2020年大学生创新训练项目

申请书

项目名称 基于模板的微信公众号推送内容自动生成研究

项目负责人 邬语丝 联系电话 18347943868

所在学院 计算机与网络空间安全学院

学 号 201811113003

专业班级 2018计算机科学与技术

指导教师 李春芳

E-mail yusi@cuc.edu.cn

申请日期 2019年11月26日

* 1. 基本情况

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 项目  名称 | | 基于模板分析的微信公众号活动类推送的自动内容生成模型 | | | | | | | | | | |
| 选题  来源 | | □规划项目 一般选题5 新媒体前沿技术改造与实践 | | | | | | | | | | |
| 所属  学科 | | 数据科学 计算机科学与技术 传播学 新闻学 网络与新媒体 | | | | | | | | | | |
| 负责人  姓名 | | 邬语丝 | | | 性别 | 女 | 民族 | 汉 | | 出生年月 | | 2000 年 9 月 |
| 学号 | | 201811113003 | | | 联系  电话 | 18347943868 | | | | | | |
| 项目负责人曾经参与科研的情况 | | |  | | | | | | | | | |
| 指导  教师 | | 李春芳 | | | 联系  电话 | 13269636506 | | | | | | |
| 指导教师介绍 | | | | （1.职称/职务 2.研究方向 3.所在学院 4.相关经历） | | | | | | | | |
| 项  目  组  主  要  成  员 | 姓 名 | | | 学号 | | 专业班级 | | | 所在学院 | | 项目中的分工 | |
| 王泽宇 | | | 201811153023 | | 18数据科学与大数据技术 | | | 计算机与网络空间安全学院 | | 软件开发 | |
| 章颖 | | | 201811153012 | | 2018级数据科学与大数据技术 | | | 计算机与网络空间安全学院 | | 软件开发-前端页面 | |
| 黄婧怡 | | | 201801213019 | | 2018级传播学(媒体市场调查与分析方向) | | | 新闻学院 | | 数据新闻设计 | |
| 谢清扬 | | | 201801113029 | | 2018级新闻学 | | | 新闻学院 | | 新闻模板设计 | |
| 项目组成员参与科研的情况 | | **黄婧怡：参与教育部哲学社会科学研究重大委托项目“高等教育大众化与媒介融合时代菁英女性培养与领导力提升研究”（项目批准号：15JZDW002），参与专著《连接力与粘合力：平台型媒体与全球菁英女性领导力自我呈现研究》的数据分析。**  **谢清扬：调查研究《微信健康推送对特定群体传播影响》，对微信内容传播模板、传播方式有一定研究经验。** | | | | | | | | | | |

* 1. 立项依据（可加页）

|  |
| --- |
| 1. **研究目的** 2. **了解已有机器新闻和自动写作相关技术。**   **通过查询各大团队机器写稿发展情况，对该行业进行全面认知，并对用户市场需求进一步分析。**   1. **实现一个图文、视频融合的内容自动生成系统。**   **依据新闻稿件倒金字塔式结构、螺丝杆式结构、金字塔式结构，分别针对动态类、单一线条内容复杂类、趣味类会议消息进行自动生成。同时依据生成内容及用户自定义风格选择，广泛检索互联网图片与视频，最终完成个性化排版。同时，以“短、平、快”为特点，为内容生产者提供便利，为新媒体内容发展提供创新路径。**   1. **实现微信公众号内容自动推送。**   **通过自动填表单、爬虫及软件测试技术（存疑），在完成内容分析的基础上，达到生成内容与微信公众号发布平台的直接联动。用户可通过复制粘贴等方式，快速便捷地推送消息。**  1-了解现有的机器新闻和自动写作相关技术。  2-实现一个图文、视频融合的自动内容生成系统。  3-实现一个微信公众号的内容自动推送。（自动填表单，爬虫，软件测试）   1. **研究内容**   **1. 需求分析及信息来源**  **运用焦点小组访谈和问卷调查法研究用户对于即时发布内容推送的形式、排版的需求，获取动态需求信息。并通过内容分析法整合特定领域官方公众号平台的会议类推送，获取会议新闻各基本要素；将百度图片、必应图片、百度贴吧、豆瓣、花瓣、堆糖等图片网站作为图片检索信息源，以bilibili、爱奇艺、腾讯视频等主流视频平台作为视频检索信息源，与生成内容结合形成推送整体内容。**  软件与算法设计（网站，实现类似今日头条类的根据海量数据的内容生成）  （1）NLP自动摘要生成；词频分析；爬虫；  （2）图像理解：用于选择配图  （3）视频的摘要生成，找一段扣题的视频。（借助已有成果，人工辅助）   1. 软件实现（Python+Flask+MySQL+D3）   系统的搭建   1. 基于软件测试的自动表单填写，推送公众号（文本、图、视频链接传到公众号里）。   二、针对活动类微信公众号推送内容生产的程序化模型构建  计划实现的功能：  1.文本素材生成：提供推送使用的文本。  （1）使用MySQL建立活动类推送文本模板数据库。  （2）使用Python的NLTK类库和SnowNLP类库，设计活动类推送中关键指标、文本关键字与（1）所建立数据库中文本模板的匹配算法。  2.图像素材生成：提供推送中非用户上传部分的辅助配图。  （1）使用MySQL建立应用于活动类推送的配图数据库。  （2）使用Python的SnowNLP类库和Pillow类库，设计活动类推送中关键指标与（1）所建立数据库中配图的匹配算法，并将图像与文本中段落关键字按顺序进行绑定。  3.图文样式生成：对文本与图片进行排列与样式编辑。  （1）使用Python的文件读取和写入功能，将文本与图像写入HTML文件，并设计编辑图文排列样式（包括字体、字号、配图尺寸、对齐方式、颜色）的简化算法。  （2）使用UEditor富文本web编辑器读取HTML文件，实现图文排版在网页上的展示和自定义编辑。  ~~后期可能实现的功能：~~  1.优化图文样式生成  （1）使用MySQL建立推送排版样式数据库  （2）使用Python的NLTK类库和SnowNLP类库，设计活动类推送中关键指标与数据库中排版样式的匹配算法。  2.功能性网站搭建  使用HTML，PHP，JavaScript设计承载上述功能的网站，允许用户通过网页对自动生成的图文内容进行自定义编辑，并实现（以用户复制粘贴或与微信公众平台建立接口的方式）将图文内容添加到微信公众号平台。   1. **国、内外研究现状和发展动态**   随着人工智能技术的不断发展进步，其应用也逐渐渗透到日常生活中的方方面面。其中，“机器新闻写作”就是其在新闻传播领域的重要应用。2016年里约奥运会上，由北大计算机所万小军团队研发的“张小明”新闻机器人横空出世，13天内，撰写了457篇关于乒乓球、羽毛球、网球等项目的消息简讯和赛事报道，一时间吸引了世人的眼球；而实际上，在“张小明”出现之前，新闻机器人就已经被不少世界上的主流媒体关注并加以使用——    贴图  其实，早在2006年，美国商业信息集团汤姆森金融就已经开始运用电脑程序自动生成部分财经新闻，可在上市公司公布业绩短短0.3秒内就发布出一篇盈利报道。这也被看作是“机器新闻写作”的初探，只可惜，在更加注重质量而非速度的报纸时代，它并未引起太大关注。  在那之后，西北大学的科技精英们研制出了新一代智能写作软件Narrative Science，这时的机器写作已经不仅仅局限于对数据的呈现，还已经具有了强大的运算能力，展示出一篇具有“人情味”的新闻报道——此时，已经具有了如今“机器写作”的雏形。  随着社会的飞速发展，传统的报纸、电视等传播媒介日渐衰微，数字化、碎片化的阅读成为人们最主要的信息来源。基于人们这样的需求，新一代“新闻机器人”应运而生，在不断的改进和充实中，逐渐满足了人们对新闻报道的各种需求——如今的“新闻机器人”，已经可以结合最新的自然语言处理、机器学习、大数据分析和视觉图像处理技术，通过语法合成与排序，基于现有的文本模板生成完整的新闻——此处所说的“完整”，不仅仅是指可以生成一篇完整的文本，更可以通过图片检索自己选择合适的配图，甚至模仿人类的语气，使新闻文本读起来更加亲切、抓人眼球。  如今，新一代的机器写作的应用更加广泛。由最开始的集中于数据化、模板化的财经新闻和体育新闻报导为主，到现在，已经可以根据数据的宏观对比，得出变化、趋势等结论并加以呈现；同时，同一则新闻可以生产出多种个性化的版本。而国外由于机器写作研究的起步较早，拥有相对多的技术优势和实战经验，除了能够实现国内的应用外，还可以在房地产、商业等更多领域得到使用，对数据的追踪、复杂信息的消化、关系的挖掘等方面也要比国内更加完备与成熟，甚至还可以写出极具个性化的新闻报道。  不过，由于相关技术仍然不够完备，当下的机器写作依然有一系列问题亟待解决，比如机器人对信息的理解深度还不够、新闻成品模板化明显，扁平而又千篇一律的文章依旧是绝大多数机器写作产品的常态，缺乏亮点；对信息的提炼和概括能力也还有所不足。这些也是研究人员们接下来将要致力于解决的方向。  一言以蔽之，当前，“新闻机器人”是目前机器写作在我们日常生活中最为常见的应用，且已经证明了其高效性和可信赖性，但由于相关技术还不够成熟，其依旧有十分广阔的发展前景，有望在未来成为跨领域的多面手、人类记者编辑的好帮手，而本课题团队也想要以目前已经比较完备的基于模板生成完整新闻这一技术为基础，完成活动类微信公众号推文的实现，提高生产推文的效率，使机器写作这一技术在新媒体领域得到更加普遍的应用。  **参考文献**  **专利**  **网页**  **爱奇艺视频**   1. **创新点与项目特色**   **3条:内容生成；图文视频融合媒体；自动填充**   1. **探索智能生成内容的新路径**   **智能生成内容发展虽有相关路径可循，但针对会议类报道信息生产的研究较少，此次探索是一次合理尝试。基于传统新闻写作结构，从用户范围广泛的微信公众号入手，融合为新媒体用户所接受的阅读模式。并以会议类内容生成为核心，探索会议类新闻内容模板，开发“短平快”的信息写稿路径。**   1. **媒介融合实践，打造图文视频融合媒体。**   **此次研究计划在自动形成文案的基础上，依据生成内容进行全网智能检索，为推文匹配合适的图片与视频，实现推文文字、图片、视频三合一，为受众带来更优越的视听体验。**   1. **自动填充**   **依据新闻稿结构，形成内容模板。用户可选择适当的结构并输入新闻要素，系统自动生成新闻稿件，提高内容生成效率，降低人才培训成本。**   1. **技术路线、拟解决的问题及预期成果** 2. **技术路线** 3. **拟解决的问题** 4. **预期成果** 5. 针对活动类微信公众号推送的模板分析   技术路线：  问卷调查、内容分析及焦点小组访谈  问卷调查法指的是将需要调查的具有公共性的内容分解为各个要点，通过选择题、填空题、排序题、量表题等多种形式指向最终结果。问卷制作需要经过多次实验和修改，以达到最佳的数据效果。研究小组拟运用问卷调查法找到吸引人的活动类推送的篇幅、风格、形式等等。  内容分析法指在一定的范畴内对活动类推送包含的各个方面进行分类，并确定这些信息单元在这一范围内出现的频次，对各篇推送的差异性和共同性进行定量分析，以观察、分析某篇活动类推送的可能出现的样式的一种方法。  焦点小组访谈法指的是一组人按照规定的程序有序地提供信息，主持人引导一组参与者进行一种开放的和深入的讨论，而不是简单地提出问题并征得参与者回答。研究小组拟运用焦点小组访谈法，围绕已搜集的推送资料挖掘其共同性及人工制作此类推送的习惯。  拟解决的问题：  已有的研究活动类微信公众号推送的书籍和资料近乎没有，打算先学习更深入的搜集资料的方法，以及向有类似研究经验的老师请教。  画图  1-设计模板3套：体育；影视；经济；IT  2-软件设计、算法  3-软件实现  4-自动表单  5-人工审核  写一段文字论述你的图   1. 针对活动类微信公众号推送内容生产的程序化模型构建   技术路线：  前期利用Python爬取数据，结合MySQL建立数据库；再通过Python的NLTK库和处理中文的第三方类库（例如SnowNLP等）对文本数据进行处理、分析；由分析的结果总结模板；在基于模板的基础上选择“模板填空模式”或“机器学习/深度学习”的方式实现特定类型推送文章的自然语言生成。若进展顺利，将在上述步骤已实现的基础上进一步实现推文的简单排版功能，届时将借助Python针对HTML文件的操作生成文字+配图的网页展示给用户或利用如百度等公司提供的富文本编辑器UEditor为用户提供实时编辑功能。    拟解决的问题：  将获取的数据分类建立语料库  通过自然语言处理的现有手段对中文文本进行合理的分析并得出有价值的结论  生成有效的文本模板  尝试实现基于模板的自然语言生成  初探利用人工智能技术实现自然语言生成  制作可供使用的由关键信息生成微信推文的程序  预期成果：  1.通过NLP技术对推文进行分析后总结出一定的规律  2.实现基于模板的自然语言生成  3.制作出可以自动生成推文的程序   1. **项目研究进度安排**   立项－2020.1  研读与微信活动类推送有关的文献并搜集至少三百条活动类推送。  2020.2  组建焦点小组，根据搜集的活动类推送编制访谈内容，进行开放、深入的访谈，对所得访谈结果进行梳理。根据访谈结果制作问卷，问卷经修改通过信度效度后，进行问卷的发放、回收及数据分析。  学习自然语言处理相关知识，使用Python的NLTK和SnowNLP等类库，建立短文本处理简单模型。  2020.3－2020.5  在阅读相关文献的基础上，参考已有的严谨研究，编制内容分析法中使用的编码表，并对编码表进行试编码及信效分析，得出最终的编码表；培训编码员，检测编码员间的信度，再对几百条活动类推送进行编码，整理编码所得数据，运用内容分析法得出构成活动类推送的关键指标和建立文本模板所需的语料集。  使用Python爬虫技术与MySQL，建立配图数据库，并设计活动类推送关键指标与配图数据库的匹配算法。  学习设计将图文素材简单排版为HTML文件的Python程序。  学习Ueditor的API文件，实现用户通过Ueditor可视化操作文章排版的功能。  2020.6-2020.9  根据调查得出的活动类推送关键指标与语料集，使用MySQL建立文本模板数据库，并设计关键指标与文本模板数据库的匹配算法。  后端设计基本完成。  2020.10-2020.12  搭建网站，设计网页，并实现前后端联调。   1. **已有基础**    1. **与本项目有关的研究积累和已取得的成绩** 2. 本课题团队已研究今日头条、新华社万小军等机器新闻写稿前例，对基本流程有一定掌握；同时，在新媒体传播形式、内容及影响中有相关的研究经验，例如通过对微信公众号平台内容在一定群体中的传播及影响的研究。 3. 小组成员可以熟练掌握并应用定量分析、采访访谈技巧等研究方法。 4. 能够使用爬虫爬取网页，并对爬取下来的内容进行分词处理、词频分析等，即已经可以抓取文本并且对文本进行最基本的分析操作。 5. 数理统计知识准备充足，已经可以对大量数据进行结构化的归类和统计分析，整理出想要的有用数据信息。    1. **已具备的条件，尚缺少的条件及解决方法**   已具备条件：   1. 掌握MySQL基础知识与操作，能够熟练使用数据库并进行数据库的管理等，比如建立图片库，为文章生成合适的配图； 2. 具备Python编程基础，为对NLTK类库和SnowNLP类库的学习和实际应用做好了准备； 3. 有扎实的数学、概率、数理统计基础，对人工智能领域的基本算法已有初步了解并能够合理应用，为更进一步的研究打好了坚实的理论基础； 4. 初步学习过机器学习相关知识，为日后生成模板文本和完整连贯文章等研究成果提供理论基础；   缺少条件：2条   1. 课题团队目前缺少对市场需求细化的把握、对活动类微信推文大数据资源的掌握，尤指论文资料、课程书籍等； 2. 本课题缺少现有模式借鉴，本次研究也需要针对模式进行； 3. 目前尚缺乏对NLP技术的了解，暂时难以通过自然语言处理的现有手段对文本进行分析； 4. 本小组缺少人力资源，难以进行大数据量的内容分析及编码工作； 5. 对用户的具体需求了解还有所不足。   解决方法：2条   1. 针对目前问题，团队将通过市场调查分析及模式探索创新解决已有问题； 2. 招募编码员加入团队，共同完成编码及内容分析，完善内容模板； 3. 学习NLP自然语言处理技术、进一步深化机器学习领域的探索和研究，以解决生成文本模板、自动生成连贯的文章、完成完整排版的技术； 4. 采用问卷、当面采访等方式，了解用户需求，以实时更新团队的具体研究方向，呈现出符合用户需求的研究成果。 |

* 1. 经费预算

|  |  |  |
| --- | --- | --- |
| 开支科目 | 预算金费（元） | 主要用途 |
|
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| 预算经费总额： |  |  |

* 1. 项目组成员承诺

|  |
| --- |
| 我们承诺所填报的内容真实有效。如果获得立项资助，我与本项目组成员将严格遵守学校有关规定，认真开展项目，按时报送材料。如违反承诺，将自行承担后果。  承诺人（全体成员）签名：  年 月 日 |

* 1. 指导教师意见

|  |
| --- |
| **导师（签章）：**  **年 月 日** |

* 1. 大学生创新训练计划专家组意见

|  |
| --- |
| **负责人（签章）：**  **年 月 日** |