Journal of Image and Signal Processing 图像与信号处理, 2020, 9(3), 137-145
Published Online July 2020 in Hans. http://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2020.93017

Hans汉斯

Research on News Recognition of "Clickbait" Fusing Image Information

Lin Yang, Jichao Ding, Sheng Zhu, Shuai Wang

School of Mechanical Electronic & Information Engineering, China University of Mining and Technology (Beijing), Beijing
Email: d.jichao.dp@gmail.com

Received: May 13th, 2020; accepted: Jun. 4th, 2020; published: Jun. 11th, 2020

Abstract

In recent years, online news has gradually replaced the traditional paper news, becoming the main way for people to get news daily. Internet news has thus become an industry. The main body of the industry is composed of news producers, users, and click-through rates. News producers convert users' clicks into click-through rates to obtain benefits, which leads to the emergence of the "Clickbait". The "Clickbait" is a huge hazard to users, industries and even society. In the past, the identification method of "Clickbait" ignored the feature that the main body of online news was composed of two parts: image and text. It only detected the text information of the news and ignored the image information of the news. At present, there is a lot of news on the Internet that uses eye-catching or irrelevant images to attract users to click. This article designs an image information extraction model using deep learning related technology for the image information in the news. Using this model, the news Extract the information of the image, design the features of the extracted information, integrate the image features into the "Clickbait" recognition model, and finally verify the necessity and effectiveness of using the image information to identify the "Clickbait" news.

Keywords

Image Caption, Clickbait, Machine Learning, Text Similarity

融合图片信息的"标题党"新闻识别研究

杨　林，丁继超，朱　胜，王　帅

中国矿业大学(北京)机电与信息工程学院，北京
Email: d.jichao.dp@gmail.com

收稿日期：2020年5月13日；录用日期：2020年6月4日；发布日期：2020年6月11日

文章引用: 杨林, 丁继超, 朱胜, 王帅. 融合图片信息的"标题党"新闻识别研究[J]. 图像与信号处理, 2020, 9(3): 137-145.
DOI: 10.12677/jisp.2020.93017

摘　要

近年来，网络新闻逐渐取代传统的纸质新闻，成为人们日常获取新闻的主要方式。网络新闻因此而成为一个产业，产业的主体是由新闻制作者、用户、和点击率构成，新闻制作者通过用户的点击转换为点击率获取利益，由此导致了“标题党”的产生。“标题党”对用户、行业乃至社会都有巨大的危害。以往的“标题党”识别方法都忽略了网络新闻的主体是由图像和文本两个部分组成这一特点，只针对新闻的文本信息进行检测，而忽略了新闻的图片语义信息。目前网络上有大量利用引人眼球或者与文章毫不相关的图片吸引用户点击的新闻，本文针对新闻中的图像信息利用深度学习相关技术设计了图像语义描述与信息提取模型，使用这一模型，对新闻中的图片进行信息提取，对提取到的信息进行特征设计，将图片特征融合进“标题党”识别模型中，最后通过实验验证了使用图片信息识别“标题党”新闻的必要性和有效性。

关键词

图像描述，标题党，机器学习，文本相似度

1. 引言

网络新闻在传播方式上具有流量化、去中心化、题文分离的特点，新闻标题既承担文章内容的概括作用，同时还作为文章链接负责将用户引导到新闻正文部分[1]。网络新闻内容多、数量大、更新速度快的特点，使用户逐渐养成快速阅读新闻标题并根据标题内容决定是否点击链接的“读题”式浏览习惯，也在素材趋于同质化的新闻行业中催生了“标题党”的产生[2] [3]。“标题党”新闻常使用过度夸张、歪曲事实、制造虚假新闻等手段，加工出引人眼球、耸人听闻的标题，以迎合、吸引用户。“标题党”新闻的泛滥，不仅影响用户获取所需的新闻内容，而且对网络新闻行业的健康发展以及网络和谐环境的构建造成恶劣影响。“标题党”现象已受到国家相关部门和社会各界的关注，相关部门积极采取各种措施对“标题党”新闻进行打击和消除，并加强在“标题党”新闻识别方面的研究。

2. “标题党”新闻识别方法的介绍及分析

2.1. “标题党”新闻识别方法研究现状

目前，针对“标题党”新闻识别的研究主要以“标题党”写作中使用的语言特征为主要研究对象，常用的“标题党”新闻识别方法可分为基于文本相似度的识别方法和基于机器学习模型的识别方法两大类。

基于文本相似度的识别方法主要针对“标题党”新闻中“题文不符”的现象，利用文本相似度计算对“标题党”新闻进行判断。2011 年，王志超提出基于主题句相似度的“标题党”新闻鉴别技术，利用正文主题句集合与标题的文本相似度对“标题党”新闻进行识别[4]。2015 年，罗佳提出基于潜在语义的“标题党”新闻识别技术，利用正文中的词频构建向量空间模型，奇异值分解后得到正文的塌陷矩阵，生成标题的文档坐标，通过计算文档坐标和与各段落对应向量间的余弦相似度进行判断[5]。2018 年，赵帅提出基于改进型 VSM-How Net 融合相似度算法的“标题党”新闻识别方法，使用同义词词组的向量替

3.1. 图像语义描述与信息提取

本文使用融合场景信息的图像描述模型[11]获取新闻图片中的语义描述信息，获取新闻图像的图像描述，模型主要结构如图 2 所示。

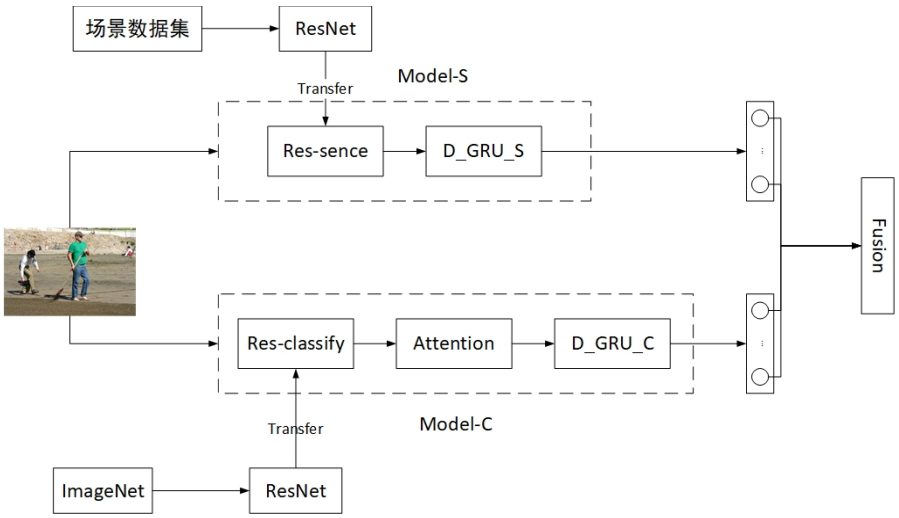

Figure 2. Scene image description structure diagram
图 2. 场景图像描述结构图

Res-sence 是用于提取图片场景信息的模型，Res-classify 是用于提取图片中物体信息的模型，两者在结构上使用的都是 ResNet 101 网络结构。区别在于 Res-sence 是在场景数据集 places-365 和 AI challenge 的场景分类模型数据集上预训练的，而 Res-classify 是在 ImageNet 物体分类数据集上训练的。两个模型分别连接双层 GRU 解码器，得到侧重场景信息和侧重物体信息的图像语义表达，最后对概率进行加权平均计算，得到最终的输出。Attention 代表软注意力机制，考虑到本文研究方法中，新闻图像信息的提取需在最大程度上与新闻内容相匹配，因此，本文在使用 Res-classify 模型对非文本图像进行物体信息提取时，融入软注意力机制的思想[12]。Fusion 代表加权融合操作，在 t 时刻 Model-S 和 Model-C 的输出概率值为 p_S^t 和 p_C^t，α_S 和 α_C 表示权重，得到 p_{Fusion}^t 表达式：

$$p_{Fusion}^t = \frac{\alpha_S \cdot p_S^t + \alpha_C \cdot p_C^t}{\alpha_S + \alpha_C}$$

最终模型在 AIC-ICC 数据集(AI challenger 图像描述竞赛数据集)上训练得到本文的场景图像信息提取模型，AIC-ICC 数据集中数据分布表 1 所示。

Table 1. AIC-ICC data set distribution table
表 1. AIC-ICC 数据集分布表

数据集	训练集	测试集	验证集
AIC-ICC	210000	30000	30000

图像描述常用的客观评价指标有 BLEU，CIDEr 等，但是考虑到客观评价指标主要是评价模型生成语句与参考语句的相似程度或者是评价生成语句与人类表达的差异性，这与本文的应用场景并不完全符合，本文选择人工评价的方式，具体的是设立四个评价等级，分别是非常好、好、一般、差，具体等级及对应等级描述如表 2 所示。

针对C_I中一副图片i_j，因为i_j可能同时包含图像描述、文本信息以及人脸信息。将i_j经过图像语义描述与信息提取模型进行处理，得到一个三元组$\{\{O_j\},\{ic_j\},\{face_j\}\}$，其中$\{O_j\}$表示这幅图片经过OCR 模型转换得到的文本信息；$\{ic_j\}$表示这幅图片进过图像描述模型处理得到的图像描述；$\{face_j\}$表示这幅图片经过人物信息提取模型处理得到的人物信息。因为考虑到$\{O_j\}$和$\{ic_j\}$是直接与新闻正文相关的文本信息，且本身具有一定的语义信息，我们可以将得到的$\{O_j\}$和$\{ic_j\}$分别与新闻的正文的主题句进行直接的文本相似度计算，可以得到这两个部分与新闻正文的相关度的度量值*sim*1，*sim*2。

对于人物图像信息的使用，本文在对大量“标题党”新闻分析后发现，一些“标题党”新闻往往使用名人、政要的人脸图像吸引用户点击，让用户产生一种这篇新闻是与这个人物有关系的错觉，但是往往新闻的正文文本信息和使用的人物图像相关度很低。针对这个特点，本文考虑对新闻正文进行主题句提取后，对主题句进行命名实体识别操作，识别句子中的人名得到一个包含人名的集合*CH _Theme*，对ch_j和*CH _Theme*进行比较，具体如下：

判断ch_j是否是*CH _Theme*的子集，如果ch_j是*CH _Theme*的子集，则认为这张图片和文章有关联，令*sim*3的值为1；如果*CH _Theme*中并未包含ch_j则sim3的值为0；如果两个集合中有交集，则利用Jaccard相似度计算方法，得到$sim3=\dfrac{|ch_j \cap CH_Theme|}{|ch_j \cup CH_Theme|}$，由此可以得到关于一张图片中的人物信息的相关度的度量*sim*3，且*sim*3的取值范围为[0, 1]。

考虑到一副图片中并不一定全部包含这三类图像信息，故这三个值可能为空值，所以单独将这三个相似度值分别作为图片和文本的相似度度量是不合适的，会导致一幅图片因为某个方面的信息的缺失而被模型认为这张图片与文本的相关度为0。因为*sim*1和*sim*2本身是通过文本相似度计算得到，故本身的取值范围是[0, 1]，而我们从sim3的计算过程也可以发现，*sim*3本身的取值范围也为[0, 1]，我们考虑使用计算三个值的均值，将得到的均值作为一幅图片和文本的相关度的度量值。计算方法如下：

$$sim\left(i_j,C\_T\right)=\left(sim1+sim2+sim3\right)/N$$

其中，*N*为一幅图片中存在的*sim*值的个数，又因为$C_I=\{i_1,i_2,\cdots,i_m\}$，所以

$$sim\left(C\_I,C\_T\right)=\frac{1}{m}\sum_{j=1}^{m}sim\left(i_j,C\_T\right)$$

最终提取的新闻图片特征如表5所示。

Table 5. News image features
表5. 新闻图片特征

符号	解释
Num_Image	新闻中包含图片的数目
CI_CT_Sim	新闻中图片与正文文本的相关度

4. 实验过程及结果分析

4.1. 实验设计与评价标准

4.1.1. 实验方案设计

本文设计了对比实验，验证本文提出的融合图像信息的“标题党”新闻识别方法的有效性。本文

Table 8. Model training features and importance ranking
表 8. 模型训练特征及重要度排序

符号	解释	贡献度
TT_CT_Sim	新闻正文文本和标题文本之间的相似度	1.9211
Have_S_Word	新闻标题中是否包含高频词汇，高频词汇是指有具有吸引眼球效果的词汇	1.4416
CT_CI_Sim	图片文本相关度	1.1455
Have_Pron	新闻标题中是否含有“她”，“他”这类指代词	1.0762
Have_Symbol	新闻标题中是否具有情感倾向性的标点符号	1.0263
CT_Length	新闻正文长度	0.9376
Have_Num	新闻标题中是否包含数量词汇	0.8824
Is_Origin	新闻是否是原创新闻	0.7698
TT_Length	新闻标题的长度	0.6936
Have_Names	新闻标题中是否使用名人姓名	0.6182
News_Tag	新闻标签	0.5928
Nums_Image	新闻中图片的数量	0.5496
Nums_Review	新闻评论数	0.4096
Avg_Sen_Length	正文中平均句子的长度	0.2651

4.3. 模型训练

在模型选择和训练方式上，为了保持单一变量，我们在 baseline 上采用与文献中相同的采样方式和分类模型，即采用有放回随机抽取的采样方式，分类算法使用随机森林分类算法。本文的分类算法模型实现主要使用的是 Scikit-learn 机器学习库，Scikit-learn 是一个开源的机器学习 Python 库，其中集成了包含分类、聚类、回归等众多机器学习算法。为了避免模型训练过拟合同时可以使模型学习到更多的有效信息，本文使用 K 折交叉验证法训练模型避免过拟合，K 值取 10。

4.4. 实验结果

Table 9. Experimental data of “Clickbait” news classification
表 9. “标题党”新闻分类实验数据

Model	精确度	召回率	F_1-score
baseline	0.862	0.842	0.852
Baseline + sim(CI, CT)	0.896	0.878	0.887

从表 9 中我们可以直观的看出，baseline 在同时含有文本信息和图像信息的多模态新闻数据集上精确度是 0.862，召回率是 0.842，F_1-score 是 0.852，这个结果与文献[16]中使用随机森林的最好成绩 0.87 相比，精确度有所下降，导致这个现象的原因应该是数据集差异，本文使用数据集一共是 3356 条，而 baseline 使用的数据集为 808 条，这对最后结果有所影响。而在 baseline 融合入图片特征信息后精确度是 0.896，召回率是 0.878，F_1-score 是 0.887，相对于 baseline，模型识别效果在精确度上有所提高，但是提高的幅度并不是很大，经过分析，本文认为可能有两个方面的原因，第一是因为数据集中部分图片“标题党”新闻中也具有“标题党”的文本特征，这一部分的信息，在忽略图片信息以后，依然可以使用新闻的文本特征去识别，而在加入图像信息之后，模型的检测效果有所提升，被提升出来的这一部分新闻，是仅