第35卷 第18期 计 算 机 工 程 2009年9月
Vol.35 No.18 Computer Engineering September 2009
·博士论文· 文章编号：1000—3428(2009)18—0028—03 文献标识码：A 中图分类号：TP311

面向互联网新闻的在线话题检测算法

程 葳，龙志祎
(北京城市学院人工智能研究所，北京 100083)

摘 要：针对互联网新闻报道冗余多、议题发散、易漂移等特点，提出一种面向互联网的在线话题检测算法。该算法针对冗余问题提出子话题概念，针对议题发散问题建立双层检测结构，针对话题漂移问题提出基于滑动窗口的跟踪策略。应用该算法建立网上话题检测系统，通过来源于互联网的真实数据进行测试。结果表明，算法性能优于传统的单路径聚类算法，其最小错误代价率低于0.14。

关键词：在线话题检测；话题检测与跟踪；文本聚类

Online Topic Detection Algorithm for Internet News

CHENG Wei, LONG Zhi-yi
(Institute of Artificial Intelligence, Beijing City University, Beijing 100083)

【Abstract】 This paper analyses the Internet news reports and finds their characteristics such as redundancy, low centralization of the discussions and the topic drift. An Online Topic Detection(ODT) method for Internet is proposed. It defines the sub-topic to ignore the redundancies reports, presents the double-lays configuration for the low centralization of the discussions, and advances a topic tracking algorithm based on the sliding window. A topic detection system is build according to the method. The system is tested by the real data from the Internet. The results present that this method is better than the single-pass method for ODT. The C_{Det} of the method is under 0.14.

【Key words】 Online Topic Detection(ODT); Topic Detection and Tracking(TDT); text clustering

1 概述

互联网因即时性强、互动性好成为人们日常获取新闻的重要途径之一，并受到政府、金融、企业、情报等各领域关注。由于网络中的重要信息常会被海量数据淹没，因此建立以话题为主线的信息组织模式、快速有效地检测出网上新话题和热点话题成为该研究领域的新热点。

话题检测与跟踪(Topic Detection and Tracking, TDT)是近年来兴起的一项针对新闻报导进行信息识别、语言挖掘和组织的新技术。在线话题检测(Online Topic Detection, OTD)是其中的一个重要研究课题，其特点在于系统必须在对所有话题毫不了解的情况下构造话题检测模型，并根据该模型检测陆续到达的报道流，从中识别出最新话题；同时收集已识别话题的后续相关报道[1]。

目前，OTD系统最常用的方法是单路径聚类算法[2]。不过，该算法对互联网新闻报道并不能取得预期效果[3]，因为互联网新闻具有以下特点：(1)冗余多。网络中存在大量内容相似的报道。(2)议题发散。网络新闻在语言上较为灵活，议题的发散性也更强。一些话题不仅可以持续几个月，而且涉及诸多领域。(3)易漂移。热门话题会引发大量相关报道，而话题在不同历史阶段的论述中心也将有所漂移[4]。

本文根据以上分析提出一种改进的在线话题检测算法，并通过实际互联网数据对其进行测试。实验表明，该方法对互联网新闻的在线话题具有较好的检测效果。

2 主要核心技术

2.1 子话题

传统OTD算法只考虑话题和报道2个层次。不过，互联网新闻中存在大量的冗余信息，这些信息不仅浪费人们的阅读时间，而且占用系统资源、影响话题检测效率。因此，本文在原有话题层和报道层之间增加了子话题层，见图1。

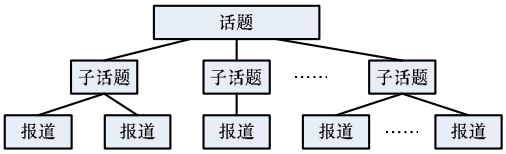

图1 话题、子话题和报道之间的关系

具体定义如下：

定义1 关于同一事件或活动的相似报道集合称为子话题。一个子话题可以包含一篇或多篇报道；但唯一从属于一个话题。一个话题可以包含一个或多个子话题。

可以看出，子话题反映的是话题内部的信息冗余现象。同时，由于子话题中报道包含的信息基本相似，因此可只选择其中一篇作为代表，忽略其余报道。这样就能在几乎不损失信息的情况下有效压缩数据、提高处理效率。为表述方便，进行如下定义：

定义2 被选择用于代表一个子话题的报道称为有效报道，其他报道称为冗余报道。

2.2 双层检测结构

据分析，互联网新闻的议题发散性主要表现在子话题内部的相对集中和从属于一个话题的子话题间的相对发散。因

基金项目：北京市教育委员会科技发展计划面上基金资助项目(KM200600006002)
作者简介：程 葳(1973－)，女，副教授、博士，主研方向：互联网内容分析，自然语言处理；龙志祎，讲师、硕士
收稿日期：2009-02-24 **E-mail：**wcheng@bcu.edu.cn1

此，本文建立了基于双层检测结构的互联网新闻检测系统，见图 2。该系统针对子话题层和话题层的不同特点，分别采用不同的特征抽取或相似度计算公式，以兼顾系统的漏检率与误检率。

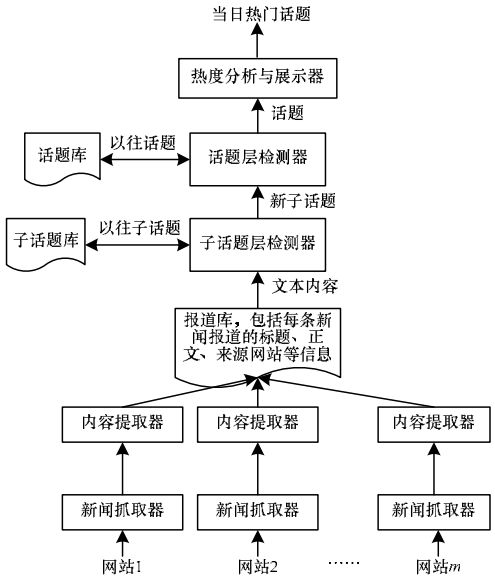

图 2　互联网新闻检测系统框图

2.3　基于滑动窗口的跟踪模型

跟踪过程主要是通过新数据与以往数据的比较，发现已有话题(或子话题)的后续报道。其最大困难在于互联网新闻的数据量大和易漂移。因此，本文建立了基于滑动窗口式的跟踪模型。如图 3 所示，以当前时间为基准，沿时间轴向前开设固定时长的窗口，仅跟踪窗口内所包含的话题(或子话题)，而对窗口外信息(如 a, b)不再进行处理。

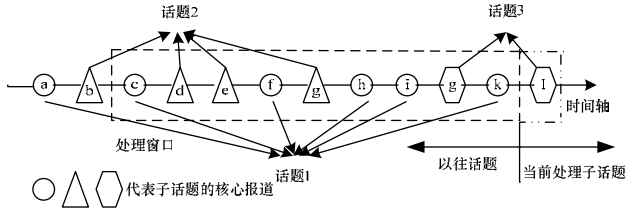

图 3　基于滑动窗口的跟踪模型

该模型具有如下特点：

(1)将有限的资源用于处理最需要处理的部分。模型只对新近增添了报道的活跃话题(或子话题)进行跟踪，有效避免了大量已被网民遗忘的话题(或子话题)浪费系统处理时间。

(2)同时兼顾大话题和小话题。模型中只有当一个话题(或子话题)在一定时段内没有出现新报道才被认为是自动关闭，从而可以有效地兼顾持续时间较长的大话题和持续时间较短的小话题。

(3)体现话题的漂移。对于话题层，模型应用构成一个话题的多个子话题共同描述该话题，并随着时间推移自动调整描述话题的子话题。一些长期不添加新报道的子话题将移出处理窗口，新近添加的子话题获得更大权重。这样，模型可以自动跟踪话题的漂移，从而提高话题层检测的整体性能。

3　算法实现

3.1　子话题层检测

图 4 给出子话题层的在线检测流程，共分为 3 个阶段：子话题发现，有效报道获取和子话题跟踪。其中，子话题跟踪采用 2.3 节所述的滑动窗口模型，$\theta 1$ 为子话题发现阈值。

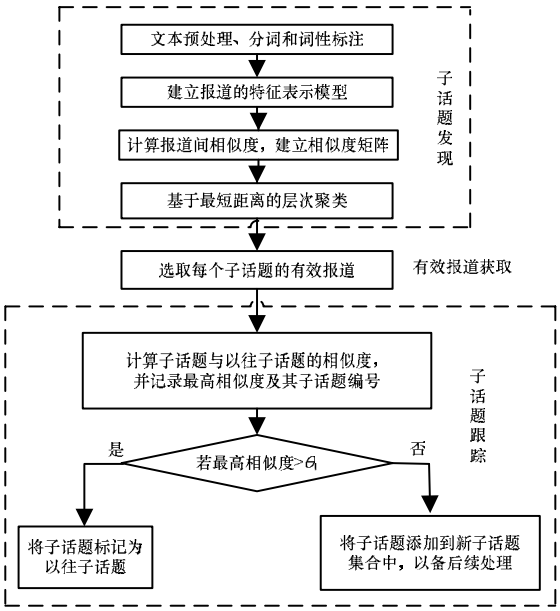

图 4　子话题层检测流程

(1)子话题发现

子话题发现采用自底向上的层次聚类算法。其中，特征权重采用标准的 TFIDF 算法；特征选择按照权重由大到小提取前 200 个作为报道的特征词；报道间相似度采用式(1)的余弦距离；子话题间相似度采用式(2)的最短距离。

$$sim(\bar{x},\bar{y})=\frac{\sum_{i=1}^{m} wx_i wy_i}{\sqrt{\sum_{k=1}^{m} wx_k^2}\bullet\sqrt{\sum_{k=1}^{m} wy_k^2}} \tag{1}$$

其中，$sim(\bar{x},\bar{y})$ 是报道 $\bar{x}$ 和 $\bar{y}$ 间的相似度；wx_i 是特征词 i 在报道 $\bar{x}$ 中的权重；wy_i 是特征词 i 在报道 $\bar{y}$ 中的权重；m 是单词总数。

$$sim(c_j,c_l)=\max_{d\in c_j,\tilde{d}\in c_l}\{sim(d,\tilde{d})\} \tag{2}$$

其中，d 和 $\tilde{d}$ 分别表示子话题 c_j 和 c_l 所包含的报道。

(2)有效报道获取

根据式(3)可抽取类内平均相似度最大的报道作为有效报道。

$$d=\max_{d_i\in c_k}\left\{\sum_{\substack{j=1\\ j\neq i\\ d_j\in c_k}}^{M_k} sim(d_i,d_j)\right\} \tag{3}$$

其中，d 为有效报道；M_k 是子话题 c_k 包含的报道数。

3.2　话题层检测

话题层检测过程类似于 single-pass 算法，并进行了如下改进。

(1)基于词性的特征选择

通过分析网络新闻报道发现，不同词性在报道中所起的作用不同。实词(名词、动词、形容词、数词、量词和代词等)往往用于表达中心思想，而虚词(副词、介词、连词、助词和叹词等)主要用于体现文体和风格。因此，在子话题层检测时，同时考虑实词和虚词有助于准确找到内容相近的报道；但在话题层检测中，报道间内容差别较大，所包含的信息也不尽相同，通常只有一些核心词语将其关联在一起，这种情况下实词发挥的作用更大。因此，本系统在标准 TFIDF 公式的基础上增加了基于词性的权值信息，见式(4)，以提高话题检测

的召回率。

$$w_{ik} = \text{weight}_j \times tf_{ik} \times \text{lb}(\frac{N}{df_i}) \quad (4)$$

其中，w_{ik} 是单词 i 在报道 k 中的权重；weight_j 表示单词 i 所属词性 j 的权值，本系统定义 $\text{weight}_{实词}$=0.99，$\text{weight}_{虚词}$=0.01。tf_{ik} 是单词 i 在报道 k 中出现的频率；N 是训练集合的总报道数；df_i 是训练集合中包含单词 i 的报道数，即单词 i 的文档频率。训练集合是指专门用于统计单词文档频率的大规模语料库，该库中的数据来源于 13 个权威的国内新闻网站，包含 N=626 133 篇报道。

(2)基于平均距离的相似度计算

传统的 single-pass 算法采用最短距离作为报道与话题间的相似度计算。但研究发现，对于互联网新闻报道，应用最短距离往往会产生一个内容非常杂乱的特大话题[3]。因此，本系统采用平均距离消除此现象，即子话题与话题间相似度等于子话题与该话题所包含子话题间相似度的平均。子话题间相似度为子话题有效报道间相似度，见式(5)。

$$sim(c,T_j) = \frac{\sum_{i=1}^{N_j} sim(cd,cd_i)}{N_j} \quad (5)$$

其中，$sim(c,T_j)$ 是子话题 c 与话题 T_j 间相似度；cd 和 cd_i 分别是子话题 c 和 c_i 的有效报道；c_i 是话题 T_j 包含的子话题；N_j 是话题 T_j 包含子话题的个数。

(3)基于滑动窗口的话题跟踪

话题跟踪的具体步骤为：

1)应用话题包含的有效报道组成沿时间轴分布的话题空间。

2)定义处理窗口，处理窗口内的子话题为待跟踪子话题，其所属话题被添加进话题集合 I。

3)计算当前处理子话题与话题集合 I 中所有话题间的相似度。该相似度近似采用当前子话题与话题在处理窗口内所有子话题的相似度的平均，如图 3 中子话题 l 与话题 1 的相似度为

$$sim(\text{l},T_1) = \frac{sim(l,c)+sim(l,f)+sim(l,h)+sim(l,i)}{4}$$

4)选取相似度最大的话题，并记录最大相似度和话题编号；如果最大相似度大于话题层检测阈值 θ_2，则将当前子话题归入已有话题；否则以当前子话题为种子建立新话题，并在话题集合 I 中添加新话题。

5)将当前子话题添加进待跟踪话题空间。重复步骤 3)~步骤 5)，直到处理完所有子话题。

4 实验与结论

本实验选择某时段从国内 13 个新闻网站实际抓取的报道构造测试集，共包括报道 463 篇、人工标注话题 349 个。实验采用美国国家标准与技术研究院(NIST)为 TDT 建立的公认的评测体系[5]进行评价。该评价体系采用式(6)所示的错误代价规范化指标$(C_{\text{Det}})_{\text{Norm}}$。一般最小的$(C_{\text{Det}})_{\text{Norm}}$ 代表系统的最佳性能，简写为 $\min(C_{\text{Det}})_{\text{Norm}}$。

$$(C_{\text{Det}})_{\text{Norm}} = \frac{C_{\text{Miss}} \times P_{\text{Miss}} \times P_{\text{target}} + C_{\text{FA}} \times P_{\text{FA}} \times P_{\text{non-target}}}{\min\left(C_{\text{Miss}} \cdot P_{\text{target}}, C_{\text{FA}} \cdot P_{\text{non-target}}\right)} \quad (6)$$

其中，C_{Miss} 和 C_{FA} 分别代表漏检率和误检率的代价系数(C_{Miss}=10, C_{FA}=1)；P_{Miss} 和 P_{FA} 分别是系统漏检和误检的条件概率；P_{target} 和 $P_{\text{non-target}}$ 是先验目标概率(P_{target}=0.02, $P_{\text{non-target}}$=1-P_{target})。

图 5 给出当子话题层发现阈值 θ_1 取某一指标时，$(C_{\text{Det}})_{\text{Norm}}$ 随话题层检测阈值 θ_2 的变化趋势。从中可看出，当 θ_1=0.28, θ_2=0.25 时，系统达到最佳性能：$\min(C_{\text{Det}})_{\text{Norm}}$= 0.138 8。

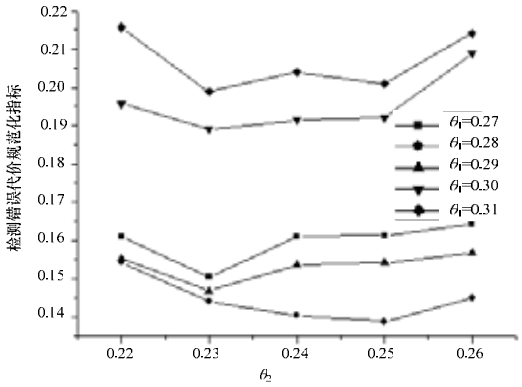

图 5 系统参数估计曲线

为了比较系统性能，应用上述测试集对 single-pass 算法进行测试，得到表 1 所示实验结果。可以看出，本文方法对互联网在线话题的检测效果远好于 single-pass 算法。

表 1 实验结果对比

	本文算法	single-pass 算法
$\min(C_{\text{Det}})_{\text{Norm}}$	0.138 8	0.371 9

5 结束语

本文针对互联网新闻的特点提出了一种面向互联网的在线话题检测算法。该算法的主要特色是：提出子话题概念缓解信息冗余问题；建立包含子话题层和话题层的双层检测结构缓解议题发散问题；建立基于滑动窗口的跟踪策略缓解话题漂移问题等。互联网实际数据的测试表明，本文方法的最小错误代价为 0.138 8，远低于传统 single-pass 算法的 0.371 9。下一步将进一步提高算法效率和添加更多语言学知识。

参考文献

[1] 洪 宇, 张 宇, 刘 挺, 等. 话题检测与跟踪的评测及研究综述[J]. 中文信息学报, 2007, 21(6): 71-85.

[2] Allan J, Papka R, Lavrenko V. On-line New Event Detection and Tracking[C]//Proceedings of SIGIR'98. Amherst, USA: [s. n.], 1998.

[3] Allan J, Harding S, Fisher D, et al. Taking Topic Detection from Evaluation to Practice[C]//Proceedings of the 38th Hawaii International Conference on System Sciences. Big Island, Hawaii, USA: [s. n.], 2005.

[4] 洪 宇, 张 宇, 范基礼, 等. 基于子话题分治匹配的新事件检测[J]. 计算机学报, 2008, 31(4): 687-695.

[5] Allan J. Topic Detection and Tracking: Event-based Information Retrieval[M]. Norvell, MA, USA: Kluwer Academic Publishers, 2002.

编辑 张正兴