实际验证其效果。

5 结束语

目前来看，TDT 的研究呈现以下特点：

（1） 多数已公开系统采用的方法主要还是传统的文本分类、信息过滤和检索的方法，专门针对话题发现与跟踪自身特点的算法还未形成；

（2） 要取得整体上比较满意的效果并不太困难，但对某个用户感兴趣的特定话题，现有系统都无法保证取得满意的效果，比如对于用户当前最为关注的“伊拉克战争”，系统不能保证取得高于平均值的准确率；

（3） 综合使用多种相对成熟的方法，从长期来看在实际应用中可能效果最佳，同时这也是将来的一个研究发展方向。

总之，TDT 是自然语言处理领域中的一个新兴的研究课题。通过评测驱动的方式，TDT 的研究已经取得了相当大的进展。但当前的研究主要还是基于传统的统计方法，这些方法在文本分类、信息检索、信息过滤等领域得到广泛的应用。将来的发展应主要关注话题本身的特性，并考虑多种方法的综合运用。TDT 的发展和实际应用息息相关，在国家信息安全、企业市场调查、个人信息定制等方面都存在着实际需求。随着现有系统性能的不断提高，TDT 在各个领域必将得到越来越广泛地应用。

参 考 文 献

[1]S.A. Lowe. “The Beta-Binomial Mixture Model and Its Application to TDT Tracking and Detection,” *Proceedings of the DARPA Broadcast News Workshop,* February 1999.

[2]W. Lam and H. Meng and K. Hui. “Multilingual Topic Detection Using a Parallel Corpus”. In *Proceedings of the DARPA TDT 2000 Workshop*, November 2000.

[3]Jin. H., R. Schwartz, S. Sista and F. Walls, “Topic Tracking for Radio, TV Broadcast and Newswire,” *Proceedings of the DARPA Broadcast News Workshop*, Herndon, Va, 1999.

[4]Schwartz, R., Imai, T., Nguyen, L., and Makhoul, J., “A maximum Likelihood Model for Topic Calssification of Broadcast News,” in *Proc. Eurospeech*, Rhodes, Greece, September,1997.

[5]Walls, F., Jin, H., Sista, S., and Schwartz, R., “Topic Detection in Broadcast News,” in *Proceedings of the DARPA Broadcast News Workshop*, Herndon, Va, 1999.

[6]J.P. Yamron, I. Carp, L. Gillick, S.A. Lowe, and P. van Mulbregt,“Topic Tracking in a News Stream”, *Proceedings of the DARPA Broadcast News Workshop,* February 1999.