识别网络新闻标题党

张晓春

内容摘要 :随着网络的迅猛发展和媒体竞争的加剧以及市场化步伐的加快,“标题党”成为一个越来越突出的现象。不少“标题党”为了提炼成“语不惊人死不休”的标题,不惜断章取义、张冠李戴甚至歪曲事实,危害极大。本文首先对标题党做了解释,然后根据目前网络新闻标题存在的问题提出识别标题党的必要性,接着对识别网络新闻标题党的主要技术做了简要分析和比较,指出存在的不足之处,最后提出了一些自己的建议。

关键词 :标题党 网络新闻 识别技术 方法介绍

DOI:10.16692/j.cnki.wxjys.2018.02.077

我们通常认为,在网络上,一篇新闻报道要想获得较高的点击量,必须有吸引人的标题、优质的内容、广泛的推广渠道,相比于内容和渠道,对标题的再加工无疑是成本最小、收效最大的。随着网络信息的膨胀与快速变化,仅仅为吸引人眼球的标题党谈不上任何创新创意,也绝无过人之处,有的只是罔顾事实,误导舆论。网络新闻标题党虽然能够博人眼球,但是却失去了新闻报道最重要的功用——保障知情、舆论监督、促进公正,对于新闻舆论的公信力也是极大损伤。

本文首先介绍了网络新闻标题党出现的原因及其危害,通过对几个典型的标题党新闻的分析来探究标题党新闻识别技术的主要目的,介绍当前国内主要标题党新闻识别技术,厘清标题党新闻识别的主要方法,指出这些方法存在的问题和局限性。最后,针对如何更好地进行标题党新闻识别提出了一些自己的见解,希望标题党识别技术能够在将来变得更加完善、高效,在一定程度上阻止标题党新闻的传播,提高受众者的阅读品质。

一.何为“标题党”

众所周知,标题是新闻的眼睛,在当前信息化快速发展的时代中,人们接受新闻的方式呈现出多样性,出现5秒效应或者看新闻看题的提法,这种情况下在一定程度上体现出了新闻标题的作用。此外,在网络新闻中标题同样需要加以关注,只有如此才能让快速抓住受众,完成引导活动,形成良好阅读,换而言之,现阶段网络新闻的竞争主要体现在网络新闻标题的竞争之上。

《现代汉语词典》(第六版)中并没有收录“标题党”这一词条。一般认为,“标题党”是发端于网络论坛,发帖者为吸引人气,提高帖子的点击量而制作博人眼球标题的网络贴主群体或行为。“标题党”可以说是这样一些信息发布和转发行为、人群的总称。“标题党”的目的不是为传播信息本身,而是为了引起关注,获取更多的经济利益。网编们为了在海量的信息中异军突起,获取点击率,争取到好的排位,有更多的广告收益,纷纷仿效广告业“标题党”的手法将大量新闻标题重新包装,挖空心思“哗众取宠”,大量“题不对文”的新闻标题就此出笼。从特点上分析,标题党具有两个特征,第一是夸大性:主要是对文章内容进行夸大,以此起到吸引受众眼球的作用;第二是作假型:标题与文章的内容有所差别,内容描述的是另外一个事实,但是为了提高受众的注意力,则将标题设定为其它内容,起到吸引的作用。无论哪一个特点,均在一定程度上反映出了标题党这种现象没有遵循新闻的实际情况,也没有符合新闻发展的要求,长此久往则会导致新闻呈现出缺陷与不足,甚至在新时期让人们对新闻形成不良情绪。

二.标题党新闻识别的必要性

原标题:“大胸”比“平胸”更易患乳癌

记者获悉,中国女性超过一半是致密性乳腺,患乳腺癌的风险比脂肪性乳腺高4.7倍,而且,密集的腺体易掩盖早期癌症病症。这是因为,亚洲女性的乳房相对较小,且以致密性乳房为主。以致密性乳腺为主的亚洲女性,如果只用手动超声做为第一线做乳腺癌筛查,可能会存在一定的漏诊。

析因:超过50%的中国女性是致密性乳腺

在中国,超过50%的女性具有致密性乳腺。拥有致密性乳腺的女性,相对于脂肪性乳腺的女性,罹患乳腺癌的风险高4.7倍。(节选)

这种新闻就是典型的“标题党”!全篇新闻没有一次出现“大胸”、“平胸”,但在标题中却出现了“‘大胸’比‘平胸’更容易患乳腺癌”的结论,简直是惊为天人。本来一个“中规中矩”的标题,在他们手下,就变了大样。“标题党”抓住人们“扫视”新闻的这一心理,误导大众,不明事实真相的“吃瓜群众”将这种不经过大脑分析的消息通过网络传播给他人,以讹传讹,谣言就产生了。从另外一个角度分析,如果这种文章传输给受众,则会让受众无法辨别其真实性,甚至还会产生抵触心理,严重影响了新闻的真实性以及全面性。除此之外,在当前的发展背景下,需要清楚的认识到标题党所带来的影响,并且能够从本质出发,从新闻识别的方式出发,对标题党新闻的识别方法进行对比与分析,这样才能真正提高新闻的可行性以及创新性。

三.标题党新闻识别方法介绍

及对比

1. 基于主题句分布的标题党新闻识别算法

在新闻标题与主题内容相关程度研究方面,国内有学者对“标题党”类新闻的识别进行了研究,王志超提出中提出了一种基于内容主题句相似度的“标题党”新闻识别方法:首先从正文中提取出可能反映正文主题的句子集合,再分别计算它们与标题的相似度,并以最大相似度作为评价参数。这种方法归根到底是网页信息抽取→主题句提取→句子相似度计算。但这个方法对于“以偏概全”的新闻难以达到较好的识别作用。它对一些同义词以及未登录词、专有名词无法很好地识别。比如林俊杰演唱会的报道中同时出现“林俊杰”和“JJ”,他们是一个人,但是基于主题句的识别方法认为两个词不相关,句子相似度计算出现误差。这种方法终究没有避免TF-IDF只考虑上下文统计特性而不考虑语义信息的局限。汉语句子的表达形式是多种多样的,如果要准确地刻画一个句子所表达的意思,还应该结合语法结构信息。所以在进行新闻报道的时候,需要多角度的分析与研究,并且要从本质商除法,对主题句分布的标题党新闻识别加以重视,如此才能实现新闻报道的有效与全面。

2. 基于主题词分布的识别算法

考虑到主题句识别算法的短板,又有学者引入了基于主题词分布的新闻识别算法。首先从新闻标题中提取出最能反映标题中心含义的主题词。再分析主题词在新闻正文中的分布情况,最后根据分布情况计算出是正常新闻的概率,从而判断是否为标题党新闻。这个算法考虑到短语结构分析和依存句法分析,分析句子结构对主题词集合K的形成提供了帮助。但是这个方法也存在一些问题,根据依存句法所构建的24种关系词典都可以进行二次处理吗?在对多名词短语处理时如何确定这个名词短语是否过滤?

北京饭馆老板换大招牌迎接奥运

“北京饭馆老板”为并列名词,处理后变为:“老板换大招牌迎接奥运

万一作者强调的重点就是这个“北京老板”喜迎奥运呢?这种情况该如何识别?

3. 基于潜在语义的标题党新闻识别算法

这种标题党识别技术以潜在语义分析算法为理论基础、以矩阵的奇异值分解为核心。这种方法的优点有以下几点:1.可消除无关词语的干扰。2.抽取正文简单有效。3.从与新闻标题相关的段落数占总段数的比值,以及这些段落内容总长度占新闻正文总长度的比值两个角度对目标新闻内容进行双重判定。这种方法有是有明显的局限性的,⑴ 网页新闻布局结构造成识别误差 ⑵ 分词词典具有局限性,人名、地名、网络词汇等未登录词无法正确识别。

从总体上看,这三种方法都有各自明显的优势和缺点,如果能将主题词识别新闻技术和潜在语义识别新闻技术结合起来使用,应该是很有帮助的。在提取新闻正文和进行分词的过程中我认为还有一些问题需要注意:

⑴ 现在的网页不只包含新闻正文,网页两边还会有广告,正文下方还会有“相关推荐”。这部分信息肯定会干扰关键词的识别,那该如何筛除这部分信息?

⑵ 在对标题、正文进行切词时,应该采用多种分词算法相结合的方法。如切分“北京大学生运动会今日开幕”这样的句子,可以使用正、逆向最大匹配方法,得到粒度更细的分词结果。但是双向最大匹配无法发现链长为偶数的交集型歧义,那么可以增加回溯机制。

⑶ 分词词典是进行标标题识别不可缺少的部分,未登录词、新词语、方言词难以识别是造成识别率较低的原因之一。建立动态分词词典和语料库就是我们必须要提上日程的事。同时,现在的新闻标题,尤其是网络新闻标题中大量使用字母词,对字母词的识别也是我们需要注意的。

⑷ 在《基于潜在语义分析的标题党新闻识别技术研究》中,作者使用了基于词频统计的方式。但是,仅仅考虑新闻的分词单位和段落之间的关系是不够的,还需要了解分词单位与其上下文之间的关系。这时可以利用TF·IDF算法来计算各单位对于各个段落的重要程度。所以无论从哪一个角度分析,均可以清楚的认识到在网络新闻标题是十分关键的,需要多加关注与研究,加强分析与研究,避免出现标题党现象的发生。

总而言之,在当前时代的不断发展下,网络新闻标题语中存在很多不规范的地方。这些不规范的现象产生了许多消极的影响,严重破坏了语言的纯洁性,尤其是标题党的出现,没有遵循新闻的基本要求,也没有从本质出发,实现新闻传播的有效性,故此本文通过对网络新闻标题党识别算法进行简单分析和比较,对网络新闻标题的识别情况有一个整体性的了解和认识,发现了新闻标题党识别技术在使用上存在的一些问题,并提出相关的意见和建议。虽然现在已经出现多种识别网络新闻标题党识别的技术,但是它们都存在或多或少的问题,对于这方面的研究还需要继续深入,我们能做的还有很多。

参考文献

[1]王志超,翁楠,王宇.基于主题句相似度的标题党新闻鉴别技术研究 [J]. 北京:现代图书情报技术,2011,(11):48-53.

[2]朱青,李贞昊.基于主题词分布的低价值新闻识别技术研究[J].上海:计算机应用与软件,2015,(7):190-195.

[3]罗佳.基于潜在语义分析的标题党新闻识别技术研究[J].武汉:湖北工业大学硕士论文,2015年6月.

[4]常鹏,马辉.高效的短文本主题词抽取方法 [J] .北京:计算工程与应用,

(作者单位:厦门大学人文学院中文系)