基于改进的TF-IDF算法的微博话题检测

陈朔鹰[1],金镇晟[2]

1. 北京理工大学网络信息中心,北京 100081
2. 北京理工大学计算机学院,北京 100081

摘要 中文微博具有更新快、时效性强等特点,产生的热点话题均具有一定的突发性,与此同时文本中有代表性的特征词也会随之激增。利用这一特性,在传统的TF-IDF(term frequency-inverse document frequency)基础上提出一种改进的特征权重算法,称之为TF-IDF-KE(term frequency-inverse document frequency-kinetic energy),用以解决突发性热点话题在聚类时特征不明显的问题。该算法结合物体的动能原理,将特征项的突发值用动能的概念进行描述,加入权值计算,提高突发性特征项的权重,最后使用CURE(clustering using representatives)算法,实现微博的话题检测。该方法描述了文本和特征项所具有的动态属性,实验结果表明,该方法能够有效地提高话题检测的效果。

关键词 微博;TF-IDF;话题检测;TDT;文本聚类

随着信息技术的迅猛发展,在互联网上,每天会产生海量的媒体信息,如何快捷、准确地获取感兴趣的内容成了人们面临的一个新的难题。建立以话题为主线的信息模型,可以有效地帮助人们解决信息过载问题。在这种情况下,话题检测与跟踪[1]技术应运而生。TDT[2~4]是近年来提出的一项信息处理技术,包括对信息流进行新话题的自动识别和已知话题的持续跟踪两部分。话题检测,就是在文本流中把讨论相同主题的文本聚类到一起,完成话题的自动识别。

2009年,随着Twitter的风靡全球,中国也出现了微博,人们可以在微博平台上发布不超过140字的文本信息。微博以方便、快捷的发布方式迅速抢占了用户市场,成为了继人人网之后最大的网络社交平台。现如今,微博已成为民众表达舆情的重要窗口,研究微博内容对舆情监控、话题发现、信息安全等领域的决策具有重要作用。

针对长篇文本的研究已经取得了许多成果,并被广泛应用于业界。但微博篇幅极短、特征少且不明显、表达内容不完整,往往会造成信息的无组织现象,导致现有的诸多方法在微博分析中效果不理想;另外微博的传播性极强,产生的话题时效性明显,而现有的许多方法都没有有效地利用这一点。所以此处以微博短文本为研究对象,利用微博的时间属性,改进特征权重算法,优化文本聚类效果。

通过对传统TF-IDF算法以及话题检测系统的研究,本文提出一种基于TF-IDF的新的特征权重算法,用词语的动能来衡量词语的突发值,提高热点特征项的权重,克服短文本特征不明显的缺点。

1 相关研究

目前,国内外已有很多学者在话题检测方面做了大量的研究和尝试。

Allan等[5]利用VSM(vector space model)[6]描述文本,并对诸如人名、地名等命名实体赋予更高的权重,完成话题检测任务。贾自艳等[7]提出增量聚类方法,基于动态模型并不断更新聚类中心,已完成TDT任务,实验取得了很好的效果。闵可锐等[8]则采用支持向量机算法并结合知识库实现TDT,实验表明该方法具有很高的准确率。

但以上方法均忽略了热点话题的突发性特点。为了利用时间这一重要的文本属性,部分学者进行了以下尝试。Fung等[9,10]尝试描述词语的时序分布情况,通过统计时间上的分布情况来表达文本的突发性。Wang等[11]和Suba等[12]通过建立突发性模型来筛选热点词,然后依据热点词进行话题检测。薛峰等[13]通过χ^2检验公式计算词语的突发值,建立一种新的动态文本模型来描述文本动态属性,与传统方法相比实验结果得到了大大提升。翟东海等[14,15]将词频累加和作为文档统计篇数的影响因子引入χ^2检验公式,改进了原有方法,提高了度量热点词突发性的精确度。

这些方法尝试从话题的时间属性入手挖掘突发性热点

收稿日期:2015-04-23;修回日期:2015-06-08

作者简介:陈朔鹰,副教授,研究方向为计算机网络、数据挖掘、物联网,电子信箱:chensy@bit.edu.cn

引用格式:陈朔鹰, 金镇晟. 基于改进的TF-IDF算法的微博话题检测[J]. 科技导报, 2016, 34(2): 282-286; doi: 10.3981/j.issn.1000-7857.2016.2.048

话题。经过实验表明,时间属性对于话题检测任务具有重要的研究价值。与其他估量热点词语的突发值的方法不同,本文结合物体的动能概念,认为词语的突发值就是词语在某一时刻所具有的动能,将词语的动能加入权值计算,提高突发性词语的权重,以便更好地完成文本聚类。

2　基于TF-IDF的改进算法

TF-IDF[16,17]表示的是TF*IDF,TF即词频,表示某一给定的词语在某一文本中出现的频率;IDF即逆向文本频率,表示数据集中文本总数与包含某一给定的词语的文本数之商。其主要思想是如果某个特征项在一个文本中出现频率很高,且在其他文本中出现很少,说明此特征项具有很好的类别区分能力,应该给予较高的权重,这一思想符合香农信息论。其表达式为

$$w_{ij}=tf_{ij}\cdot\lg(N/n_j) \tag{1}$$

其中,w_{ij}为特征项j在文本i中的权重,tf_{ij}为特征项j在文本i中出现的频率,N为数据集中的文本总数,n_j为数据集中包含特征项j的文本数。

但对于某个词语,在长文本里可能比短文本里有更高词频,不管这个词语是否重要,所以应避免TF对长文本的偏袒。同时又考虑到小数据集的情况,应避免lg函数为零,所以将式(1)改进为:

$$w_{ij}=\left(m_{ij}/M_j\right)\cdot\lg\left(N/n_j+0.01\right) \tag{2}$$

其中,m_{ij}表示特征项j在文本i中的频率,M_j表示文本i中的词频总数,其他参数与式(1)中相同。

后来又有研究者对TF-IDF进行了改进,命名为TFC[17]和ITC[18]。TFC与式(2)的目的类似,同样也避免对长文本的偏袒,并且将特征项的权重进行归一化:

$$w_{ij}=\frac{tf_{ij}\cdot\lg(N/n_j+0.01)}{\sqrt{\sum_{P=1}^{K}\left[tf_{ij}\cdot\lg(N/n_j+0.01)\right]^2}} \tag{3}$$

式中,K为文本i中特征项的个数。

ITC则考虑得更为全面,不仅考虑了将特征项归一化,同时也考虑了小数据集情况,ITC实际上是对式(3)进行公式化:

$$w_{ij}=\frac{\lg(tf_{ij}+1.0)\cdot g(N/n_j+0.01)}{\sqrt{\sum_{P=1}^{K}\left[\lg(tf_{ij}+1.0)\cdot\lg(N/n_j+0.01)\right]^2}} \tag{4}$$

目前,很多研究人员在实际使用过程中和论文中提到的TF-IDF算法均是进行改良后的公式,其中以式(2)和式(3)的应用较为广泛。

现实环境中产生的热点话题往往具有突发性和不可预知性,伴随爆炸式传播,文本的变化(即动态属性)十分明显。但现有的传统方法由于微博文本短小,每个词在文本中的比重相差无几,致使文本的特征项不明显,实际话题检测效果均不理想。

观察发现,当文本在急速增长时,其所具有的词语会不断重复地出现,如果是关键词,更会在转发微博或者相关微博中重复出现。简而言之,文本的增长性很明显,伴随着的关键词语动态变化也很突出。可以判定,话题内容中的关键词语(即热点词)会随着话题一样出现短时间内高频次的增长,即动态增长越剧烈的词语突发性越强、特征性越明显,应作为聚类的重要依据。然而也不能武断地将包含热点词的所有微博聚集起来,这样做错误率很高。基于以上考虑,希望利用热点词的时间属性和增长程度,科学、有效地加强文本的特征,通过提高重点特征项的权重来达到有效聚类的目的,进而解决微博特征不明显、话题检测效果不佳的问题。

但通过对以上TF-IDF算法的分析,现有改进后的TF-IDF没有体现词语的突发性,不能有效地实现话题检测的文本聚类。所以本文提出一种改进的TF-IDF算法,利用词语的时间属性,计算词语的增长速度,用动能定理为词语估量词语的突发值。具体做法:在词频统计时,对每个词语分别记录频次、包含该词的文本数、首次出现时间和末次时间,计算词语的增长速度,结合公式$E_k=(1/2)mv^2$,估量词语的突发值,并将其加入词语的特征权值中,这样就将词语的动态属性表现在了公式中,提高其在文本聚类时的作用。此处将改进的算法命名为TF-IDF-KE,公式为:

$$w_{ij}=\frac{\left[tf_{ij}\cdot\lg(N/n_j+0.01)\right]}{\sqrt{\sum_{P=1}^{K}\left[tf_{ij}\cdot\lg\left(N/n_j+0.01\right)\right]^2}}+\alpha\cdot(m_{ij}/M_i)\cdot v_j^2 \tag{5}$$

在式(3)的基础上,式(5)前半部分的参数含义与式(3)中的基本一致。

式(5)的后半部分为特征项的突发值计算部分,v_j表示特征项j的增长速度,计算方法是将特征项首次出现时间与末次时间做差,换算成小时,然后用m_{ij}除以该时间,计算出特征项的平均增长速度;α为动能权重系数,这是一个经验值,需要通过实验确定。

3　中文微博话题检测系统的设计与实现

与一般的话题检测步骤类似,实验所采用的话题检测系统也分为数据采集、预处理、分词和词频统计、特征权重计算和文本表示模型、以及文本相似度计算和聚类这5个步骤。

3.1　数据采集

以中文微博为研究对象,数据主要从国内最大的微博社交平台——新浪微博获取。为防止网络爬虫对其进行长时间的数据抓取,新浪微博对自身建立了保护机制,如每隔一段时间变更微博的URL结构、设置验证码等。

本实验利用新浪API,设计了新的采集方案。利用微博平台中关注的好友所发的最新微博会刷新在自身页面上这一机制,申请一个账号并关注多个所要获取的微博账号(上限为2000个),利用API不断获取当前账号页面上刷新的微博数据。一次API调用可获取一页微博数据,一页最多200条,一分钟内可最多获取20页。数据获取量完全可以满足本实验的要求,而且该方案不会受到新浪官方的任何阻拦。

3.2 预处理

由于微博文本的内容不规范,所以需要对原始的微博数据进行相应的处理,删除文本中的无关数据、乱码,尽可能消除噪声,以便后期进行话题检测处理。主要操作有:删除带有“@用户”格式的字段内容;删除无法识别的乱码;删除图片、表情、网址等内容。

3.3 分词和词频统计

分析文本就要先进行分词。中文与英文不同,词语之间没有明显的间隔区分,所以需要借助现有的分词技术。目前中文的分词系统有很多,如ICTCLAS、FudanNLP、ANSJ等。实验选用ANSJ中文分词系统,这是一个基于Google语义模型+条件随机场模型的开源的Java版中文分词工具。分词速度达到每秒钟大约200万字,准确率达到98%以上,且操作简单,适合作为分词系统使用。

分词后对词语进行统计,但统计之前先要筛除无意义的停用词(包括感叹词、介词、连词、象声词、方位词等)。为词语建立带有时间属性的词语统计表,存储每个词语的词频数、包含该词语的文本数、首次出现时间、更新时间等多项信息,为文本的特征权重计算提供数据支持。

3.4 特征权重计算和文本表示模型

特征权重计算是对文本分析的第一步也是重要的一步,理想情况应为越重要的特征性词语所占的权重越高,但由于微博文本特征不明显,现行的特征权重方法效果不佳,最终导致微博话题聚类效果不理想。本文选用最经典的TF-IDF算法,但这种方法没有体现词语的突发性,不能有效地支撑突发性热点话题的发现,所以对此算法提出了改进方法,即TF-IDF-KE算法,对词语的突发值进行估量,具体方法在上文中已做详细阐述。

当计算出每个词语的权重之后,将文本转化为计算机可以处理的文本表示模型。运用向量空间模型(VSM)对文本建模,将每个文本表示为一 n 维的向量:

$$\boldsymbol{d}_i=(t_1:w_1;t_2:w_2;\ldots;t_n:w_n) \tag{6}$$

式中,t_i 为特征项(词语),w_i 为特征项 t_i 的权重。

3.5 文本相似度计算和文本聚类

进行文本聚类时,使用文本的相似度值作为聚类的判断依据。此处选取应用最为广泛的余弦相似度[19]公式计算文本间的相似度值,它以两个向量的夹角作为文本相似度的衡量,夹角越小,表明2个文本越相似。其计算公式为:

$$\mathrm{Sim}(D_1,D_2)=\frac{\sum_{i=1}^{n}x_i\cdot y_i}{\sqrt{\left(\sum_{i=1}^{n}x_i^2\right)\cdot\left(\sum_{i=1}^{n}y_i^2\right)}} \tag{7}$$

式中,D_i 为某一文本,x_i 和 y_i 分别为2个文本的每一个特征值,n 为文本所含的特征个数。

文本聚类就是按照一定的规则将一个文本集合中的文本分割成不同的类簇,并使得类内文本尽可能相似,类间差异性尽可能最大。本文采用CURE[20]算法做文本聚类,因为在实际环境中,数据是杂乱无章、孤立点较多的,容易影响聚类效果,而CURE采用多个点代表一个簇的方法,对孤立点的处理更加健壮,且能够识别非球形和大小变化比较大的类。

4 实验与分析

4.1 实验环境

为了分析上述TF-IDF-KE算法在话题发现与跟踪上的效果,按照文中上述所设计的话题检测系统进行实验操作,在Eclipse上采用Java语言实现。

实验共采集新浪微博平台上的1600余条文本(筛除无字或内容少于4个字的微博)作为本次实验的数据集,采集时间为2014-07-29—2014-08-01,选取其中的3/4作为训练集,1/4作为测试集,以验证新方法的实际效果。同时对整个数据集进行人工标注话题,定义了包括周永康案、新疆莎车暴恐案、户籍制度改革、埃博拉病毒、张默吸毒等多个热门话题。

4.2 实验结果评价标准

本实验采用文本聚类的基本评测指标准确率 P 、召回率 R 和 F 测试值检测和评价实验结果的优劣程度,其中 F 的值越高说明方法的聚类效果越好。计算公式定义为:

$$P=\frac{A}{A+B} \tag{8}$$

$$R=\frac{A}{A+C} \tag{9}$$

$$F=\frac{2PR}{P+R} \tag{10}$$

式中,A 表示数据集中本来属于类别 C_i 并且被正确聚类到类别 C_i 的文本数,B 表示本来不属于类别 C_i 但被聚类到类别 C_i 的文本数,C 表示本来属于类别 C_i 但被聚类到其他类别的文本数。

4.3 实验设计及结果分析

为证明新算法的优越性和有效性,设计并实现了3个实验,采用基于传统TF-IDF(式(2))和基于归一化TF-IDF(式(3))作为baseline模型,最后和基于TF-IDF-KE(式(5))的话题检测效果作对比。

实验过程中,主要对以下参数做了观察和记录,分别是CURE算法中的类代表点个数 $cure\_c$ 、收缩因子 $cure\_\alpha$ 、相似性阈值 $cure\_sim$ (聚类结束判断条件)和TF-IDF-KE公式中的动能权重系数 α 。

从图1~图4可以得出:

1)4个参数对于话题检测系统的影响程度是大不相同的,按照 F 值得波动情况可将4个参数排序为 $cure\_sim>\alpha>cure\_\alpha>cure\_c$;

2)从图1~图3中可以看出,采用归一化TF-IDF的实验效果略优于采用传统TF-IDF的实验效果,说明作为权重计算方法,归一化TF-IDF算法略优于传统TF-IDF算法;

3)通过将采用TF-IDF-KE算法与采用传统TF-IDF和归一化TF-IDF的实验结果作比较,发现在同等参数条件下,基于TF-IDF-KE权重计算的话题检测方法明显优于其他两种baseline模型,说明本文提出的结合词语突发值的TF-IDF

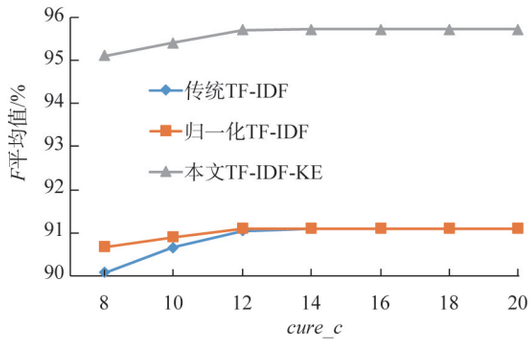

图1　实验效果随 *cure_c* 变化的情况

Fig. 1　Experimental effect varies with the *cure_c*

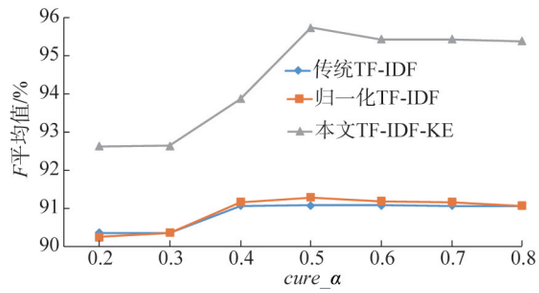

图2　实验效果随 *cure_α* 变化的情况

Fig. 2　Experimental effect against *cure_α*

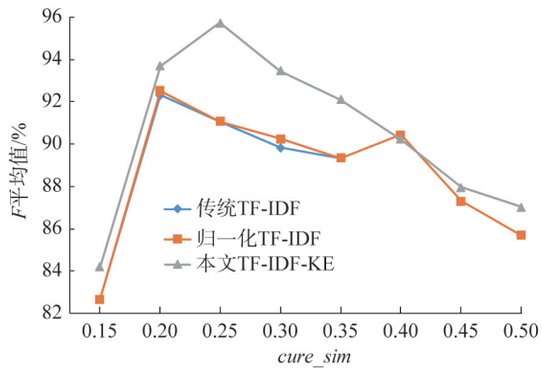

图3　实验效果随 *cure_sim* 变化的情况

Fig. 3　Experimental effect against *cure_sim*

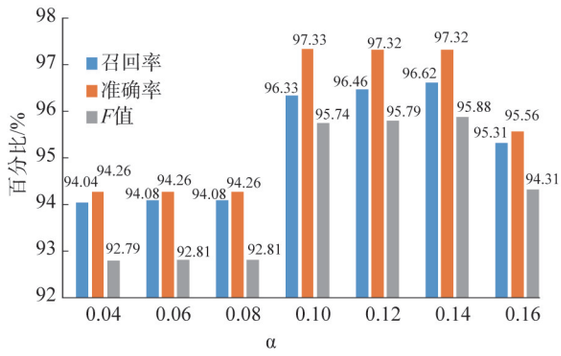

图4　实验效果随 α 变化的情况

Fig. 4　Experimental effect against α

改进算法是一种有效的优化方法;

4)在实验过程中,经过反复的试验和测试,发现当 $cure\_sim \in [0.20,0.40]$, $\alpha \in [0.10,0.16]$, $cure\_\alpha \in [0.45,0.55]$, $cure\_c \in [12,20]$ 时,话题检测系统会有最优效果。

另外在实验中还发现,越是大事件(如"周永康案"),微博信息碎片化特性越明显,信息越容易被遗漏。而TF-IDF-KE算法恰好在解决此种情况时有明显的优化作用,可以有效提升话题的准确率和召回率。

本实验中特征项的增长速度单位为小时,由于微博平台上文本产生十分迅速,直接导致式(5)中特征项增速值对于权重计算而言过大,动能权重系数过小,致使动能部分的调节优化效果不够精细,仍需进行新的尝试。

在实验过程中还发现,人名和地名对于一个事件或话题具有很好的特征和描述作用,对人名和地名增加权重有可能会起到很好的聚类优化作用,可对此做进一步研究。

5　结论

为了更好地表示文本的动态属性,针对热点话题的突发性特点,结合物体的动能概念,本文提出了一种新的特征权重算法,将突发值定义为特征项的动能加入到权值计算公式中,优化文本向量空间模型,在此基础上进行文本聚类并实现话题检测。实验结果表明,运用此方法进行话题检测,可以获得更好的准确率 P、召回率 R 和 F 值。

参考文献(References)

[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): 71-87.
Hong Yu, Zhang Yu, Liu Ting, et al. Topic detection and tracking review[J]. Journal of Chinese Information Processing, 2007, 21(6): 71-87.

[2] Li H, Wei J. Netnews bursty hot topic detection based on bursty features[C]. Proceedings of the 2010 International Conference on E-Business and E-Government.Guangzhou, China, 2010.

[3] Holz F, Teresniak S. Towards automatic detection and tracking of topic change[M]//Computational linguistics and Intelligent Text Processing. Berlin Heidelberg: Springer, 2010: 327-339.

[4] Qiu J, Liao L J, Dong X J. Topic detection and tracking for Chinese news web pages[C]. Advanced Language Processing and Web Information Technology, 2008. ALPIT'08. International Conference on Dalian, China, 2008.

[5] Allan J, Lavrenko V, Jin H. First story detection in TDT is hard[C]//Proceedings of the Ninth International Conference on Information and Knowledge Management. Atlanta, Georgia ACM, 2000: 374-381.

[6] Salton G, Wong A, Yang C S. A vector space model for automatic index-

ing[J]. Communications of the ACM, 1975, 18(11): 613-620.
[7] 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.
Jia Ziyan, He Qing, Zhang Haijun, et al. A news event detection and tracking algorithm based on dynamic evolution model[J]. Journal of Computer Research and Development, 2004, 41(7): 1273-1280.
[8] 闵可锐, 赵迎宾, 刘昕, 等. 互联网话题识别与跟踪系统设计及实现[J]. 计算机工程, 2008, 34(19): 212-214.
Min Kerui, Zhao Yingbin, Liu Xin, et al. Design and implementation of topic detection and tracking system on web[J]. Computer Engineering, 2008, 34(19): 212-214.
[9] Fung G P C, Yu J X, Yu P S, et al. Parameter free bursty events detection in text streams[C]//Proceedings of the 31st International Conference on Very Large Data Bases. Trondheim, VLDB Endowment, 2005: 181-192.
[10] Fung G P C, Yu J X, Liu H, et al. Time-dependent event hierarchy construction[C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose: ACM, 2007: 300-309.
[11] Wang X, Zhai C X, Hu X, et al. Mining correlated bursty topic patterns from coordinated text streams[C]//Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. San Jose, USA: ACM, 2007: 784-793.
[12] Subašic I, Berendt B. From bursty patterns to bursty facts: the effectiveness of temporal text mining for news[J]. Aug-2010, 2010.
[13] 薛峰, 周亚东, 高峰, 等. 一种突发性热点话题在线发现与跟踪方法[J]. 西安交通大学学报, 2012, 45(12): 64-69.
Xue Feng, Zhou Yadong, Gao Feng, et al. An online detection and tracking method for burstytopics[J]. Journal of Xi'an Jiaotong University, 2012, 45(12): 64-69.
[14] 翟东海, 聂洪玉, 崔静静, 等. 基于改进的χ^2检验的热点词突发性度量研究[J]. 计算机与数字工程, 2013, 41(11): 1788-1790.
Zhai Donghai, Nie Hongyu, Cui Jingjing, et al. Busty measurement of hot term based on improvement χ^2 test combined with TF[J]. Computer & Digital Engineering, 2013, 41(11): 1788-1790.
[15] 翟东海, 王佳君, 聂洪玉, 等. 基于互信息的热点词发现和突发性话题检测研究[J]. 西藏大学学报: 自然科学版, 2013(1): 17.
Zhai Donghai, Wang Jiajun, Nie Hongyu, et al. Research on bursty topic detection and hot word extraction based on mutual information [J]. Journal of Tibet University, 2013(1): 17.
[16] Salton G, McGill M J. Introduction to modern information retrieval[M]. New York: McGraw-Hill, 1983.
[17] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1988, 24(5): 513-523.
[18] Nigram K, Lafferty J, Mccallum A. Using maximum entropy for text classification[J]. In Proceedings of IJCAI-1999, 1999.
[19] Pang-Ning T, Steinbach M, Kumar V. Introduction to data mining[M]. Boston: Addison-Wesley, 2006.
[20] Guha S, Rastogi R. Cure: An efficient clustering algorithm for large databases[J]. Information Systems, 2001, 26(1): 35-58.

Weibo topic detection based on improved TF-IDF algorithm

CHEN Shuoying[1], JIN Zhensheng[2]

1. Department of Network Information Center, Beijing Institute of Technology, Beijing 100081, China
2. School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081, China

Abstract The topic detection and tracking (TDT) is an issue of natural language processing, which concerns with solving the problem of information explosion. The Weibo TDT is a central issue in recent years. A bad performance is usually achieved for Weibo with a short text, while the topic detection of a long text is widely used in the industry with better results. Weibo's features of short text and not very clear meaning make the clustering algorithms' effect not ideal in topic detection. So this paper focuses on finding a new way to improve the effect of clustering for Weibo. Weibo features fast renewal and strong timeliness. Hot topics produced by Weibo show burstiness, and their representative words increase in a great extent. With this feature in mind, improving the representative word's weight to a certain degree is a good way to give a prominence to the feature of short text. The burstiness of the words is a thing to consider, similar to the kinetic theory of the object. The formula of the kinetic energy theorem is used in this paper. Then an improved feature extraction algorithm named the TF-IDF-KE (term frequency-inverse document frequency-kinetic energy) is proposed. The new algorithm consists of the kinetic energy and the TF-IDF (term frequency-inverse document frequency). The formula of the kinetic energy theorem is used to evaluate the burstiness of the words and add the value to the formula. Then, the weight of some important words can be improved when extracting features. Finally, the implementation of the CURE (clustering using representatives) algorithm completes the Weibo topic detection task. The method presented in this paper describes burstiness of text and feature and solves the problem that the feature of bursty hot topics is not obvious, when clustering in a certain extent. The experimental results show that the method can effectively improve the effect of topic detection in some degree and a better accuracy rate P can be achieved, as well as the R and F values of the recall rate. So TF-IDF-KE is an effective optimization method and can well be used for the task of the TDT.

Keywords Weibo; TF-IDF; topic detection; TDT; text clustering

(责任编辑 刘志远)