ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.4, April 2008, pp.817−828
DOI: 10.3724/SP.J.1001.2008.00817

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于词元再评估的新事件检测模型*

张 阔+, 李涓子, 吴 刚, 王克宏

(清华大学 计算机科学与技术系,北京 100084)

A New Event Detection Model Based on Term Reweighting

ZHANG Kuo+, LI Juan-Zi, WU Gang, WANG Ke-Hong

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

+ Corresponding author: Phn: +86-10-62771736, E-mail: zkuo99@mails.tsinghua.edu.cn

Zhang K, Li JZ, Wu G, Wang KH. A new event detection model based on term reweighting. *Journal of Software*, 2008,19(4):817−828. http://www.jos.org.cn/1000-9825/19/817.htm

Abstract: New event detection (NED) is aimed at detecting from one or multiple streams of news stories the one being reported on a new event (i.e. not reported previously). Preliminary experiments show that terms of different types (e.g. Noun and Verb) have different effects for different classes of stories in determining whether or not two stories are on the same topic. Unfortunately, conventional approaches usually ignore the fact. This paper proposes a NED model utilizing two approaches to addressing the problem based on term reweighting. In the first approach, the paper proposes to employ statistics on training data to learn the model for each class of stories, and in the second, the paper proposes to adjust term weights dynamically based on previous story clusters. Experimental results on two linguistic data consortium (LDC) data sets: TDT2 and TDT3 show that both the proposed approaches can effectively improve the performance of NED task, compared to the baseline method and existing methods.
Key words: new event detection; information retrieval; name entity; term reweighting

摘 要: 新事件检测(new event detection,简称 NED)的目标是从一个或多个新闻源中检测出报道一个新闻话题的第一个新闻.初步实验发现,在对不同类别的新闻报道进行新事件检测时,其不同类型的词元往往具有不同的敏感程度.而传统方法往往将所有的词元等同看待.重点研究在新事件检测模型中,对于不同词元的权重设定问题.提出利用统计方法优化不同类别新闻对于不同词性词元的权重参数;提出利用已有新闻簇信息动态更新词元权重的方法,采用在新闻之间(而非新闻与新闻簇之间)计算相似度的形式,发挥两种比较形式的优点.在 Linguistic Data Consortium(LDC)公共数据集 TDT2 与 TDT3 上进行实验,实验结果表明,这两种改进方法的效果明显,性能与同类系统相比有显著提升.
关键词: 新事件检测;信息检索;命名实体;词元再评估
中图法分类号: TP181 **文献标识码**: A

新事件检测(new event detection)是话题检测与跟踪(topic detection and tracking,简称TDT)[1]研究课题中的5 项任务之一.TDT课题致力于研究对来自不同新闻源的多语言新闻文本进行有效的组织、搜索与结构化的技

* Supported by the National Natural Science Foundation of China under Grant No.90604025 (国家自然科学基金)
Received 2006-10-30; Accepted 2007-01-25

术.NED的目标是检测出报道一个新闻话题(topic)种子事件的第一个新闻(story)*.话题定义为:“一个种子事件以及所有与其直接相关的事件(event)与活动”[2].事件的定义为:“在一个确定时间,确定地点发生的事情”[3].在本文的工作中,话题和事件都表现为新闻的集合.例如,一个炸弹在一幢大楼中爆炸.爆炸本身为此话题的种子事件,后续发生的营救行动、罪犯搜索、抓捕以及审判均属于此话题的直接相关事件.NED在金融市场、新闻分析、情报收集等诸多领域有其实际应用.重要的信息常常被每天生成的海量新闻数据所淹没,NED系统可以快速识别出新的事件(种子事件),从而协助像政府分析员、金融分析员、股票交易人这样需要以最快速度了解已发生的事件的工作人员.

近年来,关于NED的一些研究工作集中在使用对命名实体的特殊加权来提高NED的效果[4−7].这些方法都没有考虑不同的命名实体和词性对于不同类别新闻描述的特殊属性.然而,不同类别新闻的重点内容往往不同,在表达这些新闻内容时所使用的词汇往往差别很大.比如,对于选举类别的新闻来说,选举人的姓名是非常重要的信息;而对于自然灾害类别的新闻来说,灾害发生的地点是非常重要的信息.此外,大多数比较成功的NED系统采用把当前新闻与之前接收到的所有新闻进行比较,当相似度大于一个阈值时,则认为此新闻为一个旧新闻.也有系统将之前的新闻组织成新闻簇(每个新闻簇对应一个话题),并将新的新闻与之前的各个新闻簇进行比较.文献[8,9]中的实验证明,前一种方式可以得到较好的结果.NED的核心问题是判断两个新闻是否属于同一个话题,而采用新闻与新闻比较的方式不能充分利用话题信息.同时,采用新闻与新闻簇的比较方式往往由于话题内容的分散而使得新闻与新闻簇的相似度过低.

本文根据从训练数据中得到的统计信息发现,不同类别的新闻对于不同词性的词元在判断新闻是否属于同一话题时有着不同的敏感程度,同时,本文提出利用统计结果优化不同类别新闻对于不同词性词元的权重.此外,本文提出利用已有新闻簇的词元分布特点动态更新词元的权重,并采用第一种方案中在新闻之间(而非新闻与新闻簇之间)计算相似度的比较形式,这样可以在充分利用新闻簇信息的同时避免话题内容分散带来的问题.实验结果表明,这两种改进方法的效果明显,性能较之同类系统有显著的提升.

本文第 1 节分析国内外相关的研究方法.第 2 节给出传统新事件检测系统使用的基本模型.第 3 节详细介绍本文提出的改进模型,包括改进的新闻描述方法及新事件检测过程.第 4 节描述实验数据、实验设计及评价标准.第 5 节是实验结果及分析.最后总结全文.

1 相关研究

在NED研究中,Papka等人提出了Single-Pass聚类的思想[10].当遇到一个新的新闻d时,对内容进行预处理并生成相应的向量表示,将此新闻与之前的所有新闻进行比较并得到相似度,若与所有新闻的相似度均小于阈值θ,则认为该新闻描述了一个新的事件.这种方式只是简单考虑新闻之间的相似度而忽略了话题的作用.Lam等人[11]则提出将新的新闻与之前的新闻簇(每个新闻簇对应一个话题)进行比较,若存在一个新闻簇相似度高于阈值,则将此新闻加入到相似度最高的新闻簇,并调整此新闻簇的向量表示;否则,生成一个新的新闻簇并添加此新闻.这种方式由于话题内重点分散也不能得到很好的效果.

近年来,NED大部分研究集中在对新闻的表示模型与新闻间的相似度模型的改进方面.Stokes等人的研究集中在对新闻表示模型的改进[12],将新闻的表示分为两个部分:第一部分为普通的文本特征向量,第二部分通过WordNet中的词汇链扩展而成.两种表示通过线性方式进行组合,但文中实验结果的改善并不十分明显.Brants等人的研究则集中在对相似度模型的改进[13]上,包括:根据不同的新闻源,建立不同的词频模型;对不同新闻的相似度进行归一化;对不同新闻源之间的相似度进行归一化;对新闻内容进行分割;使用Hellinger距离代替Cosine距离计算新闻相似度,并在TDT的测试数据上获得了较好的效果.

另外一些工作则发掘对命名实体的特殊处理方式.Yang等人在*tf-idf*模型基础上直接对地点名称给予 4 倍

* 由于历史原因,在TDT中,新事件检测的目标实际上是检测报道一个话题的第一个新闻,而非一个事件的第一个新闻.因此,若没有特殊说明,文中的“新事件”均指一个话题的第一篇新闻.

的加权[4].DOREMI研究组计算人名、地名、时间的语义相似度,并结合文本相似度得出最终的相似度[5,6].UMass的TDT研究组则将新闻中的词元分为命名实体与非命名实体两部分,并发现,一些新闻类别使用仅包含命名实体的词元向量可达到较好的效果;而另外一些新闻类别则使用仅包含非命名实体的词元向量可达到较好的效果[7].文献[4,7]的工作都使用了文本分类技术辅助事件检测.文献[7]首先对新闻分类,然后统计各个类别对于命名实体的敏感性.文献[4]首先对新闻进行分类,然后在每个类中找到频繁词,并在新闻表示中去除这些频繁词.

TDT 领域的研究在国内也逐渐受到重视.文献[14]借鉴 Single-Pass 聚类思想,并结合新闻要素给出一种基于动态进化模型的事件探测和追踪算法.此外,国内对于 TDT 的研究还包括文献[15,16]等.由于使用的实验数据不同,本文未与这些系统进行比较.

2 基本模型

本节介绍的模型为大多数 NED 系统所采用,本文以该模型作为改进和扩展的基础.NED 问题的输入为按时间顺序得到的新闻流,在处理当前新闻时的可用信息仅包括之前得到的新闻,输出为对每篇新闻是否报道了新事件的判断以及相应的支持度.一般地,一个 NED 模型包括 3 个部分,即新闻描述、新闻间相似度计算与新事件检测过程.

2.1 新闻描述

在建立新闻的描述时往往要经过新闻预处理,预处理过程包括分词、简写识别与名称归一化、词性标注.使用文献[17]中包含 418 个停用词的词表去除停用词,使用*K*-stem算法计算词根[18],随后对每个新闻生成词频向量.最后根据词频向量生成新闻的带有词元权重信息的新闻描述向量.

一般计算词元权重的基本模型采用增量的*tf-idf*模型[8].增量的*tf-idf*模型每经过一个时间窗口更新一次模型,在时间窗口t,模型的更新方式如下:

$$df_t(w)=df_{t-1}(w)+df_{D_t}(w) \tag{1}$$

其中,D_t为t窗口内的新闻集,$df_{D_t}(w)$为D_t内w词元的文档频率,$df_t(w)$为到时间t为止得到的文档频率.在本文的实际应用中,每个时间窗口包含 50 篇文档.

经过上面的处理后,在时刻 t 的每个新闻 d 可以描述为

$$d\rightarrow\{weight(d,t,w_1),weight(d,t,w_2),\ldots,weight(d,t,w_n)\},$$

其中,n 为新闻 d 中的词元数量,$weight(d,t,w)$代表词元 w 在 t 时刻新闻 d 中的权重:

$$weight(d,t,w)=\frac{\log(tf(d,w)+1)\times\log((N_t+1)/(df_t(w)+0.5))}{\sum_{w'\in d}\log(tf(d,w')+1)\times\log((N_t+1)/(df_t(w')+0.5))} \tag{2}$$

其中,N_t为到t时刻为止的新闻数目,$tf(d,w)$为w词元在新闻d中出现的次数.

2.2 新闻间相似度计算

本文使用 Hellinger 距离计算新闻之间的相似度,对两个新闻 d 和 d',它们之间的相似度表示为

$$sim(d,d',t)=\sum_{w\in d,d'}\sqrt{weight(d,t,w)\times weight(d',t,w)} \tag{3}$$

2.3 新事件的检测过程

新事件检测的过程如下:对于在时间t加入的新闻d,将d与之前获得的所有新闻文档进行比较,根据其中的最大相似度得到d为新事件报道的支持度[13]:

$$n(d)=1-\max(sim(d,d',t)),d'\text{为在 }d\text{ 之前出现的文档} \tag{4}$$

若支持度 $n(d)$大于一个阈值θ,则认为 d 报道了一个新事件;反之,则认为描述的是已报道事件.同时,支持度与阈值θ的差越大,表明决策的自信度越高.

3 改进模型

基本模型采用新闻间的相似度表示新闻的“新”或“旧”,这种情况下,对词元赋予合适的权重就成为改善效果的一个有效途径.举例来说,当两篇同一话题的新闻共同包含一个词元时,如果这个词元是两篇新闻所在话题的关键词元,则它可以对两篇新闻同属一个话题提供重要依据.因此,正确判断一个词元对于所在话题的关键程度是改进 NED 的一个重要因素.本文通过统计信息发现,不同类别的新闻对于不同种类的实体名称和词性有着不同的偏好,提出使用统计结果优化参数空间,改善 NED 效果.另一方面,本文根据词元的特点对其进行分类,并讨论各个类别词元在*tf-idf*模型下的问题与解决方法,增加对新闻簇信息的利用,动态更新词元权重,以更好地表达新闻内容.

3.1 基于词性加权的新闻描述

本文使用开源软件 open-NLP[**]标注命名实体类别与词性.命名实体包括人名(person)、组织机构名(organization)、地名(location)、日期(date)、时间(time)、货币(money)、百分比(percentage)7 种,词性则选取名词(NN)、形容词(JJ)、动词(VB)、副词(RB)、数字(CD)5 种.为了叙述方便,若不作特殊说明,后续行文则将命名实体类别和词性统称为词性.

因为NED任务需要正确判断两篇新闻是否属于同一话题,所以,本文首先使用χ^2统计方法统计词元与话题的相关度.对于一个词元w和一个话题T,首先得到一个依赖表(见表 1).

Table 1 A 2×2 contingency table

表 1 一个 2×2 的依赖表

Story number	Belong to topic T	Not belong to topic T
Include w	A	B
Not include w	C	D

χ^2统计[19]计算方法如下:

$$\chi^2(w,T)=\frac{(A+B+C+D)\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{5}$$

我们采用 LDC(linguistic data consortium)的 TDT 数据集对新闻话题给出的 11 个分类[***],对相同词性的词元和相同类别的话题的统计结果取平均值:

$$\chi^2_{avg}(P_k,R_m)=\frac{1}{|R_m|}\sum_{T\in R_m}\left(\frac{1}{|P_k|}\sum_{w\in P_k}p(w,T)\times\chi^2(w,T)\right),\ k=1...K,\ m=1...M \tag{6}$$

其中,K为词性种类数目(本文中为 12),M为新闻类别数目(本文中为 11).P_k代表第k种词性的词元集合,R_m代表第m种话题类别的话题集合,$p(w,T)$为词元w在话题T中出现的概率.表 2 给出了TDT2 数据集中部分词性种类与部分话题类别的统计结果(表中数据为对每个类别进行归一化处理之后的结果).

Table 2 Average correlation between term types and news classes on TDT2

表 2 TDT2 数据集中词元与话题相关度统计信息

	Location	Person	Date	Organization	Money	Percentage	NN	JJ	CD
Elections	0.37	1	0.04	0.58	0.08	0.03	0.32	0.13	0.1
Scandals/Hearings	0.66	0.62	0.28	1	0.11	0.02	0.27	0.13	0.05
Legal/Criminal cases	0.48	1	0.02	0.62	0.15	0	0.22	0.24	0.09
Natural disasters	1	0.27	0	0.04	0.04	0	0.25	0.04	0.02
Violence or war	1	0.36	0.02	0.14	0.02	0.04	0.21	0.11	0.02
Science and discovery	0.11	1	0.01	0.22	0.08	0.12	0.19	0.08	0.03
Finances	1	0.45	0.04	0.98	0.13	0.02	0.29	0.06	0.05
Sports	0.16	0.27	0.01	1	0.02	0	0.11	0.03	0.01

** http://opennlp.sourceforge.net/

*** http://projects.ldc.upenn.edu/TDT3/Guide/label.html

可以看出,对于不同类别的话题,不同种类的词性对于话题的区分起着不同程度的作用.其中,自然灾害(natural disasters)、军事冲突(violence or war)、金融(finances)这 3 个类别对于地名较为敏感;而选举(elections)、犯罪(legal/criminal cases)、科学发现(science and discovery)这 3 个类别对于人名较为敏感.同时还可以看到,绯闻/听证(scandals/hearings)类别对日期的敏感度较高;犯罪和金融类别对货币的敏感度较高;科学发现则对百分比的敏感度较高;而非命名实体则对于不同类别较为平均.从对表 2 数据的分析可知,对于不同类别的话题,不同词性的词元在计算相似度时应具有不同的权重.为此,我们对词元权重计算进行如下改进:

$$weight_T(d,t,w)=\frac{weight(d,t,w)\times\alpha_{type(w)}^{class(d)}}{\sum_{w'\in d} weight(d,t,w')\times\alpha_{type(w')}^{class(d)}} \tag{7}$$

其中,$type(w)$为词元 w 的词性,$class(d)$为新闻 d 所属类别,α_k^c 为对应于新闻类别 C 和词性 k 的加权参数.

基于针对不同类别新闻、不同词性词元的χ^2统计数据,本文考虑如下 5 种 α_k^c 加权参数设置方案:

(i) $\alpha_k^c=\chi^2_{avg}(k,C)$

(ii) $\alpha_k^c=\ln(\chi^2_{avg}(k,C))$,若$\chi^2_{avg}(k,C)<1$,则$\alpha_k^c$置0

(iii) $\alpha_k^c=(\chi^2_{avg}(k,C))^{1/2}$

(iv) $\alpha_k^c=(\chi^2_{avg}(k,C))^2$

(v) α_k^c=1(退化为基本模型)

其中,第(ii)种、第(iii)种方案不同程度地削弱了方案(i)的加权参数效果,方案(iv)则增强了加权参数效果.方案(v)退化为基本模型,便于效果比较.本文在第 5 节给出了 5 种参数设置方案在训练数据中的实验结果.

本文使用BoosTexter[20]预先对新闻按 11 种话题类别进行分类.BoosTexter是一种基于Boosting的机器学习算法,它从训练数据中学习出用于建立分类器的一系列简单规则.本文将最初始的*tf-idf*模型生成的词元权重作为分类的特征.使用TDT2 的 12 000 篇标注数据作为训练集,对TDT2 和TDT3 的所有数据进行分类.分类结果用于式(7)中词元权重调整的计算.

3.2 基于分布距离的动态词元权重调整

*tf-idf*模型原本是在信息检索领域所广泛使用的模型,其基本思想是,在整体文本集出现越少而在某些特定文本中出现相对集中的词元,相对于一个特定的查询区分文档的能力越强(与查询相关或不相关).而在TDT领域,我们需要确定的是相对于一个特定话题的区分文档的能力(属于此话题或不属于),而单纯采用新闻文档内词元的分布不能很好地描述话题的意义.因此从直觉上讲,用话题(表现为一个新闻簇)的向量模型与后续的新闻进行比较更为合乎情理.但是,已有实验证明,这种方法不能得到很好的效果[8,9].根据对数据的分析发现,这是因为一个新闻话题往往包含着很多互相直接或间接关联的事件,而这些事件大都有着各自不同的重点.因此,如何利用好话题信息,从而又可以避免话题内重点分散带来的问题是本节讨论的重点.

本文根据词频信息将词元分为 5 类:

A 类词:在整个文档集中频繁出现,如 year,people.这类词不具备区分新闻是否属于同一话题的能力,因此应给予很低的权重.

B 类词:在某个话题类别内集中并广泛地出现,这类词如 election,storm,具备区分属于不同话题类别新闻的能力,但是不具备区分同类新闻中两个不同话题的能力.如 election 和 storm 可以帮助区分一个新闻的类别(是描述 election 或者 storm),但却无法帮助区分两次不同的 election 或 storm.因此,此类词应给予较低权重.

C 类词:在一个话题内集中并频繁地出现,如一次空难的航班的名称、一次飓风的名称.此类词可以很好地进行话题间的区分,且出现得越频繁,越应当给予更高的权重.

D 类词:在一个话题内集中出现,但不广泛出现.如一次火灾中一个消防员的名字,可能只在两三个报道此消防员事迹的新闻中出现.此类词应给予较高权重,但因为在话题中出现得并不广泛,故权重也不宜过高.

E 类词:出现次数很少,且不集中在同一个话题中.此类词应给予很低的权重.

对于上文所提到的 5 类词,逐一分析 *tf-idf* 模型是否可以合理地给出相应的权重值.显然,A 类词的低权重可以在 *tf-idf* 模型中得到满足.B 类词的权重与其所在的话题类别的新闻数目有很强的关系,因此,*tf-idf* 不能总是很好地满足 B 类词的权重需求.对于 C 类词,如果在一个话题内出现得越多,*tf-idf* 模型会赋予越低的权重,这与上文对于 C 类词的权重需求不符.对于 D 类词,*tf-idf* 模型会给予很高的权重,这与 D 类词的权重需求相符.对于 E 类词,*tf-idf* 模型会给予很高的权重,这也与 E 类词的权重需求不符.可见,B,C 和 E 这 3 类词的权重需求与 *tf-idf* 模型不符,因此,本文提出一个修正模型.修正模型的新事件检测过程如图 1 所示.

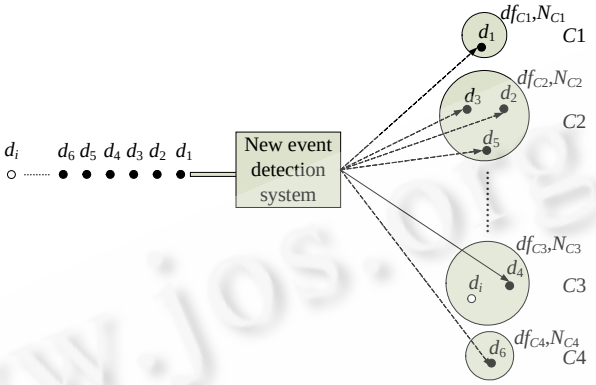

Fig.1 Modified NED procedure

图 1 修正的 NED 过程

在修正的新事件检测过程中,将当前新闻d_i(由白色圆圈代表)与前面所有新闻$d_1 \ldots d_{i-1}$(由黑色圆圈代表)一一比较.如图 1 所示,若新闻d_4与新闻d_i拥有最高相似度且高于阈值θ,则将d_i加入到d_4所在新闻簇$C3$,同时,更新$C3$ 所对应的文档频率df_{C3}与文档数目N_{C3};若最高相似度低于阈值θ,则建立一个新的新闻族并加入此新闻,例如新闻d_6与新闻簇$C4$.

我们利用词元在新闻簇内部分布与在整体文集分布的Kullback Leibler(KL)距离[21],动态调整词元权重:

$$weight_D(d,t,w)=\frac{weight(d,t,w)\times(1+\lambda\times KL(P_{cw}\,\|\,P_{tw}))}{\sum_{w'\in d} weight(d,t,w')\times(1+\lambda\times KL(P_{cw'}\,\|\,P_{tw'}))} \tag{8}$$

其中,

$$p_{cw}(y)=\frac{\mathrm{df}_c(w)}{N_c},\ p_{cw}(\bar{y})=1-\frac{\mathrm{df}_c(w)}{N_c} \tag{9}$$

$$p_{tw}(y)=\frac{\mathrm{df}_t(w)}{N_t},\ p_{tw}(\bar{y})=1-\frac{\mathrm{df}_t(w)}{N_t} \tag{10}$$

其中,$df_c(w)$为新闻d所在新闻簇C中包含词元w的新闻数,N_c为d所在新闻簇C包含的新闻数,N_t为到时间t为止的所有新闻数,λ为调和参数.其中,$df_c(w)$和N_c根据新闻簇的变化动态调整.

其基本思想为:在话题内出现的次数越多,而在话题外出现的次数越少的词应给予越高的权重.显然,公式(8)的权重模型可以兼顾所有 5 种类型词元的权重需求.

4 实验准备

4.1 数据集

本文使用LDC[22]提供的数据集TDT2[****]和TDT3[*****]作为实验数据.TDT2 数据集包含了从 1998 年 1 月~6

[****] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2001T57

[*****] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2001T58

月的大约 54 000 篇英文新闻,新闻来自ABC,Associated Press,CNN,New York Times,Public Radio International, Voice of America等媒体机构.TDT3 数据集包含从 1998 年 10 月~12 月的大约 31 000 篇英文新闻,新闻源在TDT2 的基础上增加了National Broadcasting Co.和MS-NBC.TDT2 数据集包含约 100 个人工标注的话题,约 12 000 条英文新闻属于标注话题中的至少一个.TDT3 数据包含约 120 个人工标注的话题,约 8 000 条英文新闻属于标注话题中的至少一个.所有话题被人工分为 11 个类别:(1) Elections;(2) Scandals/Hearings;(3) Legal/Criminal Cases;(4) Natural Disasters;(5) Accidents;(6) Ongoing Violence or War;(7) Science and Discovery News;(8) Finance;(9) New Law;(10) Sports News;(11) MISC. News.

4.2 实验设计

为了测试本文提出的改进模型的效果,实现并测试了如下 4 个系统:

SYSTEM-1:此系统为基线系统,采用第 2 节介绍的基本模型,即使用增量的 *tf-idf* 模型生成词元权重 $weight(d,t,w)$,作为 t 时刻词元 w 在新闻 d 中的权重.

SYSTEM-2:采用第 3.1 节提出的新闻描述方法,即对于不同类别的新闻赋予不同词性词元不同的权重,使用$weight_T(d,t,w)$作为t时刻词元w在新闻d中的权重.

SYSTEM-3:采用第 3.2 节提出的新闻描述方法及相应的新事件检测过程,根据词元在话题内部与整体文档集的分布距离动态更新词元的权重,使用$weight_D(d,t,w)$作为t时刻词元w在新闻d中的权重.

SYSTEM-4:综合采用第 3.1 节和第 3.2 节提出的改进模型,即首先对于不同类别的新闻赋予不同词性词元不同的权重,并在$weight_T(d,t,w)$基础上根据词元在话题内部与整体文档集的分布距离,动态更新词元的权重.

以上 4 个系统均在 TDT2 数据集上进行训练,并在 TDT2 和 TDT3 数据集上进行测试.

同时,为了便于比较,这里列出同类系统及简单介绍:

SYSTEM-5:在比较两篇新闻时计算 3 个相似度,分别对应命名实体、非命名实体、所有词元,并使用这 3 个相似度作为特征,利用支持向量机分类器判断新闻的“新”或“旧”[23].

SYSTEM-6:根据不同的新闻源建立不同的词频模型;对不同新闻的相似度进行归一化;对不同新闻源之间的相似度进行归一化;对新闻内容进行分割[13].

SYSTEM-7:根据新闻类别选择使用命名实体或非命名实体计算相似度,并去除类别内频繁词[7].

4.3 评价标准

TDT使用C_{Det}代价函数对结果进行评价[24]:

$$C_{Det}=C_{Miss}\times P_{Miss}\times P_{Tar}+C_{FA}\times P_{FA}\times P_{Nontar} \tag{11}$$

其中,C_{Miss}表示失报一个新事件新闻的代价,P_{Miss}表示对于一个新事件新闻失报的概率,P_{Tar}表示新事件新闻出现的概率;C_{FA}表示误报一个新闻为新事件新闻的代价,P_{FA}表示对于一个新闻误报为新事件新闻的概率,P_{Nontar}表示非新事件新闻出现的概率.一般使用标准化的代价函数作为最终评价标准:

$$Norm(C_{Det})=\frac{C_{Det}}{\min(C_{Miss}\times P_{Tar},C_{FA}\times P_{Nontar})} \tag{12}$$

对于一个完全判断正确的系统,$Norm(C_{Det})$为 0,将所有新闻都判断为新事件新闻或都判断为非新事件新闻中较好的一种情况$Norm(C_{Det})$为 1.

系统对于一个新闻是否为新事件新闻的判断包括两个部分:第 1 部分为“是”或“否”:第 2 部分为系统对一篇新闻为新事件新闻的确信度.确信度可用来描绘Detection Error Tradeoff(DET)曲线图,从而表现出误报率和失报率之间的关系,并可从图中得到最小标准化代价(min $Norm(C_{Det})$).最小标准化代价为不同阈值情况下所得到的标准化代价最小值.

5 实验结果及分析

表 3 给出了 SYSTEM-2 中 5 种加权参数设置方案在 TDT2 数据集中的最小标准化代价.方案(v)退化为

SYSTEM-1,方案(iv)由于对某些词性的词元给予过高的权重,从而导致代价大幅度升高.方案(ii)与方案(iii)的效果强于方案(v),但由于削弱了统计结果的作用,效果差于方案(i).方案(i)的代价低于其他所有 4 种方案.后续的实验中,SYSTEM-2 均使用方案(i)的加权参数设置方式.

Table 3 Min $Norm(C_{Det})$ of five α_k^c setting methods in SYSTEM-2 on TDT2

表 3 SYSTEM-2 中 5 种 α_k^c 设置方案在 TDT2 数据集上的最小标准化代价

α_k^c	$\chi^2_{avg}(k,C)$	$\ln(\chi^2_{avg}(k,C))$	$(\chi^2_{avg}(k,C))^{1/2}$	$(\chi^2_{avg}(k,C))^2$	1
Min $Norm(C_{Det})$	0.509 7	0.569 2	0.541 6	0.785 8	0.588 4

表 4 给出了式(8)中λ参数从 0~5 取值范围下 SYSTEM-3 在 TDT2 数据集中的最小标准化代价.可以看出,当λ取值为 3 时可以达到最好的效果.如果λ取值过大,则反而会降低性能.根据对实验结果的分析发现,这主要是由于两个原因造成的:(a) 由于用于计算词元分布距离的话题是动态自动构成的,因此并非完全准确,所以导致有些词元的分布距离存在噪声;(b) 对于话题内外分布相同或相近的词元,如果λ过大,则权重会趋近于 0,从而出现对词元权重低估的现象.因此,λ也不宜过大,使式(8)中的常数 1 可以起到润滑的作用.后续的实验中,均设定 SYSTEM-3 中λ参数为 3.

Table 4 Min $Norm(C_{Det})$ using different λ values in SYSTEM-3 on TDT2

表 4 SYSTEM-3 中λ参数不同取值范围下在 TDT2 数据集上的最小标准化代价

λ	0	1	2	3	4	5
Min $Norm(C_{Det})$	0.588 4	0.561 2	0.537 9	0.521 7	0.568 2	0.635 5

表 5 给出了在 TDT2 数据集上本文实现的一个系统与 SYSTEM-5 的 NED 结果.由于没有数据可以作为 TDT2 的训练数据用以得到最佳阈值θ,所以本文只提供最小标准化代价.图 2 给出了 4 个系统在 TDT2 数据集上的 DET 曲线.由于 TDT2 同时为训练集和测试集,在 TDT2 上 SYSTEM-2 和 SYSTEM-4 的结果较好,基于词性的加权与 TDT2 数据更加吻合.

Table 5 C_{Det} results on TDT2 and TDT3

表 5 在 TDT2 和 TDT3 数据集上的代价结果

Systems	TDT2		TDT3	
	$Norm(C_{Det})$	Min $Norm(C_{Det})$	$Norm(C_{Det})$	Min $Norm(C_{Det})$
SYSTEM-1	–	0.588 4	0.602 9	0.570 9
SYSTEM-2	–	0.509 7	0.535 1	0.518 3
SYSTEM-3	–	0.521 7	0.523 4	0.486 0
SYSTEM-4	–	0.468 2	0.499 3	0.476 5
SYSTEM-5	–	0.530 0	–	–
SYSTEM-6	–	–	–	0.578 3
SYSTEM-7	–	–	–	0.522 9

表 5 右半部分给出了TDT3 数据集上的标准化代价(由TDT2 的最佳阈值得到)与最小标准化代价.表中数据可得到如下结论:(1) SYSTEM-2 的最小标准化代价比基线系统降低了 0.052 6,说明对不同类别新闻进行不同的词性加权可以有效改善效果;(2) SYSTEM-3 比基线系统降低了 0.084 9,说明用KL距离方法可以有效发现话题中的关键词元,从而改善效果;(3) 两种改进方法的组合SYSTEM-4 可以达到最好的效果,其最小标准化代价比基线系统降低了 0.094 4.但SYSTEM-4 只比SYSTEM-3 改善了 0.009 5,说明两种改进方法发现的话题关键词元存在一定的重合;(4) SYSTEM-4 比其他现有系统[7,13,25]的最好结果SYSTEM-7 代价降低了 0.046 4,说明本文提出的模型改进效果显著.图 3 给出了 4 个系统在TDT3 数据集上的DET曲线,其中,SYSTEM-4 的最小标准化代价点在误报率 0.022 4 和失报率 0.366 7.SYSTEM-3 与SYSTEM-4 在低失报率部分可以得到更低的误报率,是因为权重偏向话题内部的关键词元,因此,属于不同话题的新闻的相似度大为降低,因为它们所共有的词元大都不是话题内部的关键词元.

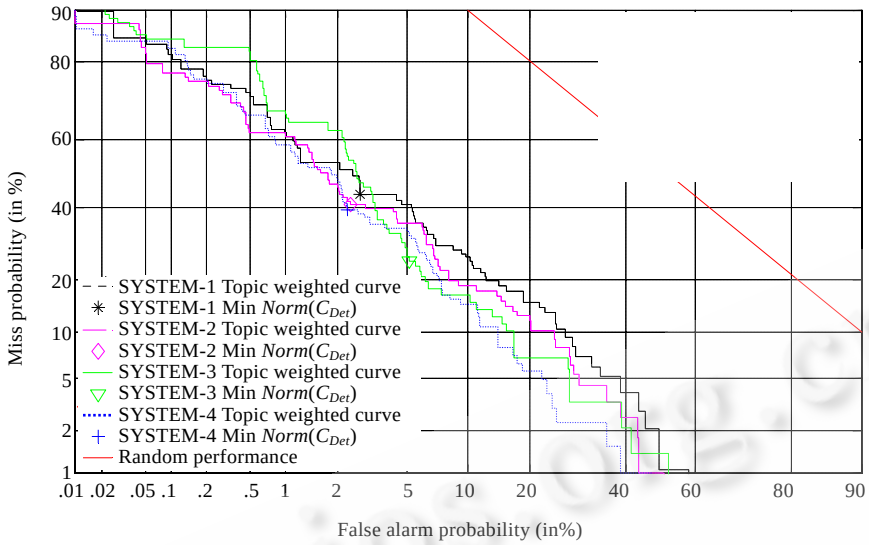

Fig.2 DET curves on TDT2

图 2 TDT2 数据集上的 DET 曲线图

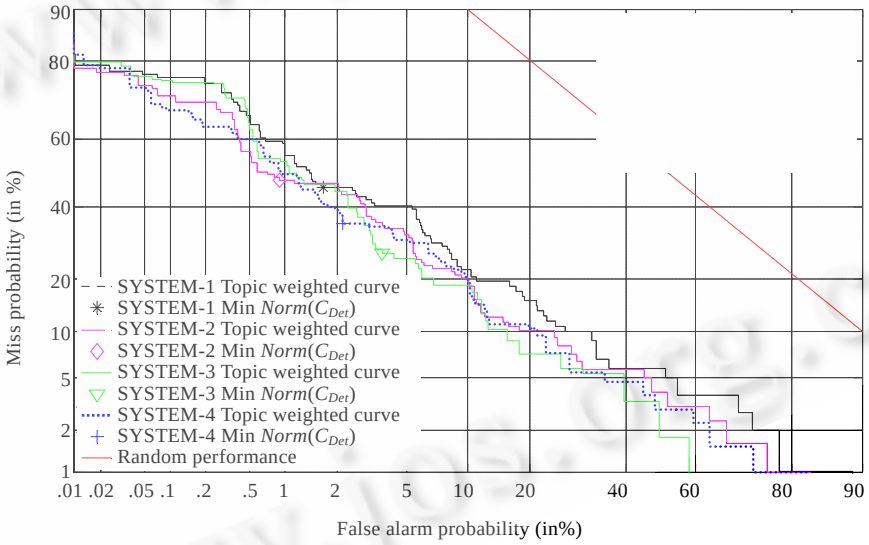

Fig.3 DET curves on TDT3

图 3 TDT3 数据集上的 DET 曲线图

为了从直观上看到各个系统之间的区别,我们使用相似度矩阵来展示各个系统的效果,并使用 MATLAB 绘制矩阵的等高线图.相似度矩阵定义如下:

$$S=\begin{bmatrix} s_{N1} & \cdots & s_{Ni} & \cdots & s_{NN} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ s_{i1} & \cdots & s_{ii} & \cdots & s_{iN} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ s_{11} & \cdots & s_{1i} & \cdots & s_{1N} \end{bmatrix}=\begin{bmatrix} S^{M1} & \cdots & S^{Mj} & \cdots & S^{MM} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ S^{j1} & \cdots & S^{jj} & \cdots & S^{jM} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ S^{11} & \cdots & S^{1j} & \cdots & S^{1M} \end{bmatrix} \tag{13}$$

其中,N为新闻数量,M为话题数量.新闻$\{d_1,d_2,\ldots,d_N\}$首先按话题排序,在话题内部按时间排序.s_{ik}为新闻d_i与d_k的相似度,S^{jh}为话题j与话题h中新闻的相似度子矩阵.限于篇幅,下面只给出在TDT3测试集中5个话题类别的相似

度等高线图,如图4~图8所示.其中,(a),(b),(c),(d)分别为SYSTEM-1,SYSTEM-2,SYSTEM-3,SYSTEM-4的相似度矩阵等高线图.图中的直线为不同话题之间的分隔线.在对角线上的矩形表示话题内部的新闻相似度子矩阵,在对角线外的矩形表示不同话题之间的新闻相似度子矩阵.可以看出,SYSTEM-4 在所有类别中都有最好的效果,表现在不同话题之间的相似度更低,而话题内部的相似度更高.其中,Finance类与Sports News类的效果更加明显;SYSTEM-2 在除Ongoing Violence or War与Finance之外的类别也取得了较好于SYSTEM-1 的效果;SYSTEM-3 在所有类别中取得了好于SYSTEM-1 或相当的效果.

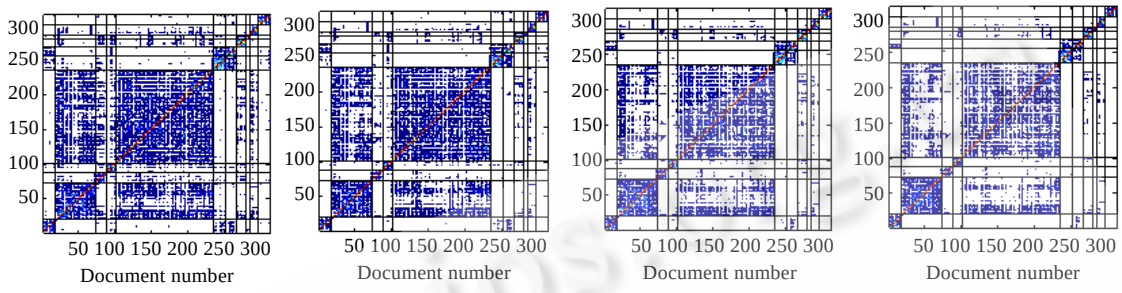

Fig.4 Elections news similarity matrix figure

图 4 Elections 新闻类的相似度矩阵图

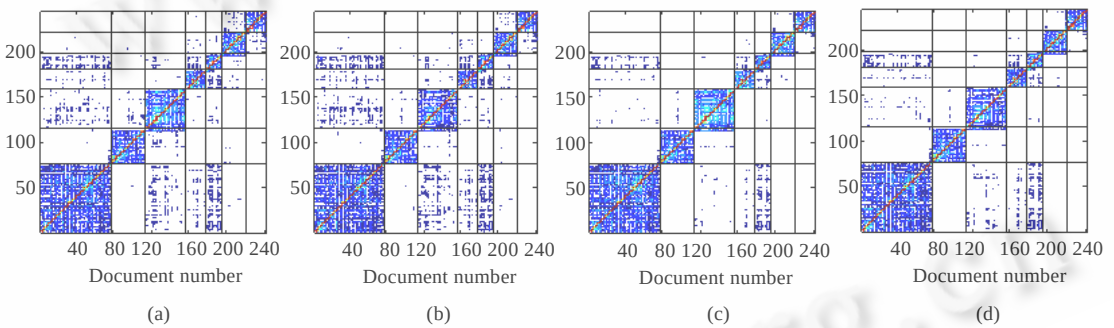

Fig.5 Ongoing Violence or War news similarity matrix figure

图 5 Ongoing Violence or War 新闻类的相似度矩阵图

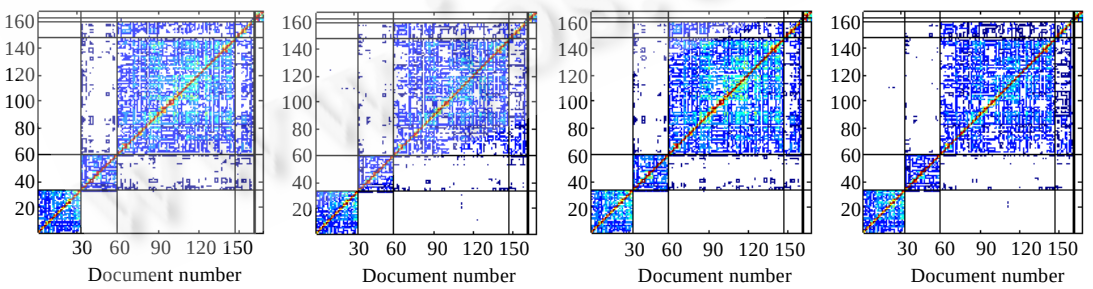

Fig.6 Science and Discovery news similarity matrix figure

图 6 Science and Discovery 新闻类的相似度矩阵图

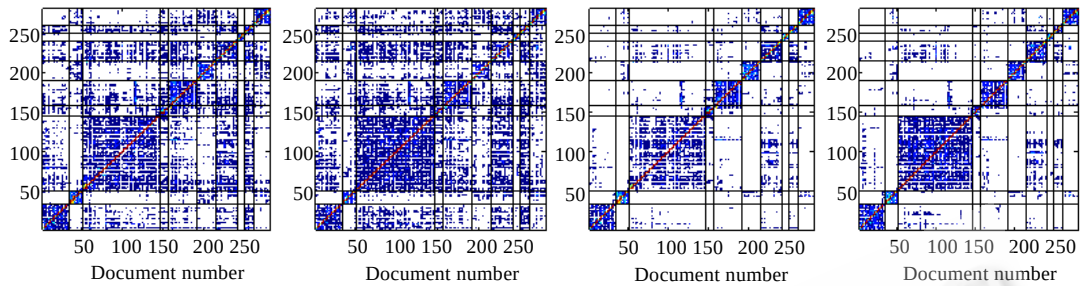

Fig.7 Finance news similarity matrix figure

图 7 Finance 新闻类的相似度矩阵图

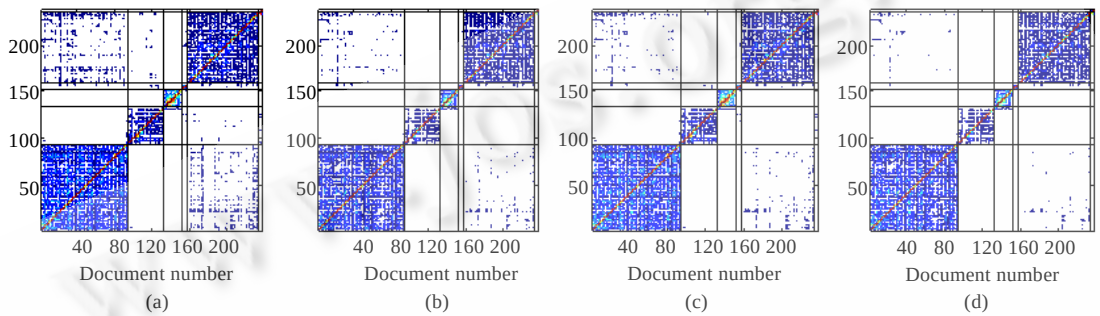

Fig.8 Sports news similarity matrix figure

图 8 Sports 新闻类的相似度矩阵图

6 结 语

本文对 NED 问题提出了一种改进模型.该模型从训练数据中得出不同词性的词元对于不同类别的新闻在区分是否属于同一话题中发挥作用的大小,并根据统计结果改进词元权重.之后,根据已处理的新闻及话题信息得出词元对于某些话题的重要程度,从而动态调整词元权重.这种方法既避免了在新闻间进行比较的方法未有效利用话题信息的问题,同时也避免了在新闻与话题之间进行比较的方法中话题重点分散带来的问题.实验表明,在 TDT2 和 TDT3 数据集上,本文提出的模型可以使结果得到显著的改善;在 TDT3 数据集上,最小标准化代价较之同类系统的最好结果降低了 0.046 4.

由于TDT数据集时间跨度较短,因此,本文未涉及到对新闻时间信息的应用.下一步工作需要从网络上收集时间跨度更长的新闻数据,研究如何利用时间信息辅助新事件发现.此外,提高 NED 系统的效率也是一个重要的课题,大多数系统都将新报道与之前的所有新闻进行比较,而这在实际应用中是不可接受的.

References:

[1] http://www.nist.gov/speech/tests/tdt/. 2003.

[2] Allan J. Topic Detection and Tracking: Event-Based Information Organization. Kluwer Academic Publishers, 2002.

[3] Yang Y, Carbonell J, Brown R, Pierce T, Archibald BT, Liu X. Learning approaches for detecting and tracking news events. IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval, 1999,14(4):32−43.

[4] Yang Y, Zhang J, Carbonell J, Jin C. Topic-Conditioned novelty detection. In: Proc. of the 8th ACM SIGKDD Int'l Conf. ACM Press, 2002. 688−693. http://www.stat.purdue.edu/~jianzhan/papers/sigkdd02.pdf

[5] Juha M, Helena AM, Marko S. Applying semantic classes in event detection and tracking. In: Sangal R, Bendre SM, eds. Proc. of the Int'l Conf. on Natural Language Processing (ICON 2002). 2002. 175−183. http://www.cs.helsinki.fi/u/jamakkon/papers/icon02.pdf

[6] Juha M, Helena AM, Marko S. Simple semantics in topic detection and tracking. Information Retrieval, 2004,7(3-4):347−368.

[7] Giridhar K, Allan J. Text classification and named entities for new event detection. In: Jarvelin K, Allan J, Bruza P, Sanderson M, eds. Proc. of the 27th Annual Int'l ACM SIGIR Conf. New York: ACM Press, 2004. 297−304.

[8] Yang Y, Pierce T, Carbonell J. A study on retrospective and on-line event detection. In: Croft WB, Moffat A, van Rijsbergen CJ, Wilkinson R, Zobel J, eds. Proc. of the SIGIR'98. Melbourne, 1998. 28−36.

[9] Allan J, Lavrenko V, Malin D, Swan R. Detections, bounds, and timelines: Umass and tdt-3. In: Proc. of the Topic Detection and Tracking Workshop (TDT-3). Vienna, 2000. 167−174. http://ciir.cs.umass.edu/pubfiles/ir-201.pdf

[10] Papka R, Allan J. On-Line new event detection using single pass clustering. Technical Report, UM-CS-1998-021, University of Massachusetts, 1998.

[11] Lam W, Meng H, Wong K, Yen J. Using contextual analysis for news event detection. Int'l Journal on Intelligent Systems, 2001,16(4):525−546.

[12] Nicola S, Joe C. Combining semantic and syntactic document classifiers to improve first story detection. In: Kraft DH, Croft WB, Harper DJ, Zobel J, eds. Proc. of the 24th Annual Int'l ACM SIGIR Conf. New York: ACM Press, 2001. 424−425.

[13] Thorsten B, Francine C, Ayman F. A system for new event detection. In: Clarke C, Cormack G, Callan J, Hawking D, Smeaton A, eds. Proc. of the 26th Annual Int'l ACM SIGIR Conf. New York: ACM Press, 2003. 330−337.

[14] Jia ZY, He Q, Zhang HJ, Li JY, Shi ZZ. A news event detection and tracking algorithm based on dynamic evolution model. Journal of Computer Research and Development, 2004,41(7):1273−1280 (in Chinese with English abstract).

[15] Wu PB, Chen QX, Ma L. Research on extraction and integration of developing event based on analysis of space-time information. Journal of Chinese Information Processing, 2005,20(1):21−28 (in Chinese with English abstract).

[16] Lei Z, Wu LD, Lei L, Huang YY. Incremental *K*-means method based on initialization of cluster centers and its application in news event detection. Journal of the China Society for Scientific and Technical Information, 2006,25(3):289−295 (in Chinese with English abstract).

[17] Callan JP, Croft WB, Harding SM. The INQUERY retrieval system. In: Proc. of the DEXA'92, the 3rd Int'l Conf. on Database and Expert Systems Applications. 1992. 78−83. http://www.cs.cmu.edu/~callan/Papers/callancroftdexa92.ps.gz

[18] Krovetz R. Viewing morphology as an inference process. In: Korfhage R, Rasmussen E, Willett P, eds. Proc. of the ACM SIGIR'93. New York: ACM Press, 1993. 61−81.

[19] Yang Y, Pedersen J. A comparative study on feature selection in text categorization. In: Fisher DH, ed. Proc. of the 14th Int'l Conf. on Machine Learning (ICML'97). San Francisco: Morgan Kaufmann Publishers, 1997. 412−420.

[20] Schapire RE, Singer Y. Boostexter: A boosting-based system for text categorization. Machine Learning, 2000,39(2/3):35−168.

[21] Cover TM, Thomas JA. Elements of Information Theory. New York: Wiley, 1991.

[22] The linguistic data consortium. http://www.ldc.upenn.edu/

[23] Giridhar K, Allan J. Using names and topics for new event detection. In: Mooney RJ, ed. Proc. of the Human Technology Conf. and Conf. on Empirical Methods in Natural Language. Vancouver, 2005. 121−128.

[24] The 2001 TDT task definition and evaluation plan. 2001. http://www.nist.gov/speech/tests/tdt/tdt2001/evalplan.htm

[25] TDT 2001 evaluations. 2001. http://www.nist.gov/speech/tests/tdt/tdt2001/index.htm

附中文参考文献:

[14] 贾自艳,何清,张海俊,李嘉佑,史忠植.一种基于动态进化模型的事件探测和追踪算法.计算机研究与发展,2004,41(7):1273−1280.

[15] 吴平博,陈群秀,马亮.基于时空分析的线索性事件的抽取与集成系统研究.中文信息学报,2005,20(1):21−28.

[16] 雷震,吴玲达,雷蕾,黄炎焱.初始化类中心的增量 *K* 均值法及其在新闻事件探测中的应用.情报学报,2006,25(3):289−295.

张阔(1981－),男,北京人,博士生,主要研究领域为文本挖掘,信息抽取,信息检索.

吴刚(1978－),男,博士生,主要研究领域为数据仓库,半结构化数据与Web数据集成,数据挖掘.

李涓子(1964－),女,博士,副教授,CCF 高级会员,主要研究领域为语义网,中文信息处理,网络环境下的知识发现和知识管理.

王克宏(1942－),男,教授,博士生导师,CCF高级会员,主要研究领域为知识工程,分布式知识处理.