种子事件与新颖事件演化关系的话题检测与追踪*

胡耀斌[1],林培光[1],聂培尧[1],耿长欣[1],文 卉[2]
(1.山东财经大学 计算机科学与技术学院,山东 济南 250014;
2.山东财经大学 财政税务学院,山东 济南 250014)

摘 要:在话题检测和追踪过程中,话题漂移的产生往往降低话题检测和追踪的准确率。为了克服这个问题,通过分析新闻报道中种子事件与后续的新颖事件之间的演化关系,强调命名实体词的贡献度,并及时调整话题的重心向量,建立了一种动态的话题检测和追踪模型。实验证明,该模型有效地降低了话题漂移现象在话题检测与话题追踪中的影响。

关键词:种子事件;新颖事件;话题漂移
中图分类号:TP18 **文献标识码**:A **文章编号**:1674-7720(2013)06-0065-03

Topic detection and tracking based on the evolution between the seed events and subsequent novel events

Hu Yaobin[1], Lin Peiguang[1], Nie Peiyao[1], Geng Changxin[1], Wen Hui[2]
(1.School of Computer Science and Technology, Shandong University of Finance and Economics, Ji'nan 250014, China;
2.School of Finance & Taxation, Shandong University of Finance and Economics, Jinan 250014, China)

Abstract: Topic drift will reduce the accuracy of topic detection and tracking. In order to overcome negative effect of topic drift in the process of topic detection and tracking, this paper proposes a new dynamic algorithm. This paper analyses the evolution between the seed events and subsequent novel events about the same topic. The contributions of named entity words are especially emphasized. Meanwhile, the vector of the topic should be timely adjusted. Experiments show that this algorithm can effectively reduce the effect of topic drift in the process of topic detection and tracking.

Key words: seed events; subsequent novel events; topic drift

随着信息技术的不断发展,互联网已经成为人们获得信息的主要来源之一,然而面对互联网的海量信息,用户要集中关注某个感兴趣的主题时,往往感到无所适从。在这种情况下,话题检测与追踪 TDT(Topic Detection and Tracking)应运而生。TDT 概念由美国国防高级研究计划委员会 DARPA(Defense Advanced Research Project Agency)在 1996 年提出,近些年关于 TDT 的研究得到迅速发展,目前已经发展到第五代 TDT 技术评价阶段[1]。

TDT 对话题的定义为: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。话题检测的主要任务是检测识别出系统未知的话题。话题追踪被定义为用一个或几个报道定义一个话题,在一个报道流中顺序检测出对该话题的所有相关报道[2]。

在话题检测和追踪领域存在一种现象,话题会随着时间的变化转移重心,例如当某自然灾害发生时,新闻报道的内容主要是该自然灾害造成的伤亡和损失,随着事件的发生,新闻报道的重心则转移到了灾后的救助和灾后重建,这种话题的动态演变称为话题漂移。本文旨在提出一种能够解决话题漂移的话题检测与追踪模型,使得 TDT 系统在追踪话题漂移的报道时,能够准确地将其划分到所属的话题中。

1 国内外研究现状

国外对 TDT 的研究起步较早,最初的研究参与者不仅包括卡耐基梅隆大学、马萨诸塞大学、宾州大学等一流的大学,还包括 IBM、GE、Dragon Systems 等实力雄厚的公司。马萨诸塞大学采用 Rocchio 算法,利用特征词的不同权重组合实现自适应的话题追踪算法,证明了采用

*基金项目:教育部人文社科一般项目(10YJC880076);山东省自然科学基金项目(ZR2010FL008)

“NUC”权重计算方法可取得最优性能[3]。卡耐基梅隆大学的研究者提出了一种基于决策树的组合系统 BORG TRACK，该系统在话题追踪领域表现优异[4]。IBM 公司在话题检测和追踪系统中采用了两次聚类的策略，使得系统在准确率方面得到很大的提高[5]。另外还有多种不同方法在这项研究中被尝试使用，如 Single-Pass 方法、贝叶斯算法、K-最近邻居方法等，其中比较成功的有 K-最近邻居方法以及多种方法的组合。

话题检测和追踪已经成为国内信息处理领域的热点问题，虽然国内对 TDT 的研究相对国外起步较晚，但经过多年的发展也取得了一些突出的研究成果。贾自艳[6]把命名实体融入到 TDT 系统中，并将时间信息考虑到文本相似性计算的阈值中，有效地提升了 TDT 系统跟踪话题的性能。赵华[7]在 TDT 系统中考虑时间信息的重要性，提出了一种基于时间信息的动态阈值模型，证明了时间信息应该在话题检测系统中得到充分利用。

2 相关技术

2.1 报道模型

文本表示模型共有 3 种：布尔模型、概率模型及向量空间模型，其中向量空间模型应用最为广泛。本文采用向量空间模型作为文本表示模型。众所周知，新闻报道包含 4 个要素：时间、地点、人物和事件。其中时间、地点、人物和组织机构名等 4 种命名实体词对新闻报道内容的贡献度大于其他特征词。为区分命名实体词和其他特征词对文章的贡献度，在构造新闻报道向量空间模型时，提取上述 4 种命名实体词作为命名实体向量，提取除命名实体词外的其他特征词作为内容向量。

一个新闻文本可以表示为：$R=(N_R, C_R)$，其中 $N_R=(t_1, x_1; t_2, x_2; \cdots; t_i, x_i; \cdots; t_n, x_n)$ 表示命名实体变量；$C_R=(l_1, y_1; l_2, y_2; \cdots; l_j, y_j; \cdots; l_m, y_m)$ 表示内容向量。t_i 表示命名实体特征词，x_i 表示命名实体特征词对应的权重，l_j 表示内容特征词，y_j 表示内容特征词对应的权重。

2.2 话题模型

为了方便报道和计算话题的相似度，话题模型应该采取与报道模型相同的表示方法（向量空间模型）。一个话题模型可以表示为：$T=(N_T, C_T)$，其中 $N_T=(t_1, z_1; t_2, z_2; \cdots; t_i, z_i; \cdots; t_n, z_n)$ 表示命名实体变量；$C_T=(l_1, h_1; l_2, h_2; \cdots; l_j, h_j; \cdots; l_m, h_m)$ 表示内容向量。t_i 表示命名实体特征词，z_i 表示命名实体特征词对应的权重，l_j 表示内容特征词，h_j 表示内容特征词对应的权重。

在报道模型中，某特征词的权重等于模型中所有报道的对应特征词权重的加权平均值，即：

$$z_i=\frac{\sum_{i=1}^{num} x_i}{num}, h_i=\frac{\sum_{i=1}^{num} y_i}{num}$$

其中 num 为话题模型中所含报道的个数。

每当有新的报道被划分到话题模型中后，都需要重新计算话题模型的权重，以完成话题模型的更新，目的是让更新后的话题模型能够体现出新加入报道对该模型的影响。

2.3 新闻报道中特征词权重的计算

经过一些文本预处理（去噪分词）后，新闻文本被表示为一系列的词，而词与词之间对文本的贡献是不同的，如何计算这些词的权重显得很重要。显而易见的是，出现次数越多的特征词对文本的贡献越高，表现形式越突出的特征词对文本的贡献越高，例如各级标题中的特征词或加粗后的特征词要比那些普通的特征词具有更高的贡献。

本文在计算特征词权重时，将特征词分成两部分。第一部分是命名实体特征词和特殊内容特征词，命名实体特征词指表示时间、人物、地点、组织机构名的词；特殊内容特征词指那些加粗或出现在各级标题中的内容特征词。第二部分为除第一部分外，无明显表现特征的普通内容特征词。

本文特征词权重计算基于目前应用最为广泛的 TF*IDF 权重计算方法，$\mathrm{tf}(t_i)$ 表示特征词在文档中出现的次数，即词频 TF（Term Frequency），$\mathrm{idf}(t_i)$ 表示 t_i 反文档频率 IDF（Inverse Document Frequency），$\mathrm{idf}(t_i)=\lg\left(\frac{N}{\mathrm{df}(t_i)}+1\right)$，其中 N 表示文档总数，$\mathrm{df}(t_i)$ 表示文档集中含有 t_i 的文档数目，则权重计算公式为：$w_i=\mathrm{tf}(t_i)\times\mathrm{idf}(t_i)$。

关于第一部分特征词的权重计算，本文定义了一组权重辅助值 w_λ 来表示对第一部分特征词中特征信息（命名实体）和表现形式信息（存在于各级标题）的考虑，则该部分特征词的权重计算公式为：$w_i=\mathrm{tf}(t_i)\times\mathrm{idf}(t_i)+w_\lambda$。

第二部分特征词的权重直接利用 TF*IDF 的权重计算方法，即 $w_i=\mathrm{tf}(t_i)\times\mathrm{idf}(t_i)$。

2.4 特征选择

由于新闻文本中含有丰富的词汇量，而能够表示话题核心的词汇却只占一小部分，如果不对特征词加以选择，那么空间向量的维数会变得非常高，加大了相似度计算的复杂度，系统的性能也会随之下降，这就要求对特征项加以选择。本文采用隐含语义分析 LSA（Latent Semantic Analysis）技术对文本向量实施降维，经过验证 LSA 是目前最好的降维方法之一。

隐含语义分析的核心思想是将特征项和文本映射到一个二维的向量空间（矩阵 $A_{i\times j}$）中，假设这个矩阵的秩为 r，其中每行代表一个特征词的权重，每列代表一个文本。然后对矩阵进行奇异值分解，即：$\boldsymbol{A}=\boldsymbol{UBV}^T$，其中 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 均为正交矩阵，$\boldsymbol{B}=\mathrm{diag}(\beta_1, \beta_2, \cdots, \beta_r)$，然后在这 r 个特征值中取前 k 个。

2.5 相似性计算

根据新闻报道 R 与话题 T 的相似性计算结果判断报道是新话题或是已存在话题。本文将相似度的计算分成两部分，即新闻报道的命名实体向量与话题的命名实体向量二者之间的相似度（Sim_n）、新闻报道的内容空间

向量与话题的内容空间向量二者之间的相似度(Sim_c)。

报道 R 与话题 T 之间的相似度为：$Sim=Sim_n+Sim_c$，其中 Sim_n、Sim_c 采用余弦公式进行计算：

$$Sim_n=\cos(N_R,N_T)=\frac{\sum_{i=1}^{n}x_i z_i}{\sqrt{\sum_{i=1}^{n}x_i^2\times\sum_{i=1}^{n}z_i^2}} \quad (1)$$

$$Sim_c=\cos(C_R,C_T)=\frac{\sum_{i=1}^{m}y_i h_i}{\sqrt{\sum_{i=1}^{m}y_i^2\times\sum_{i=1}^{n}h_i^2}} \quad (2)$$

3 话题检测与追踪算法

根据 TDT 对话题的定义，可知话题是由种子事件引起的，新闻报道的内容则是围绕种子事件进行描述的。随着事态的发展，种子事件可能会产生新的状态或情况，新闻报道的内容重心也产生了漂移，这种新的事态情况称为新颖事件，这个过程即为种子事件到新颖事件的演化。新颖事件仍然属于原始话题。在实际生活中，话题发生演变过程中，即新闻报道中产生新颖事件时，常常会有对种子事件或前一个新颖事件的回顾性描述，而且新颖事件一定是发生在种子事件之后。

根据上述内容可以得出一个结论，即新颖事件的报道中常常会有对种子事件或前一个新颖事件的回顾性描述，而事件描述的主要内容是时间、地点、人物等命名实体，这就意味着新闻报道的命名实体向量与所属话题的命名实体向量具有高相似性。设定阈值为 λ_1，若不属同一话题二者的命名实体向量的相似性则低于阈值 λ_1，然后比较新闻报道的内容向量与模型的内容向量相似度，若该相似度大于阈值 λ_2，则仍然认为新闻报道属于该话题。

该算法将报道按时间先后进行排序，依次处理报道流中的报道。具体算法实现如下：

```
Input: R={R1, R2, ··· , Rn}   Output: T={T1, T2, ··· , Tm}
                          //其中 Ri 为新闻报道，Tj 为话题
Begin
    T1={R1}; num(T1)=1; k=1;
    While(i<n){
        if(Sim_n(Ri,Tj)> λ1||Sim_c(Ri,Tj)>λ2) {
            Tj=Tj+{Ri};          //将报道 i 划分到话题 j 中
            num(Ti)++;
            update(Tj);          //更新话题模型，重新计算
                                 话题模型中特征词的权重
        }
        else {
            k++;
            Tk={Ri};                        //创建的新话题
            create(Tk)                      //创建新话题模型
        }
    }
    return{T1, T2, ··· Tm}
End
```

4 实验和结果分析

本文语料以日本政府购买钓鱼岛事件为例，选取了自 2012 年 4 月 16 日起 900 多篇语料，利用中科院分词系统 ICTCLAS 进行分词和词性标注，计算出特征词的词频以及相应的特征词权重。从中抽取 200 篇新闻报道作为样本，发现每篇命名实体中的特征词平均有 121 个，每篇内容特征词平均有 224 个，结合前面所述的特征选择方法，选取命名实体特征词前 80 个，选取内容特征词前 160 个。表 1 是 4 月 17 日新闻报道中部分高频词的权重。

表 1 部分特征词的词频与权重

特征词	词频	特征词权重
钓鱼岛	9	0.908 3
日本	7	0.774 9
中国	7	0.655 0
购买	5	0.713 9

本文 TDT 系统采用美国国家标准技术研究院制定的 TDT 评测体系，即采用准确率、召回率以及二者的综合指标（F1-measure）来评价话题追踪的效率。三个指标的计算公式如下：

准确率：$P=\frac{A}{A+B}$；召回率：$R=\frac{A}{A+C}$；

$$F1-measure=\frac{2PR}{P+R}$$

其中，A 表示系统追踪到的相关新闻报道数；B 表示系统追踪到的不相关新闻报道数；C 表示系统未追踪到的相关新闻报道数；D 表示系统未追踪到的不相关新闻报道数。

通过对前 20 个样本的学习，得到参数的最优值分别为：$w_\lambda=0.07$，$\lambda_1=0.39$，$\lambda_2=0.44$，利用所得参数对剩余报道进行话题追踪，最终得到准确率为 95.24%，召回率为 93.02%，F1-measure 为 94%。从评价指标中可以看出本文提出的基于种子事件和新颖事件时序关系的话题检测和话题追踪模型实现了较好的效果，有效地解决了话题漂移带来的问题。

本文首先介绍了 TDT 系统的相关技术，包括向量空间模型、特征词权重计算、相似度计算等，为体现本系统所陈述的算法思想，并对这些相关技术在一定程度上进行了改进。另外，本文提出了种子事件和后续的新颖事件之间的时序关系，并在此基础上提出了新的话题探测和追踪模型。通过实验证明，该模型能够有效地解决话题漂移带来的问题，保证了 TDT 系统的有效性。

参考文献

[1] ALLAN J.Topic detection and tracking-event based information organization[M].Boston: Kluwer Academic Publisher, 2002: 1241-1253.

[2] CIERI C, STRASSEL S, GRAFF D.Corpora for topic detection and tracking[A].In: ALLAN J.Topic detection and

tracking-event based information organization[M].Boston: Kluwer Academic Publisher, 2002: 33-66.

[3] ROECHIO J.Relevance feedback in information retrieval[A]. In: SALTON G.The smart retrieval system: experiments in automatic document processing[M].New Jersey: Prentice Hall, 1971: 313-323.

[4] MITCHEN T M.机器学习[M].曾华军, 张银奎, 译.北京: 机械工业出版社, 2003.

[5] ALLAN.Topic detection and tracking-Event-based Information Organization[M].Dordrecht: Kluwer Academic Publishers, 2002.

[6] 贾自艳, 何清, 张海俊, 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.

[7] 赵华, 赵铁军, 赵霞.时间信息在话题检测中的应用研究[J].计算机科学, 2008, 35(1): 221-223.

(收稿日期: 2012-10-24)

作者简介:

胡耀斌, 男, 1986 年生, 硕士研究生, 主要研究方向: 网络舆情监控分析。