DOI:10.3969/j.issn.1673-4785.201111014

基于文本的新闻事件多版本发现模型

肖融,孔亮,张岩
(北京大学 教育部机器感知重点实验室,北京 100871)

摘 要:信息时代的发展让越来越多的新闻事件充斥人们的生活,对于一件特定的新闻事件,目前已有很多算法可以帮助人们进行事件追踪和发现.提出一种CDW算法,帮助读者对于一件具有多个版本描述的新闻事件进行多个不同版本的发现.这个算法将文档集映射到话题层,通过提取每个话题的流行词,以得到文档集中具有高区分度的特征.然后根据这些特征对文档集进行聚类,最后得到事件的多个版本.通过在2个实际数据集上进行实验,实验结果表明,该算法与以往的相关算法相比是十分有效的.

关键词:多版本事件;高区分度;聚类模型;话题分析
中图分类号:TP18 **文献标志码**:A **文章编号**:1673-4785(2012)04-0307-08

A text clustering model for diverse versions discovery

XIAO Rong, KONG Liang, ZHANG Yan
(Key Laboratory on Machine Perception of MOE, Peking University, Beijing 100871, China)

Abstract: The development of information technology brings numerous news and events to our daily life. Although previous researches have provided various algorithms to detect and track events, few of them focus on uncovering the diversified versions of an event. In this paper, a novel algorithm CDW which is capable of discovering different versions of one event according to the news reports was proposed. First, documents were mapped to the topic layer to get the information of each topic. Then the highly-differentiated words of each topic were extracted to cluster the documents. At last, various versions of one event were got. Experiments conducted on two data sets show that the algorithm given in this paper is effective and outperforms various related algorithms, including classical methods such as K-means and linear discriminant analysis (LDA).

Keywords: diverse versions discovery; highly-differentiated words; clustering model; topic analysis

人们生活在信息时代,每天都在接收大量的信息,从各种媒体渠道浏览各种新闻事件.有些新闻事件只有基于事件本身的客观性报道,比如《哈利波特7》上映,莫斯科发生大规模球迷骚乱,欧盟呼吁欧洲共同应对危机等.这类新闻报道主要是对所发生的新闻事实进行客观描述,一般所有的报道都相近似,不会众说纷纭.而有一些新闻事件由于具有开放性或者模糊性,导致各方面口径不一,就会出现所谓的"罗生门"现象.比如流行天王Michael Jackson的死因,有报道说是心脏病意外死亡,有报道称是自杀,有报道称是私人医生误杀或谋杀等.再比如对于韩国天安舰沉船事件,有报道称是朝鲜所为,有报道称是美国的阴谋,还有报道说是南北交火时沉没等.这一类新闻事件的众多报道就会出现对于同一事件有多个不同版本说法的现象,也就是本文所研究的多版本事件.

随着话题发现与追踪技术(topic detection and tracking, TDT)[1-2]的发展日益成熟,很多网站都可以提供为用户组织归纳新闻事件的应用.通过话题发现与追踪,用户可以清楚地知道新闻事件的发生和衍化过程,也可以看到关于事件的各种报道.TDT源于1996年美国国防高级研究计划委员会提出的需要一种能自动确定新闻报道流中话题结构的技术[3-5].随后,DARPA、卡内基·梅隆大学、Dragon系统公司以及马萨诸塞大学的研究人员定义了TDT的相关内容,并检验信息检

收稿日期:2011-11-24.
基金项目:国家自然科学基金资助项目(61703081).
通信作者:肖融. E-mail:xrsmile@gmail.com.

索中基于主题的技术在TDT中的应用情况,这些研究及评测被命名为TDT pilot[6]. TDT是一项综合的技术,需要较多的自然语言处理理论和技术作为支撑. 话题发现技术可以看作是一种按事件的聚类,研究者常采用的算法有agglomerative聚类、增量K-means聚类、增量聚类等. 话题追踪的常用技术有Rocchio分类方法、决策树方法、基于HMM的语言模型等[7-10].

然而,对于多版本的新闻事件,简单的组织归纳难以满足用户对于不同版本报道的信息获取的需求. 对于存在多个版本的事件,读者很难面对庞大的新闻数据而自行鉴别事件的版本,如果存在一个算法可以为读者找出一共存在多少个版本,每一个版本的描述是什么,那么对于读者获取相关新闻信息将会十分有用.

遗憾的是,目前关于事件多版本发现的研究很少,没有太多有价值的相关文献. 对于多版本发现最直接的考虑就是进行简单的聚类分析. 聚类是数据挖掘技术中极为重要的组成部分,是在事先不规定分组规则的情况下,将数据按照其自身特征划分成不同的簇(cluster),不同簇的数据之间差距越大、越明显越好,而每个簇内部的数据之间要尽量相似,差距越小越好. 常见的聚类算法有K-means算法、Birch算法、Dbscan算法、Clique算法、神经网络方法等[11-14]. 但是单纯的聚类方法具有很多局限性. 由于对于同一事件的新闻报道在内容主体上通常具有高度的相似性,简单的聚类方法无法将其中不同的"声音"有效地区分开来. 文献[15]提出了一种基于图模型的事件多版本发现算法. 该算法是基于语义的迭代算法,通过提取流行词并将之过滤来降低同一特定事件的文档之间的紧密联系性. 然后构建词图以发现词与词之间的层次关系. 根据社区发现算法[16],构建虚拟文档来表示每一版本的中心. 最后根据Rocchio分类算法[17]来进行多版本的分类.

尽管本文在内容上借鉴上述一些前人的工作,但无论从算法思想还是效果上都有很大创新. 一方面,它提出了话题层的概念,建立了文档集与话题层的映射关系,利用LDA将文档集合引申到话题空间,然后对每一话题进行特征提取. 另一方面,它提出了一种有效的提取高区分度特征的方法. 该方法过滤掉了文本集之间相似性的部分,有效地提取出文档集之间的差异性特征,从而提高多版本发现的效率和准确度.

1 基本定义

关于事件的多版本发现,这里首先要讨论的就是一个有效的事件多版本发现算法需要具备的性质. 并且,为了使多版本发现的工作更有意义,本文认为这样的算法必须是足够强健的(qualitatively strong)[18].

首先,这里先要声明几个符号表示的意义. 令$\boldsymbol{D}=\{\boldsymbol{d}_i,\boldsymbol{d}_{i+1},\cdots,\boldsymbol{d}_n\}$表示对于某一特定新闻事件所搜集的$n$个文档的集合,其中每一篇文档$\boldsymbol{d}_i$, $i=1:n$,用bag of words表示$(w_1,w_2,\cdots,w_d)$. 多版本发现的目标在于发现m个不同的版本$\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_m\}$来描述一个事件,其中每一种版本$\boldsymbol{v}_i$($i=1:m$),用一种词的分布表示. 对于某一事件的多版本发现也就是找到关于这一事件的不同方面、不同说法或不同观点等,让用户能够一目了然地看到这一事件的不同角度和层面.

为了得到有效的多个版本描述,一个关于新闻事件的多版本发现算法需要满足以下3个特性:

1)多样性. 即给定一个文档集合$\boldsymbol{D}$作为输入,多版本发现算法需要在不改变相似度函数的情况下,找到$m(m>1)$个不同的版本$\boldsymbol{v}_i$, $i=1:m$. 也就是说算法不依赖于相似度函数的形式.

2)区别性. 得到的每一个版本$\boldsymbol{v}_i(i=1:m)$应该是显著不同的. 这里指的是任意2个版本之间应该具有高度不相似度.

3)高质性. 得到的每一个版本$\boldsymbol{v}_i(i=1:m)$应该是关于相似度函数表现强健的(qualitatively strong).

可以证明,本文介绍的多版本发现算法满足以上提到的多样性、区别性和高质性.

2 CDW:基于文本的事件多版本发现模型

2.1 CDW算法框架

CDW(clustering by highly-differentiated words)对于事件的多版本发现,最朴素、最直接的做法就是对文档进行简单的聚类. 然而,由于大数据集文档间存在复杂的语义关联和高度的相似性,仅仅简单的聚类方法无法得到区分度高的版本类别,"区别性"方面的表现很差. 为了解决这个问题,本文提出的CDW事件多版本发现算法将整个问题分为2部分:首先,需要找到具有高区分度的特征;其次,将文档进行特征向量化,并且进行文本聚类. 进一步具体分析,本文算法可以分为以下3步:

1)寻找区分度高的特征. 每一篇文档都可以被表示成a bag of words,对于大数据集来说,不经筛选无疑会造成维数灾难. 为了得到更利于区分文档的特征并去掉干扰噪音,本算法将文档集引申到话题层,

通过运用词频过滤和提取 popular words 等方法,对特征进行筛选和降维,最终得到区分度较高的特征.

2)特征向量化,构建处理后的文档. 需要将所有文档用经过筛选的高区分度特征进行向量化表示. 这里,本算法利用的是 TF-IDF 加权技术[19].

3)文本聚类. 经过特征筛选和特征表示,已经得到了经过处理的文档特征向量. 然后,用 K-means 方法[20]进行文本聚类,得到最终的多版本.

图 1 展示了 CDW 算法的流程框架. 下面将对这 3 步做进一步具体说明.

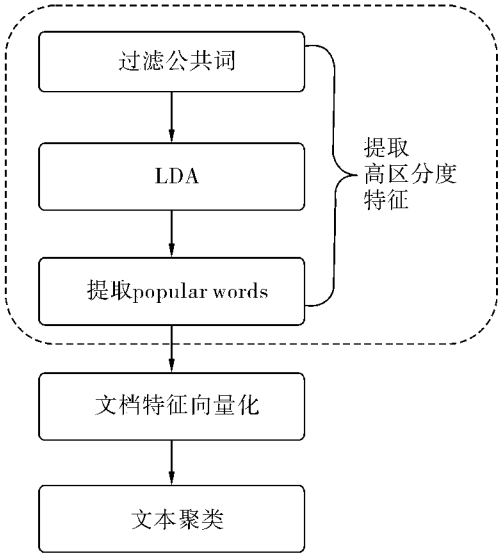


图 1　CDW 算法框架

Fig. 1　The framework of CDW algorithm

2.2　高区分度特征的生成策略

词汇是文档最基础的组成单元,也是最常用的特征表示. 然而,如果将一篇文档包含的所有词语都作为这篇文档的特征,那么对于大数据集来说可能会造成维数灾难. 所以,必须提取出对于区分文档版本最有效的词语,以进行降维.

2.2.1　根据词频过滤公共词

词频过滤是进行特征筛选时最基础的手段. 经过分析可知,对于同一事件的文档集中频率较高的词通常是描述客观事件本身的词,并不具有版本信息. 所以,本算法首先统计数据集中的每一个词出现在文档中的数目作为这个词的频率. 这里设定一个阈值 d,将频率高于阈值 d 的词全部筛掉. 另外也过滤掉频率为 1 的词.

2.2.2　基于 LDA 分析的话题映射与聚类

本算法利用 LDA 分析[21-23]对词频过滤后的特征集进行进一步特征提取,这样做的目的是进一步降维,并且将文档集映射到不同的话题. 然后,对文档集进行聚类. 这样,就得到了文档的初步类别信息.

方法步骤如下[24]:

1)初始化. 确定 K 的值,以及抽样数目和参数. 进行 LDA 分析,得到词-话题矩阵.

2)矩阵每一行为一个词的特征向量,对其进行聚类.

首先,利用 LDA 分析将文档集映射到话题层.

给定一个文档集合 D,每个文档 $\boldsymbol{d}$ 包含一个词序列$\{w_1, w_2, \cdots, w_n\}$. 在集合 D 对应的 LDA 模型中,首先假设话题数目固定为 K,然后经过 LDA 分析得到每个文档属于每个话题的概率.

然后进行文本聚类. 在 LDA 分析后,获得一个词-话题的矩阵,每一行是词在文本上的分布,每行有 K 维. 之后把词的特征向量进行聚类,根据最大隶属原则,将每篇文章划为概率最大的话题. 这样就完成了文档集到话题层的映射.

2.2.3　popular words 的提取算法

通过之前 2 步对特征集的降维和提取,至此已经将文档集进行了话题的映射并且得到了初步的聚类结果. 现在,需要对每一类进行提取 popular words,从而得到每一类最具代表性的词. 这里所用到的提取方法是基于文献[25]中提到的关键词提取方法,并加以改进运用. 本文认为,并往往在重要的句子中. 有代表性的词往往和其他有代表性的词共同出现. 并且,句子和词能够根据他们的连接结构计算排名. 所以,首先计算句子排名,找到重要的句子集,从而减少句子的影响. 构建句子连接关系图 G_S,句子 s_i 和 s_j 边的权值 $\mathrm{IF}(s_i, s_j)$ 定义如下:

$$\mathrm{IF}(s_i, s_j) = \frac{\max Co(s_i, s_j)}{\mathrm{Length}(s_j)}$$

式中:$\max Co(s_i, s_j)$ 表示 s_i 和 s_j 之间相同词的个数,$\mathrm{Length}(s_j)$ 表示 s_j 的长度. 然后构造邻接矩阵 $\boldsymbol{M}_S$,利用 PageRank[26]的思想,对 $\boldsymbol{M}_S$ 进行迭代计算得到每一个 $\mathrm{SRank}(s_i)$,其代表句子 i 的重要程度.

下一步根据句子的重要程度计算词的重要程度. 其基本思想与句子的计算和排名近似. 同样建立词链接关系无向图 G_W,词 i 与词 j 之间边的权值定义如下:

$$\mathrm{support}(i, j) = \sum_{i,j \in p} \mathrm{SRank}(s_p).$$

式中:p 代表句子 s_p 中的词集,$\mathrm{SRank}(s_p)$ 代表 s_p 的重要程度. 然后利用 PageRank 算法思想进行排名,得到每一个词的 $\mathrm{WRank}(w_i)$. 根据 $\mathrm{WRank}(w_i)$ 的大小,排名靠前的词为 popular words.

最后将每一类得到的 popular words 合并去重到

一个集合中,作为最终得到的具有高区分度的特征集合.

2.3 特征向量化

在这一步中,需要将原始数据集用得到的高区分度特征词进行表示. 对文档进行向量化最常用的方法就是计算每个词的 TF-IDF 权值,作为这一特征的特征值. TF-IDF[27-28] 实际上是 TF * IDF,TF 为词频,IDF 为反文档频率. 计算公式如下:

$$\mathrm{TF}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}},$$

$$\mathrm{IDF}_i = \log \frac{|D|}{|\{d: d \ni t_i\}|}.$$

最后,$\mathrm{TF\text{-}IDF}_{i,j} = \mathrm{TF}_{i,j} i \times \mathrm{IDF}_i$.

2.4 文本聚类

最后,对处理后的数据集进行文本聚类. 这里用的是 K-means 聚类算法[29-30]. K-均值聚类(K-means clustering)是 MacQueen 提出的一种非监督实时聚类算法,在最小化误差函数的基础上将数据划分为预定的类别数 K. 设定类别数目 K,然后将数据对象划分为 K 个聚类以便使所获得的聚类满足:同一聚类中的对象相似度较高;不同聚类中的对象相似度较小.

2.5 文档集与话题层的映射关系

图 2 展示了 CDW 事件多版本算法中文档集与话题层的映射关系以及整个算法的流程.

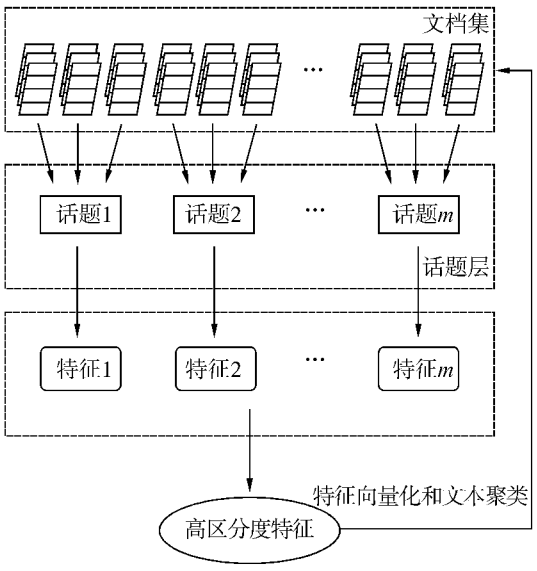


图 2 文档与话题层的映射关系

Fig. 2 Mapping between documents and topics

通过图示和之前的算法介绍,总结如下:

1)建立文档集与话题层之间的映射关系,将文档映射到不同的话题中;

2)提取每一个话题的特征;

3)合并所有话题的特征,过滤掉公共部分,找到具有特性的特征词项;

4)将原始文档集用提取出的特征表示,进行聚类.

最终得到不同版本的文档集.

3 实验与评价

3.1 实验数据集

为了展示 CDW 算法的有效性,作者在 2 个真实的数据集上进行了实验. 一个是韩国的天安舰沉没事件,包括 533 篇文档,分别来自英国广播 BBC、英国天空广播、美国之音、美国纽约时报、朝日新闻、朝鲜日报等,以下简称为 CS. 另一个是台湾连胜文枪击案,包括 391 篇文档,分别来自腾讯、雅虎、新浪、搜狐、人民网、凤凰网等,以下简称为 LSW.

韩国天安舰事件发生于 2010 年 3 月 26 日,韩国军方称其一艘导弹护卫舰“天安舰”因发生不明原因的爆炸事故而沉没. 由于确切的原因一直无法调查清楚,所以关于此次沉没事件的原因引发了很大争议. 类似地,台湾连胜文枪击案发生于 2010 年 11 月 26 日,当时正值台湾 5 市选举,连战的儿子连胜文在助选时头部遭到枪击. 由于正值政治敏感时期,关于此次枪击案凶手的动机就成了一大疑点.

表 1 CS 和 LSW 数据集说明

Table 1 Illustration of data sets CS and LSW

数据集	文档数	词条数	过滤后的词数	高区分度特征数
CS	533	9 842	6 749	879
LSW	391	7 477	4 952	650

数据集 CS 中,经过去停用词和词根还原后的词条一共有9 842 个,利用词频过滤后有6 749 个,最后提取到的高区分度特征词为 879 个. 数据集 LSW 中,经过去停用词和词根还原后的词条一共有 7 477 个,利用词频过滤后有4 952 个,最后提取到的高区分度特征词为 650 个.

3.2 评估方法

对于一个事件的新闻报道,很难通过逐篇浏览来确定每一篇报道属于哪一个版本. 所以,本文采用一个逐对判别的方法来评估 CDW 算法的效用.

在逐对判别方法中,这里关注的是某一对文档是否属于同一版本. 首先,需要构建标准测试集. 作者从 CS 数据集中随机选取了 200 对文档,从 LSW 数据集中随机选取了 150 对文档,并且确保每一对文档都不同. 然后,把每一对文档给志愿者浏览,让他们投票决定每一对文档是否属于同一版本. 如果

某一对文档很难判别是否同类,则直接将这一对文档剔除,并且添加一对新的文档到测试集中. 形式化定义如下:

$$T_\varepsilon = \{\langle\langle d_{i1}^\varepsilon, d_{i2}^\varepsilon\rangle, v_i\rangle \mid v_i \in \{0,1\},$$
$$d_{i1}^\varepsilon \in D^\varepsilon, d_{i2}^\varepsilon \in D^\varepsilon\}.$$

式中:$v_i=1$ 表示文档对 d_{i1}^ε 和 d_{i2}^ε 属于同一版本,$v_i=0$ 表示其他情况. 特别地,这里将为 CS 数据集和 LSW 数据集构建的测试集简记为 T_{CS} 和 T_{LSW}.

在这个评估方法中,用准确率来测量文档对测试. 给定一个文档对$\langle d_{i1}^\varepsilon, d_{i2}^\varepsilon\rangle$,每一个多版本发现算法都会给出一个判断 v_i'. 因此,定义文档对测试的准确率为 P_{score},即:

$$P_{\mathrm{score}} = \frac{\sum_{\langle d_{i1}, d_{i2}\rangle} v_i \odot v_i'}{|T_\varepsilon|}.$$

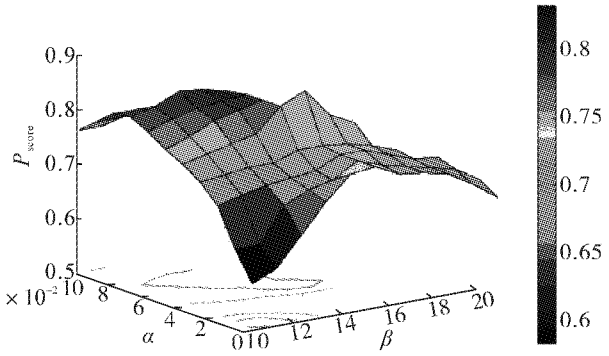
式中:$|T_\varepsilon|$表示事件 ε 的文档对测试集的大小,⊙表示异或运算.

3.3　参数设定

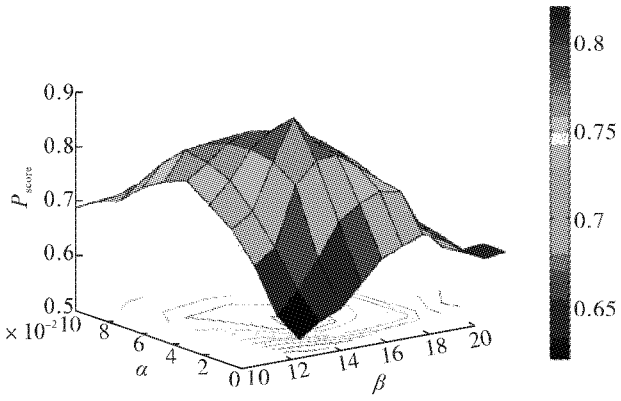
本文通过检验参数对实验结果的影响程度来确定参数的设定. 本文提出的算法中一共包括 3 个参数:α、β、K.

α 表示的是算法第 1 步中滤掉高频词的阈值,这里指的是滤掉的高频词占整个数据集词库总数的百分比. 在参数测定中,实验中让 α 从 0 变化到 10%,变化步长为 0.01. β 表示的是算法第 1 步中提取 popular words 时,每一类取的词数,这里指的是每一类中提取的 popular words 数目占这一类总词数的百分比. 在参数测定中,实验中让 β 从 10% 变化到 20%,变化步长为 0.01.

通过计算 P_{score} 值来检验这 2 个参数的变化以及它们对算法效果的影响. 图 3 展示了 2 个数据集中,不同的 α 和 β 下 P_{score} 的值的分布.



(a) 天安舰事件



(b) 连胜文枪击事件

图 3　CS 和 LSW 中特定 K 下的 α 和 β 的参数设定

Fig. 3　α and β tuning under specific K in CS and LSW

从图 3 中可以看出,得到最好的一组 α_{best} 和 β_{best} 分别是在 CS 数据集中,当 $\alpha=4\%$ 和 $\beta=15\%$ 时,得到的 P_{score} 值最优;在 LSW 数据集中,当 $\alpha=3\%$ 和 $\beta=13\%$ 时,得到的 P_{score} 值最优. 另外,也可以看出,当 α 或者 β 递增时,P_{score} 的值先增后减.

K 值表示的是 LDA 话题分析和 K-means 聚类中类别数的设定,也意味着最后得到的版本数. 提前设定 K 值是对版本数的一个预测. 这里从 2 个方面对 K 值进行设定. 1)让志愿者根据一定数量的阅读新闻报道或分析总结性的新闻报道,获取关于这一特定事件版本信息的先验知识,即志愿者通过大致的浏览分析,得到关于这一特定新闻事件版本数的模糊区间. 通过第 1 步分析后,可以得到 CS 数据集和 LSW 数据集的版本数均在 4 ~ 7 种. 2)同样通过计算 P_{score} 值来最终确定 2 个数据集分为几个版本. 以 1 为步长,让 K 值在 4 ~ 7 变化,分别得到相应的 P_{score} 值,如表 2 所示.

表 2　CS 和 LSW 中 K 的参数设定

Table 2　Parameter tuning of K in CS and LSW

K	CS			LSW		
	α_{best}	β_{best}	P_{score}	α_{best}	β_{best}	P_{score}
4	0.04	0.15	0.780	0.03	0.13	0.820
5	0.04	0.15	0.835	0.03	0.13	0.760
6	0.04	0.17	0.775	0.05	0.16	0.747
7	0.03	0.18	0.750	0.06	0.16	0.713

从表 2 中可以看出,在数据集 CS 时,当 $K=5$ 时,P_{score} 达到最优;在数据集 LSW 中,当 $K=4$ 时,P_{score} 达到最优.

3.4 实验结果及评价

在实验的部分,作者将与几种相关算法进行对比试验,以检验本文提出的 CDW 算法的效果. 相关算法研究包括:

1)K-means:根据文档之间的相似度对文档进行聚类;

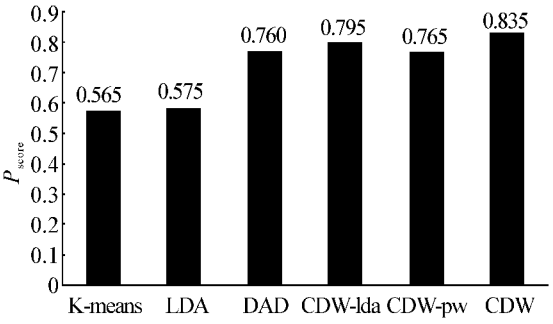
2)LDA:根据词的分布对文档进行聚类;

3)DVD:基于图模型的时间多版本发现算法;

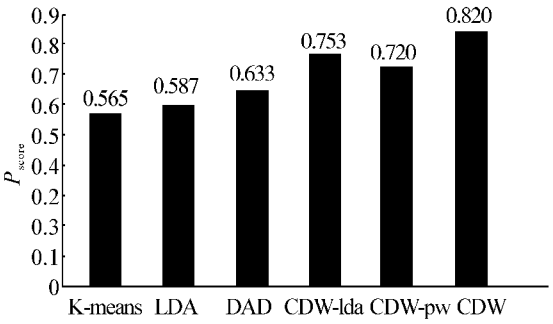
4)CDW-lda:CDW 算法的变种,过滤高频词后不进行 LDA 分析而直接提取 popular words,再进行聚类;

5)CDW-pw:CDW 算法的另一变种,LDA 分析后不用原方法提取 popular words,而是直接用每一类的高频词做 popular words,再进行特征向量化和聚类.

作者在之前构建的2个测试集 T_{CS} 和 T_{LSW} 上进行实验,以比较 CDW 模型和其他相关算法的效果. 不同算法在测试集 T_{CS} 和 T_{LSW} 上的 P_{score} 值如图 4 所示.



(a)天安舰事件



(b)连胜文枪击案事件

图 4 在数据集 CS 和 LSW 上的 P_{score} 值

Fig. 4 P_{score} performance comparison in pairwise test of CS and LSW

最后,用表 3 来展示本文的多版本发现结果. 其中,通过本文中的 CDW 模型,韩国天安舰事件共有 5 个不同版本,中国台湾连胜文枪击案共有 4 个不同版本. 从这个结果中可以看出,本文提出的多版本发现结果是比较准确可靠的.

表 3 CDW 算法对于 2 个事件的多版本发现结果

Table 3 Results of CDW model for diverse versions discovery in CS and LSW

	CS	LSW
1	遭朝鲜鱼雷攻击沉没	凶手认错人,原本袭击对象不是连胜文
2	韩美制造的骗局	凶手受人指使,幕后有操纵
3	韩国自导自演的苦肉计	政治阴谋,与 5 市选举有关
4	美军演习时误击	国民党操纵策划
5	接触水雷而沉没的意外	

从图 4 中可以看出,K-means 算法和 LDA 话题分析在 2 个测试集上的表现都是最差的. DVD 算法的结果相对较好,但是由于 DVD 算法只利用了词与词之间的层级关系,而忽略了文本信息和话题关系,所以它的结果并不如 CDW 算法. 在 CDW 的 3 个版本中,完整的 CDW 算法和 CDW-lda 的结果要好于其他所有算法. 这表明,在话题层上进行 popular words 的提取对于事件的多版本发现问题十分重要. 另外,CDW 算法的结果优于 CDW-lda 也优于 CDW-pw,这意味着文档集与话题的映射和映射之后 popular words 的提取都是十分有必要的.

4 结束语

本文提出了一种基于文本的新闻事件多版本发现的模型,能够帮助读者对某一特定新闻事件进行自动快速的多版本生成. 在论文工作中发现,简单的聚类方法具有很多局限性,无法将文本中不同的版本信息区分开来. 为了取得更好的效果,本算法建立了话题层与文档集合之间的映射关系,将文本集合引申到话题空间,在话题空间中对文本进行高区分度特征的提取. 然后,再根据这些提取出来的特征进行文本聚类,从而得到关于某一新闻事件的多个版本.

通过在 2 个真实的数据集上的实验,可以看出,与以往的相关算法相比,本文提出的方法在事件多版本发现的问题上具有更高的准确性和有效性.

虽然本文的算法取得了非常不错的效果,但是还存在一些需要改进的地方. 在算法中确定多版本类别数目时,算法采用的是提前设定版本数目. 如果可以将自动确定类别数的相关算法应用到多版本的发现问题中,那么将会产生更为准确的多版本结果.

同时,改进最后的聚类分析方法对于提高多版本发现模型的准确性也将起到一定的作用.另外,本文的算法只是对事件的多版本进行生成,而没有对生成结果做总结性概括描述.因此可以考虑加入提取摘要来完善算法,将会更具有实际应用意义.

参考文献:

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

[2]HE T T, QU G Z, LI S W, et al. Semi-automatic hot event detection [C]//Lecture Notes in Computer Science. Hongkong, China, 2006: 1008.

[3]YU M Q, LUO W H, XU H B, et al. Research on hierarchical topic detection in topic detection and tracking[J]. Journal of Computer Research and Development, 2006, 43(3): 489-495.

[4]邱立坤,龙志祎,钟华,等. 层次化话题发现与跟踪方法及系统实现[J]. 广西师范大学学报:自然科学版, 2007, 25(2): 157-160.
QIU Likun, LONG Zhiyi, ZHONG Hua, et al. Hierarchical topic detection and tracking and implementation of system [J]. Journal of Guangxi Normal University: Natural Science Edition, 2007, 25(2): 157-160.

[5]CARTHY J. Lexical chains versus keywords for topic tracking[C]//Proceedings of the 5th International Conference on Intelligent Text Processing and Computational Linguistics. Seoul, Korea, 2004: 507-510.

[6]ALLAN J, CARBONELL J, DODDINGTON G, et al. Topic detection and tracking pilot study final report[C]//Proceedings of the DARPA Broadcasting News Transcript and Understanding Workshop. [S.l.], 1998: 194-218.

[7]YANG Y, PIERCE T, CARBONELL J. A study of retrospective and on-line event detection[C]//Special Interest Group on Information Retrieval'98. Melbourne, Australia, 1998: 28-36.

[8]ALLAN J, PAPKA R, LAVRENKO V. On-line new event detection and tracking[C]//Special Interest Group on Information Retrieval'98. Melbourne, Australia, 1998: 37-45.

[9]BRANTS T, CHEN F, FARAHAT A. A system for new event detection[C]//Special Interest Group on Information Retrieval'03. Toronto, Canada, 2003: 330-337.

[10]NALLAPATI R, FENG A, PENG F, et al. Event threading within news topics[C]// International Conference on Information and Knowledge Management. Washington, DC, USA, 2004: 446-453.

[11]STEINBACH M, KARYPIS G, KUMAR V. A comparison of document clustering techniques[EB/OL]. [2011-05-14]. http://www.cs.cmu.edu/~dunja/KDDpapers/Steinbach_IR.pdf.

[12]PAUL S B, USAMA M F. Refining initial points for K-means clustering[C]//Proceedings of the Fifteenth International Conference on Machine Learning. San Francisco, USA, 1998: 91-99.

[13]JAIN A K, MURTY M N, FLYNN P J. Data clustering: a review[J]. ACM Computing Surveys, 1999, 31(3): 264-333.

[14]RYMOND T, HAN J W. Efficient and effective clustering methods for spatial data mining[C]//Proceedings of the 20th International Conference on Very Large Data Bases. Hong Kong, China, 1994: 144-155.

[15]KONG L, YAN R, HE Y J, et al. DVD: a model for event diversified versions discovery[C]//Asia-Pacific Web Conference'11. Beijing, China, 2011: 18-20.

[16]FLAKE G W, LAWRENCE S, GILES C L. Efficient identification of Web communities[C]//International Conference on Knowledge Discovery and Data Mining00. Boston, USA, 2000: 160-169.

[17]ROCCHIO J. Relevance feedback in information retrieval [C]//The SMART Retrieval System: Experiments in Automatic Document Processing. Englewood Cliffs, USA, 1971: 313-323.

[18]DASGUPTA S, NG V. Towards subjectifying text clustering[C]// Special Inspector General for Iraq Reconstruction'10. Geneva, Switzerland, 2010: 483-490.

[19]DUMAIS S T, PLATT J, HECKERMAN D, et al. Inductive learning algorithms and representations for text categorization[C]// Proceedings of the Seventh International Conference on Information and Knowledge Management. New York, USA, 1998: 148-155.

[20]FRANZ M, WARD T, MCCARLEY J S, et al. Unsupervised and supervised clustering for topic tracking[C]// Special Inspector General for Iraq Reconstruction'01. New Orlean, USA, 2001: 310-317.

[21]BLEI D M, ANDREW Y NG, MICHAEL I J. Latent Dirichlet allocation[J]. The Journal of Machine Learning Research, 2003(3): 993-1022.

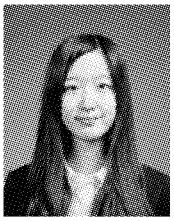
[22]WEI X, CROFT W B. LDA-based document models for ad-hoc retrieval[C]//Proceedings of the 29th Special Inspector General for Iraq Reconstruction Conference. New York, USA, 2006: 178-185.

[23]BHATTACHARYA I, GETOOR I. A latent Dirichlet model for unsupervised entity resolution[C]//SIAM International Conference on Data Mining. Bethesda, USA, 2006: 47-58.

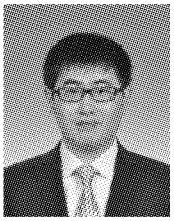
[24]JEROME R B. A novel word clustering algorithm based on latent semantic analysis[C]//Acoustics, Speech, and Sig-

nal Processing 1996. [S.l.], 1996: 172-175.
[25] SUN B, SHI L, KONG L, et al. Describing web topics meticulously through word graph analysis[C]//The IEEE Conference on Instructional Technologies '09. Xiamen, China, 2009: 11-14.
[26] PAGE L, BRIN S, MOTWANI R, et al. The pagerank citation ranking: bringing order to the web[C]//Proceedings of the 7th International World Wide Web Conference. Brisbane, Australia, 1998: 161-172.
[27] KAREN J S. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of Documentation, 1972, 28(1): 11-21.
[28] HARTIGANJ A, WONG M A. A K-means clustering algorithm[J]. Journal of the Royal Statistical Society, Series C: Applied Statistics, 1979, 28(1): 100-108.
[29] PELLEG D, MOORE A W. X-means: extending K-means with efficient estimation of the number of clusters[C]// Proceedings of the Seventeenth International Conference on Machine Learning. Stanford, USA, 2000: 727-734.
[30] MACQUEEN J B. Some methods for classification and analysis of multivariate observations[C]//Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967: 281-297.

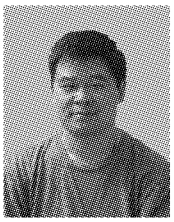
作者简介:



肖融,女,1989年生,硕士研究生,主要研究方向为数据挖掘、信息检索等.



孔亮,男,1985年生,硕士研究生,主要研究方向为数据挖掘、信息检索等.



张岩,男,1970年生,副教授,博士,主要研究方向为web信息处理、智能搜索技术、文本分析与数据挖掘、数据库性能等,发表学术论文多篇.

第4届群体智能国际会议征文通知

The Fourth International Conference on Swarm Intelligence (ICSI'2013)

The Fourth International Conference on Swarm Intelligence (ICSI2013) will be held in Harbin, China from June 12 to 15, 2013. Located in the center of Northeast Asia, Harbin is called the bright pearl on the Bridge of Eurasia Land, and it is also an important hub of Eurasia Land Bridge and air corridor. The special historical course and geographical position has contributed to Harbin, the beautiful city with an exotic tone, which not only brings together the historical culture of northern ethnic minorities, but also combines western and eastern culture. It is a famous historical and tourist city in China, with many beautiful names such as "the City of Culture", "the City of Music", "Ice City", "A Pearl under the Neck of the Swan", "Eastern Moscow" and "Eastern Little Paris".

ICSI2013 serves as a forum for scientists, engineers, educators, and practitioners to exchange the latest advantages in theories, technologies, and applications of swarm intelligence and related areas. Prospective authors are invited to contribute high-quality papers (6-10 pages) to ICSI2013 through Online Submission System. Submitting to ICSI2013 is a blind submission and the reviewing is double blind again at this year. Papers presented at ICSI2013 will be published in Springer's Lecture Notes in Computer Science (indexed by EI, ISTP, DBLP, and ISI) and some high-quality papers will be selected for special issues in SCI-indexed International Journals

Important Dates:

Special session proposals deadline: December 20, 2012

Paper submission deadline: January 01, 2013

Notification of acceptance: March 01, 2013

Camera-ready copy and author registration: March 31, 2013