基于新闻要素的在线新事件检测

李营那 阮 彤 顾春华
(华东理工大学计算机科学与工程系 上海 200237)

摘 要 在线新事件检测的主要任务是从以时间顺序到来的新闻报道中识别出未知事件。提出一种基于新闻要素的自动在线新事件检测方法。首先,构建基于新闻要素的报道和事件表示模型,该模型包括新闻报道地点、人物和内容等要素,使用多维要素的优越性在于可以区别相似事件;为计算各要素对应特征的相似度提供对应的相似度算法:使用基于地理本体树的地名相似度算法计算地点相似度,使用基于维基百科的语义相似度计算方法计算报道内容之间的相似度;为了衡量各要素的重要性,使用 SVM 模型训练得出各要素的权值;最后,以 single-pass 聚类算法为基础,在算法过程中不断修改事件的表示向量以防止事件中心的漂移,同时使用滑动的时间窗口以减少因处理大量不活跃事件引起的时间消耗。实验结果表明该方法可以有效地降低系统的漏检率和误检率,提高事件检测的性能。

关键词 新事件检测 Single-pass 地理本体 语义相似

中图分类号 TP391 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2013.12.026

ONLINE NEW EVENT DETECTION BASED ON NEWS ELEMENTS

Li Yingna Ruan Tong Gu Chunhua
(*Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai 200237, China*)

Abstract The main task of online new event detection (ONED) is to distinguish unknown events from chronological news reports. We propose an automatic ONED method which is based on the news elements. First, the method builds a news elements-based representation model for events and reports, the model includes the elements of news report including place, people and content, the use of multi-dimensional elements has the advantage in being able to differentiate similar events; it provides corresponding similarity algorithms for calculating the similarity of each element's corresponding feature: geographical ontology-based toponym similarity algorithm is used to calculate the place similarity, and Wikipedia-based semantic similarity algorithm is used to calculate the similarity between the contents of report; in order to balance the importance of each element, the weight of each element is derived from the training which uses SVM model; Finally, taking the single-pass clustering algorithm as the basis, the event representation vector is modified constantly in the process of the algorithm to prevent the drift of event centre. Meanwhile the slipped time window is used to decrease the time cost caused by dealing with a lot of inactive events. Experimental results show that the algorithm can effectively reduce the miss probability and false-alarm probability of the system, improves the performance of the event detection.

Keywords New event detection Single-pass Geographical ontology Semantic similarity

0 引 言

随着互联网的迅速发展,网络上的信息呈现爆炸式的增加,同时网络信息也面临着凌乱无序、过多冗余的困境,有价值的信息被湮没。如何快速、准确、有组织地从海量信息中获取用户感兴趣的信息,并以方便用户阅读的方式展示是所有信息科研者的目标。以关键词为主体的信息检索技术是用户从海量信息中获取需求信息的主要途径:首先基于关键词对海量信息进行组织,用户在查找相关信息时,通过关键词进行匹配获得相关信息。然而在许多情况下,用户很难使用关键词准确地表达自己的真实意图,因此基于关键词的信息检索技术很难满足人们的需求。为此,以话题为主线对信息进行组织、然后以话题的方式把相关信息展现给用户,成为信息获取的另一种重要方式。话题(Topic)是指一个种子事件或活动以及与之直接相关的事件或活动[1]。

话题检测与追踪 TDT(Topic Detection and Tracking)是从新闻专线和广播新闻等来源的新闻数据流中自动发现新话题并追踪已知话题发展动态的信息智能获取技术[2]。新事件检测 NED(New Event Detection)是 TDT 的一项重要子任务,NED 的目标是从时序新闻源中检测出一个新闻话题种子事件的第一篇报道[3]。事件(Event)是指发生在特定时间和特定地点的事情[4](如神舟九号飞船发射升空),而不是指一个广泛的概念(如中美关系)。总体而言,一个话题包含多个事件,事件则包含若干报道。NED 的研究在现实中有广泛的应用,如信息源

收稿日期:2012-08-26。国家科技支撑项目(2009BAH46B03)。
李营那,硕士生,主研领域:话题检测。**阮彤**,副教授。**顾春华**,教授。

(电视、社交网络等)的自动监控、证券市场的分析、行业调研、个性化信息定制等。

目前影响 NED 性能的主要因素有三个:(1) 由于不同事件的相关报道中会包含大量相同的词汇,导致难以区分这些相似的事件,如关于"汶川地震"和"玉树地震"的新闻报道中都会频繁包含地震、救灾、伤亡等词汇;(2) 仅使用缺乏语义的关键词作为报道之间的相似度衡量标准,容易把属于相同事件的报道误报为不同事件;(3) 在判别报道是否属于已知事件时划分的阈值,以及在使用多维特征事件表示模型时各特征的权重系数都难以通过经验进行选取。针对上述问题,本文提出一种基于新闻要素的增量式的在线新事件检测方法:采用基于新闻要素的多向量报道和事件表示方法,根据新闻要素将特征词分成三类:地点、人物和主要内容;在进行相似度计算时,采用基于维基百科的简单语义相似度计算方式;使用基于 SVM 的分类算法训练学习报道与事件的判别阈值以及各特征的系数。

1　相关工作

在线新事件检测的核心之一在于相关新闻报道的聚类算法,即在线监视后续报道数据流,如果截获到与之前事件不相关的报道,就检测到一个新事件,否则将该报道归入相关事件簇[1]。

文献[2]中 Dragon 认为在线话题检测是 k-means 算法一个自然的应用,他的研究中将语料中的第一篇报道作为一个初始的簇,其余报道按时间顺序处理,对于每个目标报道计算它和已存在簇之间的距离,当该距离大于阈值时产生一个新的簇也就是检测到一个新事件,否则把目标报道归入距离最近的簇;文献[3]提出了单路径聚类算法,当新报道达到后立即提取该报道的特征术语,建立报道内容的查询表示,然后将该查询和已存在的所有查询进行比较,如果比较结果没有超过阈值则认为检测到一个新事件;文献[4]也采用了单路径聚类方式进行在线新事件检测,当新报道达到后提取报道的特征,建立报道的向量空间表示,然后计算该向量与已有的簇的质心向量的相似度,如果相似度大于阈值则将该报道归入相似度最大的簇,否则创建以该报道为种子的新簇,即检测到新事件;文献[5]利用神经网络的思想改进聚类算法,训练了一个多层的神经网络,在新报道和已存在的簇之间出现本质不同时建立一个新簇,即检测到新事件。文献[11]扩展了基本的增量 TF-IDF 模型,新的模型包含特定源模型,基于特定文档均值的相似度标准化技术、基于特定源对均值的相似度标准均化、基于逆事件频率的术语权重调整和文档分割。NED 的核心是判断两个报道是否属于相同的话题,很明显以上方式并没有利用话题的信息,此外由于需要进行大量的相似度计算在实际的应用中这种方式并不可行。

传统的单纯基于文本聚类算法的事件检测技术难以区分相似的事件。为了解决这个问题,一些学者尝试将自然语言处理(NLP)的技术应用到事件检测中,最常用的自然语言处理技术是命名实体 NE(Name Entities)识别[1]。例如:文献[6]从通常的术语中提出七类 NEs,分别为人名、组织机构、地名、日期、时间、金钱、百分比,经过实验验证"地名"在这些 NEs 中最具有价值;文献[7]采用文本分类和命名实体相结合的策略进行事件检测,每篇文档用以下三个向量表示:文档的所有词(移除停用词)组成的向量、仅包含七种 NE 的向量,除去 NE 的词组成的向量;文献[8]将报道内容及事件使用以下四个语义类组成的向量表示:地点、名称、时间和普通术语,报道与事件的相似度由以上四个子向量的相似度加权得到。然而这些方法都没有解决报道中心的偏移问题,没有考虑特征在文档中的位置对特征重要性的影响。

2　基于新闻要素的表示模型

从事件的定义可以得知时间(when)和地点(where)为事件的两个主要特征,其实事件的另外两大要素也同样重要,即事件相关的人物(who)和事件的主要内容(what)。同样,对于一篇新闻报道,也可以使用这四大要素更加精确地表示。传统的采用单一文本向量表示报道的方法,不仅因为向量的维度过高而使计算过于复杂,同时难以准确表示与区分相似事件。因此,本研究提出了一种基于新闻要素的报道和事件表示模型,该模型由三个与新闻要素相对应的子向量表示:其中地点向量 V_P 包含报道中出现的地名,它对应与新闻报道的 where 要素;名称向量 V_N 包含报道中出现的人名、机构名,该向量对应于新闻报道的 who 要素;内容向量 V_C 描述具体所发生的事件,它对应于新闻的 what 要素。

2.1　报道模型的构建

在构建报道模型之前首先需要对报道进行预处理,预处理步骤主要包括分词、词性标注、除去停用词。本研究中中文的分词和词性标注工具采用开源的 ctbparser[12] 中文依存句法分析工具包,图 1 展示了 ctbparser 分词和词性标注的结果样例。

原文:　在今年的第4号热带风暴"海马"步步逼近广东之际。

结果:　在/p　今年/NT　的/DEG　第4号/OD　热带/NN　风暴/SP。

"/PU"　海马/NR　"/PU步步/AD"　逼近/VV　广东/NR　之际/LC。

图 1　ctbparser 处理结果图

对于一篇已经过预处理的新闻报道 S,构建 S 的表示模型需要经过以下步骤:(1) 报道时间的提取和统一化,提取报道的发布时间,并表示成"YYYY-MM-DD"形式;(2) 利用地名识别算法识别报道中的地名特征,构建地名向量 V_{SD};(3) 由于 ctbparser 将地名、人名和机构名称统一标注为 NR,因此在构建 V_{SN} 时要从被标注为 NR 的词中除去已被识别为地名的词,利用剩下的词构建名称向量 V_{SN};(4) 提取被标注为 NN、VV 的词构建内容向量 V_{SC};(5) 构建报道表示模型 $S=(V_{SD}, V_{SP}, V_{SN}, V_{SC})$,其中后三个子向量均采用向量空间模型表示($V_{Si}=(w_{s1}, w_{s2}, \cdots, w_{sn})$,其中 w_{si} 是使用利用式(1)计算得到:

$$wt_i = \partial tf_i \times \log\left(\frac{N}{n_i}\right) \tag{1}$$

其中 tf_i 是 term_i 在 S 中的词频,N 指在线检测过程中到目前为止已有报道的总个数,n_i 是到目前为止已有报道中包含特征项 term_i 的报道的个数。式(1)对传统的 TF * IDF 进行了改进,主要引进了特征项的位置信息。本研究发现报道的标题和关键词中的词更能体现报道的内容,因此,如果 term_i 出现在报道的标题或者关键词中,取经验值 1,否则取 0.6。

2.2　事件模型的构建

事件的中心随着事件的发展会发生偏移,为了更好地体现事件的动态演变,更准确地表示事件的内容,本研究中组成事件模型的特征是动态改变的,当有新的报道归入某个事件时,则把

新归入的报道的相关特征词加入事件的特征词集合以便形成新的事件模型。事件特征词的选取是在一个特定事件内部进行的,本文认为,词在该事件内的频率越高越能代表该事件。事件的表示同样采用基于新闻要素的多维向量模型:事件的时间向量 V_{ED},地名向量 V_{ED},名称向量 V_{EN},内容向量 V_{EG}。具体算法如下:

(1) 事件的时间为归入该事件的最近一篇报道中描述的事件的发生时间(本文在计算时间衰减时比较的是当前报道与事件中最近一篇报道中事件的时间);

(2) 分别计算新归入事件簇的报道的其余三项特征词的权值并归入到事件的特征词集合中,对集合中的词按照权值从高到低进行排序;

(3) 分别取出排在前面的特定个数的词组成更新后的事件的三个特征向量。

报道的特征词在事件中的权值是通过式(2)计算得到的:

$$w(\delta_i, t_j, E_n) = \frac{\sum_{S_{mj} \in E} W_{ij}(\delta_i, S_{mj})}{N_E} \tag{2}$$

其中 $w(\delta_i, t_j, E_n)$ 表示特征词 δ_i 在 t_j 时刻在事件 E_n 中的权值;N_E 是事件 E 在 t_j 时刻所包含的报道的总数;$S_{mj}(1 \leq m \leq N_E)$ 为事件 E 在 t_j 时刻所包含的报道;$W_{ij}(\delta_i, S_{mj})$ 是特征词 δ_i 在 t_j 时刻在报道 S_{mj} 中的权值,可利用式(1)计算得到。

3 报道和事件相似度的计算

报道和事件相似度分为三部分:地名子向量的相似度、名称子向量的相似度和内容子向量的相似度。在相似度计算时,对于不同的子向量采用不同的相似度计算方法。其中名称、内容子向量采用余弦相似度计算;地名子向量采用基于地理本体树的相似度计算方法。

3.1 基于维基百科的同义消解

在余弦相似度计算时,由于特征词汇缺乏语义信息,同义词之间的相似度结果为 0。在 NED 过程中,由于报道来自不同的源,经常出现表示同一语义的各种同义词,对这些词的错误识别大大降低了 NED 结果的准确度。因此,在计算相似度之前进行同义消解的步骤至关重要。

本文使用基于维基百科的同义词消解方法,利用维基百科中的各种结构化信息(如重定向机制(redirection pages)、信息模块中的别名等),归并同义词,进入相似度计算环节。

3.2 名称子向量与内容子向量余弦相似度计算

名称子向量和内容子向量的相似度使用经典的余弦相似度计算得到,两个向量的余弦相似度计算公式如式下:

$$Sim(V_S, V_E) = \frac{\sum_{i=1}^{L} W_{i,S} \times W_{i,E}}{\sqrt{\sum_{i=1}^{L} W_{i,S}^2} \times \sqrt{\sum_{i=1}^{L} W_{i,E}^2}} \tag{3}$$

其中 V_S、V_E 表示待测的报道 S 和事件 E 的子向量,L 表示 V_S 和 V_E 最大的维度,$W_{i,S}$ 和 $W_{i,T}$ 表示特征词 i 在向量 V_s 和 V_E 中的权值(可利用式(1)得到),$Sim(V_s, V_E)$ 为两个子向量的相似度。

3.3 基于地理树的地名子向量相似度计算

在属于相同事件的报道中,经常会出现不同的地名术语(例如关于"汶川地震"的报道中可能出现汶川、四川、绵阳等地名术语)。利用传统的基于关键词(或字符串)匹配的余弦相似度计算方法,这些术语之间的相似度为 0,但在地理上它们之间是有关联的,有些甚至是同义的。因此,需要将这些术语映射到某种可以体现这些术语之间关系的结构上。本文构建了中国地理树,将每个地名表示为该树中的节点,其中根节点为中国,一级子节点表示所有的省份和直辖市,每个一级节点的孩子为该省的下属市或者直辖市的下属区,以此类推,本研究中的地理树的叶子节点表示村庄或者街道。图 2 展示了地理树的部分内容,计算两个地名的相似度时,只要计算它们的覆盖率,地名 p_1 和 p_2 在地理树中的相似度 $S_p(p_1, p_2)$ 利用式(4)计算得到。

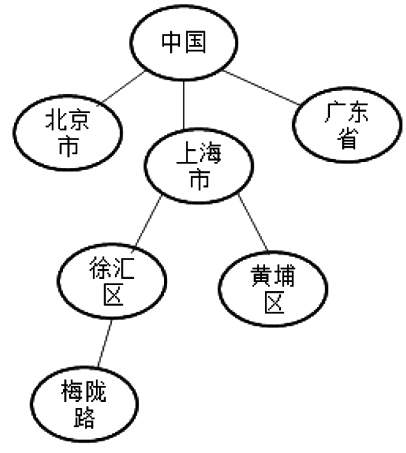

图 2 中国地理树

$$S_p(p_1, p_2) = \frac{2deep(p_1 \cap p_2)}{deep(p_1) + deep(p_2)} \tag{4}$$

其中 $deep(p_i)$ 表示地名 p_i 在地理树中距离根节点的路径长度,例如图 2 中 $deep$(北京市) =1,$deep(p_1, p_2)$ 则表示 p_1 和 p_2 的直接共同祖先距离根节点的路径的长度。报道中通常包含多个地名,报道和事件的地名子向量 $V_{SP} = \{x_1, x_2, \cdots, x_n\}$ 和 $V_{EP} = \{y_1, y_2, \cdots, y_n\}$ 的相似度利用式(5)计算得到。

$$Sim(S_P, E_P) = \sum_{i=1}^{N_S} \sum_{j=1}^{N_E} \frac{n_i}{C_S} \times \frac{n_i}{C_E} \times S_p(i, j) \tag{5}$$

其中 N_S 和 N_E 分别表示报道 S 和事件 E 的地名子向量的维度,C_S 和 C_E 分别表示报道 S 和事件 E 中的地名术语的总个数,n_i 表示地名 i 在报道或者事件中出现的频次,$S_p(i, j)$ 表示两个地名术语 i 和 j 利用地名本体树计算得到的相似度(可由式(4)得到)。

4 新事件检测算法

4.1 滑动时间窗口

在 NED 过程中通常把新报道与目前已有的所有事件进行比较,来判断是检测到新事件还是将报道归入已有的事件。这种检测方式需要的时间与新闻报道的数目成正比,而实际中互联网上的新闻是无限制增加的,因此上述方式在实际应用中不适合。本研究中经过分析大量报道发现事件存在一定的生命周期。因此,在事件检测过程中只需要将报道与一定时间段中的

事件进行比较,本文利用滑动的时间窗口来实现。如图3 所示,以当前时间(在本文中指目标报道的发布时间)t_e 为终点,沿着时间轴向前开设固定时长的窗口,设始点为 t_b,仅检测属于窗口中的事件的报道,即事件的结束时间大于 t_b 的事件,在而对窗口外的报道(如图3 中的事件 a)不再处理。采用滑动窗口的好处是将有限的资源用于处理最需要处理的部分,有效避免了大量已被网民遗忘的事件浪费系统的处理时间。

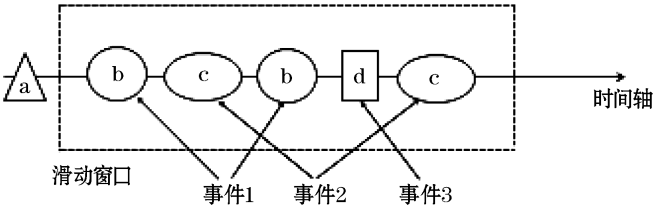

图3　滑动时间窗口示意图

4.2　基于 SVM 的 NED 算法

作为 Single-Pass 聚类算法的一个经典应用,绝大部分的 NED 算法都基于 Single-Pass 聚类算法。对于陆续到来的报道,需要与之前的事件进行比较,如果存在事件与该报道的相似度大于预先设定的某个阈值时,则把该报道归入与之相似度最大的事件。在此过程中,两个(组)参数的选择非常困难,一是各特征的权重系数,二是比较的阈值。其实对每一报道的判别都可以视为一次分类过程,对于归入的类别视为正例,其他则视为负例。因此,本文采用 SVM 分类算法对每次判别过程进行建模,所使用的特征即为前文所提到的三维向量(V_{SP},V_{SN},V_{SG})。

本文的 NED 算法具体过程如下:

(1) 对报道 S 进行预处理,建立报道 S 的表示模型;

(2) 如果 S 是报道流中的第一篇报道,则建立以 S 为中心的事件 E,按照 2.2 节的方法建立事件模型,置 E 的结束时间为 S 的发布时间;

(3) 如果 S 不是报道流中的第一篇报道,则置滑动时间窗口的终点为 S 的发布时间,对于滑动时间窗口内的事件,使用训练的 SVM 模型进行判别;

(4) 如果当前报道属于某一事件,则把它归并到所属事件并更新事件的表示模型;否则建立一个以 S 为中心的事件并建立其表示模型;

(5) 重复以上过程直到所有报道处理完为止。

5　实验结果和分析

5.1　评测语料

本研究采集了新浪网 2011 年新闻频道(内地新闻)的 14 322个报道,并从对应的专题频道中经过处理得到 48 个事件(如表 1 所示)。列表中包含各种类型的事件(如自然灾害、人为事故等);事件包含的报道数目不尽相同,最多的事件“温州动车追尾脱轨事故”包含一千余篇报道,而最小的事件“大连新港油罐起火”则只包含 4 篇报道;各事件的时间跨度也从 2 ~ 3 天到几个月不等。为了确定多少个事件能训练出可靠的判别模型,本文采用逐渐递增的事件作为训练语料:从 12 开始,增量为 3。

表 1　事件信息

事件名称	报道数	开始时间	截止时间
江苏丰县发生校车侧翻事故	106	2011 -12 -12	2012 -01 -05
济青高速汽车连环相撞	12	2011 -12 -08	2011 -12 -08
武汉建设银行爆炸	56	2011 -12 -01	2012 -01 -09
香港旺角火灾	21	2011 -11 -30	2011 -12 -06
大连新港油罐起火	4	2011 -11 -22	2011 -11 -22
甘肃庆阳幼儿园校车被撞	147	2011 -11 -16	2011 -12 -30
西安建筑物爆炸	88	2011 -11 -14	2011 -12 -01
唐山果皮中毒案	7	2011 -11 -10	2011 -11 -11
云南曲靖煤矿事故	97	2011 -11 -10	2012 -01 -17
北京遭遇十年来最大降雨	87	2011 -06 -23	2011 -06 -29
河南义马发生矿难	58	2011 -11 -04	2011 -11 -11
湖南衡阳霞流冲煤矿矿难	12	2011 -10 -30	2011 -11 -27
两岁女童遭两车碾压	172	2011 -10 -16	2011 -11 -29
十四名登山者在四姑娘山上失踪	31	2011 -09 -30	2011 -10 -13
强热带风暴尼格	96	2011 -10 -02	2011 -10 -07
台风纳沙	120	2011 -09 -27	2011 -10 -10
上海地铁十号线发生列车相撞事故	191	2011 -09 -27	2011 -10 -10
河南洛阳性奴案	25	2011 -09 -23	2011 -09 -29
河北深州监狱罪犯越狱	32	2011 -09 -14	2011 -11 -18
江西湖北交接发生 4.6 级地震	37	2011 -09 -10	2011 -09 -16
湖南邵阳沉船事故	44	2011 -09 -09	2011 -10 -09
日本首相野田佳彦访华	29	2011 -11 -17	2011 -12 -26
萨科齐访华	21	2011 -08 -23	2011 -08 -26
法国总统萨科齐访华	8	2011 -03 -30	2011 -03 -31
南京劫持人质事件	45	2011 -08 -30	2011 -09 -06
黑龙江煤矿透水事故	66	2011 -08 -23	2011 -09 -01
中石油大连石化分公司发生火灾	28	2011 -08 -29	2011 -09 -01
台风南玛都	88	2011 -08 -28	2011 -09 -02
台风梅花	356	2011 -08 -03	2011 -08 -10
故宫一级文物损坏事故	90	2011 -08 -01	2011 -09 -08
温州动车追尾脱轨事故	1283	2011 -07 -23	2012 -01 -20
京珠高速客车起火	29	2011 -07 -22	2011 -12 -06
渤海湾油田溢油事故	489	2011 -07 -03	2012 -01 -30
2011 年第 4 号热带风暴海马	65	2011 -06 -21	2011 -06 -26
云南腾冲 5.2 级地震	20	2011 -06 -20	2011 -06 -23
新安江水体遭化学品污染	44	2011 -06 -05	2011 -06 -16
我国禁止生产含双酚 A 婴幼儿奶瓶	22	2011 -04 -21	2011 -06 -01
台湾塑化剂污染食品事件	279	2011 -05 -25	2011 -09 -17
故宫博物馆院展品失窃	138	2011 -05 -10	2012 -01 -21
贫困山区小学生免费午餐计划	7	2011 -04 -19	2011 -04 -19
李庄漏罪案	93	2011 -03 -29	2011 -12 -13
上海市部分超市售染色馒头	82	2011 -04 -12	2011 -09 -26
四川炉霍 5.3 级地震	11	2011 -04 -09	2011 -04 -10
西宁百货商场火灾	17	2011 -04 -09	2011 -04 -10
乌鲁木齐公交车与列车相撞	23	2011 -03 -24	2011 -03 -28
2011 年俄罗斯总理普京访华	52	2011 -09 -30	2011 -10 -13
美国副总统拜登访华	93	2011 -08 -05	2011 -08 -26
天宫一号飞行器发射	420	2011 -04 -30	2011 -12 -29

5.2 评测标准

NIST 为 TDT 建立了完整的评测体系,在进行新事件检测的结果分析时,通常都采用此评测体系。评测标准是建立在系统漏检率和误检率的基础上的。TDT 评测公式定义如下:

$$C_{Det} = C_{Miss}P_{Miss}P_{target} + C_{FA}P_{FA}P_{non-target} \quad (6)$$

其中 C_{Miss} 和 C_{FA} 分别表示漏检率和误检率的代价系数,P_{Miss} 和 P_{FA} 分别表示漏检率和误检率的条件概率,P_{target} 和 $P_{non-target}$ 是先验目标概率($P_{non-target} = 1 - P_{target}$),$C_{Det}$ 是综合了漏检率和误检率的性能损耗代价。评价 TDT 系统性能时常采用 C_{Det} 的规范化表示$(C_{Det})_{Norm}$,定义如下:

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

5.3 实验结果分析

在实验过程中,使用逐步递增的训练数据训练 SVM 模型,并把模型在剩余的事件上进行测试。由于本文使用机器学习的方式学习算法中的各参数,因此得到的 DET 图并非一系列曲线而是对应于每一组训练语料得到相应参数所对应的点。实验结果如表 2 所示,从中可以得出以下结论:当训练的话题数量达到 24 左右时,增加训练语料并不能进一步改进实验结果。由于 SVM 模型的高效性,使用 30 个事件作为训练语料仍然具备非常快的速度,因此,在之后的实验中,都使用 30 个事件作为训练语料。

表 2 使用不同训练数据时的漏检率与误报率

训练事件	12	15	18	21	24	27	30
漏检率%	43.3	36.6	34.2	31.2	30.9	30.7	30.6
误报率%	0.105	0.081	0.063	0.051	0.47	0.46	0.46

由于某些报道内容很长,在进行特征抽取后会得到比较大的向量;而随着报道不断加入到事件中,表示事件的向量也会越来越大。为节约计算的空间,在算法中可以只选取一定数量的特征词代表向量。为了确定多少词能有效的代表报道和事件,本文分别使用 50、75、100、125、150、200 进行了实验,其结果如表 3 所示。从表中可以发现,100 词已足以表示事件与报道的信息,继续增加数量反而可能会对实验结果产生负面影响(如取 150 时的漏检率与误报率均较 100 时差)。

表 3 使用不同数量的词表示报道与事件时的漏检率与误报率

训练事件	50	75	100	125	150	200
漏检率%	31.1	30.7	30.6	30.6	30.8	30.6
误报率%	0.48	0.46	0.46	0.46	0.47	0.46

在算法过程中使用了滑动时间窗口控制算法的计算复杂度,窗口选取越小,算法速度当然越快,但是窗口太小必然会影响实验结果。为了选择合适大小的时间窗口,本文分别在时间窗口为 15 到 1(每次递减 1,单位为天)的情况下进行了实验,实验结果说明当时间窗口为 8 时得到的结果良好,并且花费的时间也相对较少。

最后,本文实现了文献[4]中提出的 NED 单路径算法和文献[9]中的简单语义算法,并把其结果与本文的算法进行了对比,如图 4 所示,本文算法的误报率低于 1%,漏检率为 32%,而在同样的误报率下文献[4,9]的漏检率分别为 50% 和 70%,可以看出本文算法的误报率和漏检率明显低于文献[4]和文献[9]中的算法,本文的算法明显具有优越性。

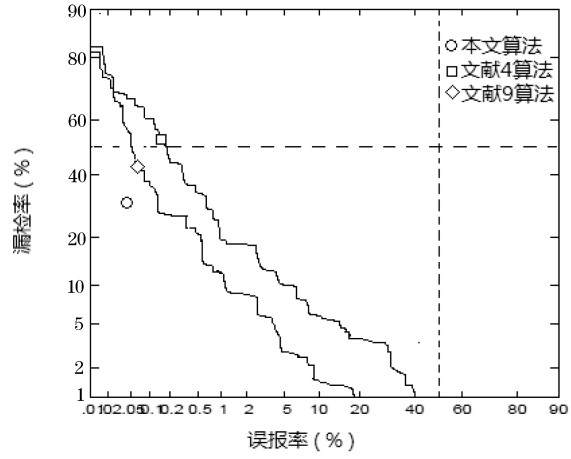

图 4 算法对比

6 结 语

本文提出的基于新闻要素的自动在线新事件检测算法,充分挖掘了新闻报道的各要素在新事件检测过程中的用途,并使用机器学习方法学习各要素的权重。在计算地点的相似度时,使用了基于地理本体树的相似度计算方法,有效地解决了使用关键词计算地名相似度时的缺陷(如同一个地点不同名称、大小地域间的相似度计算)。同时,为了克服传统的基于关键词模型的不足,本文使用了基于维基百科的语义相似度计算方法计算人际关系内容的相似度。实验结果表明该算法能有效降低误报率和漏检率。

参 考 文 献

[1] 洪宇,张宇,刘挺. 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,Carbonell J,Doddington G,et al. Topic detection and tracking pilot study:Final report[C]//Proceedings of the DARPA BroadcastNews Transcription and Understanding Workshop,February 1998:194-218.

[3] Allan J,Papka R,Lavrenko V. On-line New Event Detection and Tracking[C]//The proceedings of SIGIR 98,University of Massachusetts Amherst,1998:37-45.

[4] Yang Y,Pierce T,Carbonell J. A study on Retrospective and On-Line Event detection[C]//Proceedings of the 21st annual international ACM SIGIR conference on Research and Development in Information Retrieval,1988:28-36.

[5] Seo Y,Sycara K. Text clustering for topic detection. Pittsburgh: Robotics Institute,Carnegie Mellon University,2004:1-11.

[6] Yang Y,Zhang J,Carbonell J. Topic-conditioned novehy detection [C]// Hand Detal. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,New York:ACM Press,2002:688-693.

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

[8] Brants T,Chen F,Farahat A. A system for new event detection[C]// Proceedings of the 26th SIGIR Conference on Research and Development in Information Retrieval,2003.

(下转第 176 页)

从图 7 中可以看出,负载均衡系数的变化区间为(1,1.0012),说明进入到各个中间交换平面进行处理的信元个数的差别不大,与 VIQ PPS 相比具有相似的负载均衡性能。从图 8 中可以看出,吞吐率随着负载的增加有所下降,但下降的幅度较小,在最大负载情况下,系统吞吐率也能达到 99% 以上,系统具有较好的吞吐率性能。

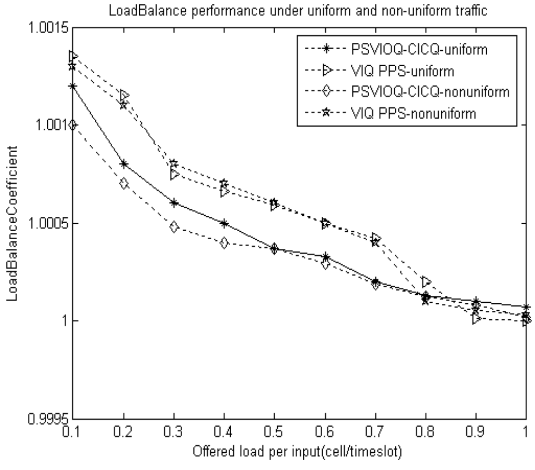

图 7 均匀和非均匀业务源下的负载均衡系数

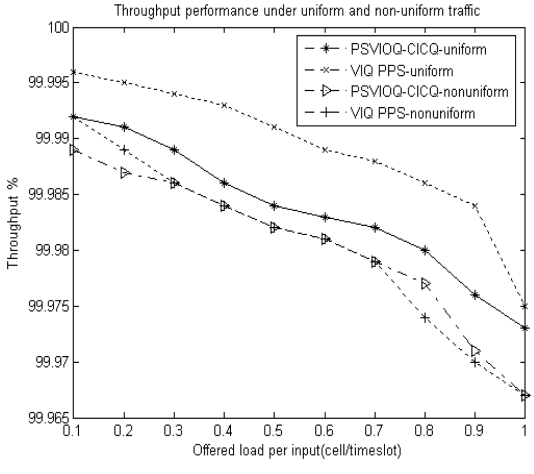

图 8 均匀和非均匀业务源下的的吞吐率

4 结 语

本文提出一种基于 CICQ 支持区分 QoS 的 PSVIOQ-CICQ 解决方案。该方案设计中,我们采用在输出缓存引入 VIQ 队列结构的办法保证信元的传输顺序,基于此设计负载均衡器和分组整合器的调度算法,能够为不同服务需求的业务提供 QoS 支持。从上述仿真实验结果的分析可看到,该并行系统解决方案能够对进入系统的负载进行均衡的分配,在无需内部加速的情况下能够获得 99% 以上的吞吐率,具有较好的吞吐性能,同时仿真结果中吞吐率、负载均衡系数以及时延性能与中间交换平面数的关系表明系统性能随着中间交换平面数目的增加未出现明显的下降,说明系统具有良好的可扩展性,仿真结果的时延性表明系统能够为不同服务需求的业务提供 QoS 保障支持,该方案基本达到了设计目标的要求。

但从分析中我们也看到,系统的吞吐率性能和时延性能在非均匀业务源下还有进一步提升的空间,特别在负载较重的情况下,时延性能有很大的改进提升空间。

参 考 文 献

[1] Iyer S, McKeown N W. Analysis of the parallel packet switch architecture[J]. IEEE/ACM Trans. on Networking, 2003, 11(2): 314 - 324.

[2] Aslam A, Christensen K J. A parallel packet switch with multiplexors containing virtual input queues[J]. Computer Communications, 2004, 27: 1248 - 1263.

[3] Khotimsky D, Krishnan S. Evaluation of open-loop sequence control schemes for multi-path switches[C]//Proc. of the IEEE ICC. Piscataway: Institute of Electrical and Electronics Engineers Inc, 2002. 2116 - 2120.

[4] M neimneh S, K Siu. Scheduling unsplittable flows using parallel switches[C]//Proc. of the IEEE ICC. Piscataway: institute of Electrical and Electronics Engineers Inc, 2002: 2410 - 2415.

[5] Khotimsky D, Krishnan S. Towards the recognition of parallel packet switches[C]//Proc. of the Gigabit Networking Workshop in Conjunction with IEEE INFOCOM. Piscataway: Institute ofElectrical and Electronics Engineers Inc, 2001.

[6] Iyer S, McKeown N. Making parallel packet switches practical[C]//Proc. of the IEEE INFOCOM. Piscataway: institute of Electrical and Electronics Engineers Inc, 2001: 1680 - 1687.

[7] 戴艺,苏金树,孙志刚,一种维序的基于组合输入输出排队的并行交换结构[J]. 软件学报,2008,19(12):3207 - 3217.

[8] 马祥杰,李秀芹,兰巨龙,等. 一种多级多平面分组交换结构中的带宽保证型调度算法[J]. 电子与信息学报,2009,31(6):1475 - 1478.

[9] 马祥杰,李秀芹,兰巨龙,等. 一种新型可扩展的多级多平面分组交换结构的图论模型与性能分析[J]. 电子与信息学报,2009,31(5):1026 - 1030.

[10] Shi L, Li W J, Liu B. Flow-based packet-mode load-balancing for parallel packet switches[J]. Journal of High Speed Networks, 2010, 17(2): 97 - 128.

[11] L Shi, B Liu, W J Li, et al. DS-PPS: A Practical Framework to Guarantee Differentiated QoS in Terabit Routers with Parallel Packet Switch [C]//Proceeding of the 25th IEEE INFOCOM 2006, Barcelona, Spain, April 23 - 29, 2006.

[12] Li Xiuqin, Li Xiuli, Lan Julong. A In-order Queuing Parallel Packet Switch Solution Based on CICQ[C]//CCCM2010 ,IEEE computer society, Yangzhou China 2010. 8. 22.

[13] Yang F, Wang Z K. A parallel packet switch supporting differentiated QoS based on weighted layer assignment[C]//Proc. of the WiCom'09, Beijing, China, 2009: 4286 - 4289.

[14] Li Xiuqin, Yang Xiliang, Lan Julong. A Differentiated QoS Supporting Scheduling Algorithm Based on Identifier[C]//ICACTE2010, IEEE computer society, Chengdu, China 2010. 8. 22.

(上接第 104 页)

[9] Juha M, A M, Marko S. Applying semantic classes in event detection and tracking[C]//Sangal R, Bendre SM. Proceedings of International Conference on Natural Language Processing(ICON). Mumbai, India. 2008: 175 - 183.

[10] Papka R. On-line New Event Detection, Clustering and Tracking[D]. Department of Computer Science. UMASS, 1999.

[11] Farahat A, Chen F, Brants T. Optimizing story link detection is not equivalent to optimizing new event detection[C]//ACL, 2003: 232 - 239.

[12] http://ctbparser. sourceforge. net/.

基于新闻要素的在线新事件检测

作者： 李营那， 阮彤， 顾春华， Li Yingna， Ruan Tong， Gu Chunhua
作者单位： 华东理工大学计算机科学与工程系 上海200237
刊名： 计算机应用与软件

ISTIC

英文刊名： Computer Applications and Software

年，卷(期)： 2013(12)

参考文献(12条)

1. 洪宇;张宇;刘挺 话题检测与跟踪的评测及研究综述[期刊论文]-{H}中文信息学报 2007(6)

2. Allan J;Carbonell J;Doddington G Topic detection and tracking pilot study:Final report 1998

3. Allan J;Papka R;Lavrenko V On-line New Event Detection and Tracking 1998

4. Yang Y;Pierce T;Carbonell J A study on Retrospective and On-Line Event detection 1988

5. Seo Y;Sycara K Text clustering for topic detection 2004

6. Yang Y;Zhang J;Carbonell J Topic-conditioned novehy detection 2002

7. Kumaran G;Allan J Text classification and named entities for new e-vent detection 2004

8. Brants T;Chen F;Farahat A A system for new event detection 2003

9. Juha M,A M;Marko S Applying semantic classes in event detection and tracking 2008

10. Papka R On-line New Event Detection,Clustering and Tracking 1999

11. Farahat A;Chen F;Brants T Optimizing story link detection is not equiva-lent to optimizing new event detection 2003

12. 查看详情