Statistics and Application 统计学与应用, 2020, 9(4), 525-532
Published Online August 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.94056

Hans汉斯

Analysis of Influencing Factors of Documentary Communication Based on Text Mining

Ziyu Huang, Jie Meng*

School of Mathematics and Statistics, Yunnan University, Kunming Yunnan
Email: 1098147472@qq.com, *691669246@qq.com

Received: Jul. 9th, 2020; accepted: Jul. 23rd, 2020; published: Jul. 30th, 2020

Abstract

After cleaning the basic influence factors as classification variables obtained from the descriptive statistics, in order to understand the factors affecting the spread of the documentary, the crawled data resources get more accuracy of Chinese word segmentation results using Python. Writing Gibbs algorithm analysis LDA model is set up to get a different degree of confusion and theme of the documentary for number and semantic network, such as word cloud image results. Through the analysis of the text, it is found that people's values and the dissemination of documentaries can show a two-way influence. Documentary filming will be based on the mainstream and demand of The Times, which is also the most direct display of people's values. In addition, people can understand what the current world shows to people by watching documentaries, which is also a guiding direction for people's thinking and has a great influence on people's practical needs, the carrying form of documentaries and the authenticity of documentaries.

Keywords

Dynamic Web Crawler, Chinese Word Segmentation, Gibbs Algorithm, The LDA Model, Semantic Analysis

基于文本挖掘的纪录片传播影响因素分析

黄梓玉，孟　捷*

云南大学数学与统计学院，云南　昆明
Email: 1098147472@qq.com, *691669246@qq.com

收稿日期：2020年7月9日；录用日期：2020年7月23日；发布日期：2020年7月30日

*通讯作者。

文章引用: 黄梓玉, 孟捷. 基于文本挖掘的纪录片传播影响因素分析[J]. 统计学与应用, 2020, 9(4): 525-532.
DOI: 10.12677/sa.2020.94056

摘　要

为了解影响纪录片传播的因素，将爬虫得到数据资源清洗后，经描述统计获得基础影响因素作为分类变量，使用Python得到较精准的中文分词结果，编写Gibbs算法建立LDA模型来进行分析得到不同纪录片适合的困惑度和主题数以及语义网络、词云图等结果。通过分析文本的结果得知人们的价值观与纪录片的传播可展现双向的影响作用，纪录片的拍摄会根据时代的主流与需求进行拍摄，而时代的主流与需求也是人们的价值观最直接的展现；再者，人们通过观看纪录片来了解当下的世界展现给人们的面目，同样这也是一种对人们思维的引领方向，影响较大的为人们的现实需求、纪录片的承载形式以及纪录片的真实性。

关键词

动态网页爬虫，中文分词，Gibbs算法，LDA模型，语义分析

1. 引言

纪录片是记录真实生活后通过特定加工形式及表现手法而体现的生活写照。纪录片市场也逐渐成熟，更多优秀的纪录片从国内走向国际，观影方式也与纪录片进行融合也展现了多元化，此次数据来源为腾讯视频纪录片片库，对片库中的纪录片进行影响因素研究。

回顾我国纪录片的而发展历程，中日合拍的纪录片《丝绸之路》较早的展现了此历程的开始，为电视与电影开创了极具影响作用的传播引领作用。国内纪录片极具多样化标准，在其中，题材的多元化选择视角、实际的故事引导性、拍摄手法多样性与真实事件的展现都可体现纪录片在人民群众中的广泛传播。基于中国目前所处于的国际地位的上升，逐渐形成走向国际市场进行竞争从而提升自身的影响程度。何莹莹通过对增加电视纪录片故事性的方法和手段研究来剖析电视纪录片故事化叙事方法[1]。孟庆龙通过对国外纪录片发展史上不同制作模式的“故事化”叙事方法考察，寻找“故事化”叙事方法演变的轨迹[2]。王鹏飞通过对题材的取舍研究来分析题材独特性与普遍性的内在统一[3]。国外纪录片的发展状况较国内而言要早一些，他们提出了纪录片的记录性与纪实性。万燕蓉用各种方法从消费者角度探讨纪录片需求[4]。张芳以受众心理学、纪实美学为基础结合案例调查法等多种科学研究方法对纪录片创作进行分析[5]。李琰从翻译中的归化和异化策略角度研究纪录片中字幕翻译引起的多种问题[6]。以上的分析大多源自文献资料的总结分析整理所得，而对纪录片中评论内容做相关性的研究。因此本文将评论内容语义集建立文本挖掘模型来及逆行相关文本类研究。

2. 数据处理

2.1. 数据来源

在本文中，利用网络爬虫，获取纪录片的相关信息，如片名、播放量、上映时间、评分、标签、简介、短评总数和短评详情等。首先提取全部纪录片的 URL 标签，再根据已获得的各纪录片 URL 对纪录

片详情进行分析提取，使用 selenium 库中的 webdriver 进行驱动浏览器，模拟用户正常上网行为，通过这样来获取异步加载后的评论信息。根据影片的 id 号通过抓包的方式获得短评内容及下一页开头评论 id 号，以此循环至全部抓取完毕或达到满足分析使用的标准。获得详细数据如下表(表 1)所示。将获得到的影片详细信息进行数据清洗及筛选工作，获得用于分析的最终数据。

Table 1. Sample content for the documentary has been downloaded
表 1. 纪录片已下载内容示例

片名	播放量	上映时间	评分	标签	短评总数	短评详情
宵夜江湖美食纯享	884.7 万	2019 8.20 发布	7.6	内地 2019 美食	5	1.真要命的视频，放在嘴里的感觉和幸福有关 2.我去，人家用竹子轩，他们用铁，吃了致癌 3.烤串位置在哪儿呢想去 4.不能合到一起放嘛。 5.羊肉串吗？？？？？？

由上表可以看出，下载得到的纪录片相关信息较为完整，查看所有的纪录片文本可以发现下载的数据完整性较高，其中包括在播的纪录片以及已下架的纪录片，选择将数据进行清洗，获得可进行分析的文本。

2.2. 数据清洗与基础分类

通过爬虫获得的纪录片数据需进行简单清洗，清洗的原因主要在于部分纪录片存在下架的情况、纪录片涉及政治因素或是纪录片带有强烈的多方因素影响，在多种情形的作用下，就会有纪录片内容缺失、评论被屏蔽或者开启评论筛选的多种选择，此类情况的评论内容受到版权限制，无法收集得到。故选择将没有评论内容的数据进行剔除处理以调整具备评论内容的数据的平衡。再通过 Excel 进行繁简体转换、空白格替换等方式得到最为基础的清洗数据。

3. 主题挖掘

不要使用空格、制表符设置段落缩进，不要通过连续的回车符(换行符)调整段间距。

3.1. 文本分词

除了一些众所周知的英文缩写，如 IP、CPU、FDA，所有的英文缩写在文中第一次出现时都应该给出其全称。文章标题中尽量避免使用生僻的英文缩写。

3.1.1. 停止词建立

停止词在分词的使用中，具备重要作用，通过建立停止词库，去除无效的词汇，可方便结果实施，若不去除无意义词，则会使关键性情感词的比例下降，所得出的结论就不具备代表性。得到停止词的方法大部分为三种，一种是网络资源进行查找，可以较快的获得停止词库，是比较简便的方法；第二种是自己编写，对于不同的文本数据，对应的无意义词也是不同的，那么在这种情况下，自己编写的停止词就更有针对性，能够更有效率的获得准确度高的分词结果；第三种一般为编程软件自带的停止词库，可调用，同时也可以通过导出来格式来方便下次使用。例如 R 语言中：write.table (stopwordsCN(x), ‘filename’)，就可以实现将 R 语言中内置的停止词写成文件格式导出，但这种方式获得的停止词只有 519 个，远远不适用于普遍使用。就一般使用而言，基本的停止词量是在 1400 左右，那么这种情况就需要在已有的停止词的基础上，进行补充来确保此次所使用的停止词库的完整性。

3.1.2. 中文分词

本文通过 Python 来进行分词，在进行分析时，同时需要导入各种库的使用，进行中文分词功能时，使用最多的就是 jieba 语库，主要通过 pandas、xlwt 这两个库来驱动实施，Python 实现的分词并没有分离开表情符号等文字，而是最大程度的保留了原始文本以确保分词的完整性，在通过自定义函数去掉停止词、除去其中包含的数字和字母，最终得到的结果分别存入 csv 以及 txt 两种格式备用，查看分词内容时，若出现其余无效字符，可以手动增加停止词文件的内容，并再次剔除，反复进行，最终可得到一份较为适合的分词结果，经过多次分词结果的比对，选择先通过 Python 中 jieba 语库进行语义分词，再导入 ROST CM6 中进行情感分析以及聊天分析，获得多种分词结果。进行查看后，发现所保留的内容有效成分高达 98%，更优于匹配率 97%，那么此份数据便成为进行文本模型构建的依据。

3.2. LDA 模型概况

3.2.1. LDA 简述

LDA 的全称是隐含狄利克雷分布(Latent Dirichlet Allocation)，一种文档主题生成模型。在机器学习中，即为线性判别，其主要功能为降维与分类[7]。若是进行文本主题类的则需要进行具备三层贝叶斯结构的文档主题模型，其内容由词、主题以及文档组成。

LDA 是基于贝叶斯模型的，而贝叶斯模型最为主要的三个关键地方就是先验分布、似然估计以及后验分布。贝叶斯推断的含义在于先预估数据的先验概率，再将其置于实验，判断实验结果对先验概率的影响是加强或是削减[8]。那么在最初的朴素贝叶斯的思想中，根据条件独立公式获得条件概率公式，并以此推得贝叶斯公式：$P(Y_k \mid X)=\dfrac{P(X \mid Y_k)P(Y_k)}{\sum_k P(X \mid Y=Y_k)P(Y_k)}$。引申获得三层的贝叶斯模型，在这其中，实现文档到主题服从多项式分布，主题到词服从多项式分布。

在贝叶斯的推断中，狄利克雷分布作为多项分布的共轭先验得到较多的应用，可运用到 LDA 模型中。在二维的分布中，可使用二项分布和 Beta 分布进行表达，类推可以获得在三维情形中，可以用三维的 Beta 分布来表达先验后验分布、三项的多项分布来表示具备似然情况的数据。

在三维的多项分布中可以写成：$\text{multi}(x_1,x_2,x_3 \mid n,p_1,p_2,p_3)=\dfrac{n!}{x_1!x_2!x_{3!}}p_1^{x_1}p_2^{x_2}p_3^{x_3}$

同时，通过二维的 Beta 分布可类推获得三维的 Dirichlet 分布，形式如下：

$$\text{Dirichlet}(p_1,p_2,p_3 \mid \alpha_1,\alpha_2,\alpha_3)=\frac{\Gamma(\alpha_1+\alpha_2+\alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}(p_1)^{\alpha_1-1}(p_2)^{\alpha_2-1}(p_3)^{\alpha_3-1}$$

通过类推，可以获得 K 维 Dirichlet 分布表达式：$\text{Dirichlet}(\vec{p} \mid \vec{\alpha})=\dfrac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)}\prod_{k=1}^K p_k^{\alpha_k-1}$；

并通过两者共轭得到与二项分布相同的结论：$\text{Dirichlet}(\vec{p} \mid \vec{\alpha})+\text{Multi}(\vec{x})=\text{Dirichlet}(\vec{p} \mid \vec{\alpha}+\vec{x})$。

3.2.2. 主题模型

LDA 主题模型的建立同样是通过先验概率+似然=后验概率的基础公式可以得到。通过已知内容进行潜在狄利克雷分布的主题挖掘[9]。形似类似于上一小节中所提到的模型表示，通过流程图(如图 1)展示的为一个文档模型的基础分布，D 表示文档，数量较多，此处表示文档集合；$\vec{\alpha}$ 表示分布的比例参数，是维数为 K 维的向量，K 为主题数量；$\theta_d=\text{Dirichlet}(\vec{\alpha})$ 为每个文档中主题所占的比例；$Z_{d,n}=\text{multi}(\theta_d)$ 表

示每一主题的比例赋值；$W_{d,n}=\text{multi}\left(\beta_{Z_{d,n}}\right)$ 表示文档中观察到的词；β_k 表示主题其中的值，同时服从狄利克雷分布，多主题的主题也能够表示为 $\beta_{K,n}$。

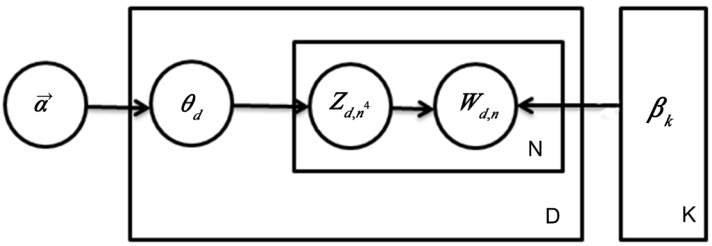

Figure 1. Subject flow graph
图 1. 主题流程图

假设在一类文档中，文档主题数有 N 个，文档主题数 N 就符合狄利克雷分布，也可形成了由 $\vec{\alpha}\to\theta_d\to Z_{d,n}$ 组成的狄利克雷分布多项共轭可推得 θ_d 得后验分布为：$\text{Dirichlet}\left(\theta_d\mid\vec{\alpha}+\overrightarrow{n_d}\right)$，$\overrightarrow{n_d}$ 为在第 d 个文档中所包含的词数，则就是文档主题的后验分布。同样道理，由主题数为 K 的主题元素形成的 $\beta_{k,n}\to\overrightarrow{W_{d,n}}$ 可形成狄利克雷多项共轭推得 $\beta_{k,n}$ 的后验分布为：$\text{Dirichlet}\left(\beta_k\mid\overrightarrow{\beta_{k,n}}+\overrightarrow{n_k}\right)$，$\overrightarrow{n_k}$ 为第 k 个主题中的次数，则产生了主题词分布的后验分布。通过推导所获得的公式，可形成初步的 LDA 思想，为后续的实施方法做更完全的准备。求解 LDA 主题分布的方法一般有两种，一种为 Gibbs 采样，另一种为变分推断 EM 算法，在本文中，则采取 Gibbs 采样的方法。

3.3. 主题数选取与困惑度

该如何选择主题数在 LDA 中一直是比较关键的问题，主题数目选的过于少的话，就会导致各个主题的信息过于简洁从而损失更多关键词；若主题数足够多过于全面的话，就会产生信息量的冗余，那么在信息的选择上难免会有相似的词产生冲突。这两种情况，无论是哪一种，在情感词分析的范畴中都会导致信息量的不平衡。那么该如何确定主题数量方法也是与计算量的直接体现，所以就选择的方法来讲，第一种方法是通过分类来做到肉眼的识别，这是比较直接的方法，但是该怎样分类以及怎么做到比较全面是一个比较困难的事情，若词汇量过于庞大，采取这种方法得到的主题就会很艰难，同时也可能会产生较大误差；第二种方法是按照一些特定的指标，例如 Perplexity 或者 MPI—score 这两种指标值，用处较多的是 Perplexity 这种方法，主要的查看方式就是根据主题数与困惑度的折线图进行查看，重点在于图中的拐点，这就需要进行代码编程；或者按照公式也可计算出困惑度。

3.3.1. 困惑度概念

困惑度的基础公式为：$P\left(\tilde{W}\mid M\right)=\prod_{m=1}^{M}p\left(\widetilde{W_m}\mid M\right)^{-\frac{1}{N}}=\exp-\dfrac{\sum_{m=1}^{M}\log p\left(\widetilde{W_m}\mid M\right)}{\sum_{m=1}^{M}N_m}$；

困惑度是用来度量一个概率分布或者一个概率模型拟合优劣的评判指标，在之后的计算当中，也产生了针对不同情形的困惑度公式。

1) 概率分布的困惑度

在定义相关概率分布的困惑度时，常用的公式为：$2^{H(p)}=2^{-\sum_x p(x)\log_2 p(x)}$。

其中，$H(p)$是概率分布 p 的熵，x 是样本点。$p(x)=\dfrac{n}{N}$。在某些特定的分布情况下，主题数与困惑度的倒数存在一定的关联；在普遍意义下，困惑度是信息熵的指数。

2) 概率模型的困惑度

分布可以构建模型，模型可以将分布可视化，那么用概率模型 $q(x)$ 来估计真实概率分布 $p(x)$ 的参数，并得到具体的模型诸值，这时需要的困惑度公式可以写作：

$$b^{-\frac{1}{N}\sum_{i=1}^{N}\log_b q(x_i)},\ H(\hat{p},q)=-\sum_{x}\hat{p}(x)\log_2 q(x)$$

在一般情况下，b 的取值为 2，$H(\hat{p},q)$ 称为交叉熵，$p(x)=\frac{n}{N}$。

3) 分词困惑度

分词的情况主要用于在分词后的句子位置的概率分布，大概可视作词语在句子上的特定位置出现的概率。其中可能包括多元模型会增加计算量，或涉及语法模型。

3.3.2. 困惑度与主题

通过 Python 来实现困惑度－主题数的分布图。主要是基于信息理论来求的某一主题对应熵的能量[10] [11]。以一个主题为基准计算其困惑度，再逐一递增直到找到困惑度较合适时相对应的主题数。可使用公式：

$$\text{Perplexity}(A)=\exp\left\{\frac{-\sum_{d=1}^{M}\lg P(W_d)}{\sum_{d=1}^{M}N_d}\right\}。$$

通过 python 进行实现，可以得到相应的主题和困惑度相关图，由于纪录片种类繁多，类型齐全，经过筛选，再经过重新合并在经过筛选，最终得出真人秀、时尚与科学证实这三种类型的纪录片在众多类型中更受欢迎。

3.4. 结果分析

根据得到的特征词分布，结合其中的比例分配，可以看出积极网络用语的比例为 65%，其中消极情绪的 35%，而消极情绪的大部分都与时尚类别相关，相关词语多以“看不懂”为主。具体主题－特征词分类下表(表 2)所示。

Table 2. Theme-Feature words
表 2. 主题－特征词

Topic	主题词	特征词
topic1	恐怖电影	僵尸、真的、\u2006、电影、喜欢、蛇、行尸走肉、美剧、发生、好看
topic2	多种类型	真的、僵尸、电影、美剧、喜欢、科学、好奇、\u2006、节目、宇宙
topic3	演技	干、假、挖出、湿、泥土、演技、东西、太、感觉、厉害
topic4	自然求生	鱼、鲨鱼、求生、吐、想、好看、大自然、喜欢、太、生活
topic5	节目效果	完、改、蓝色、想、效果、节目、粗糙、银色、老板、干活
topic6	芭蕾节目	袖珍、🐼、节目、说、芭蕾、🐠、挑选、想、片子、😳
topic7	刺激好看	更新、鱼、喜欢、刺激、好看、想、紧、做人、真的、吃
topic8	魔术假	魔术、假、真的、哈哈哈、想、说、逼、厉害、太、神奇
topic9	减肥养生	减肥、运动、想、吃、报名、针灸、广告、太、节目、节食

Continued

topic10	节目假	假、节目、不错、两个、真实性、怀疑、疑点、金属、逗比、☺
topic11	鞋子	\u2006、收藏、&#、鞋、沙发、鞋子、更新、挖掘机、呜呜、国家
topic12	钓鱼	喜欢、金枪鱼、鱼、钓、日本、好样、鱼钩、一条、钓鱼、资源
topic13	魔术假	魔术、假、说、玻璃、真的、逼、鞋子、克里斯、太、🐼
topic14	青春靓丽	穿、真的、青春、当年、选美、想、衣服、容颜、美貌、美丽
topic15	希望好看	哈哈哈、哈、喜欢、克里斯、感觉、好看、第三季、纪录片、希望、挺
topic16	探索科学	宇宙、节目、科学、知识、世界、袖珍、喜欢、探索、东西、洞
topic17	出乎意料好看	出乎意料、挺、好看、一双、买、上档次、人生、骗到、一丝、节目
topic18	青年喜欢玩	害羞、青年、图、终究、玩儿、帅帅帅、记录、喜欢、玩、生活
topic19	高跟鞋	乘客、高跟鞋、穿、☹、首歌、□、喜欢、手机、铃声、好看
topic20	巴铁兄弟	傻、兄弟、巴铁、丢人、脑残、说、东西、一点、欺骗、性格
topic21	出乎意料好看	出乎意料、好看、特别、节目、骗到、☹、一丝、一双、上档次、买
topic22	穿高跟鞋	穿、□、☹、高跟鞋、好看、假、买、这鞋、点、女人
topic23	确定穿高跟鞋	穿、☹、□、高跟鞋、好看、漂亮、美女、真的、女人、肯定
topic24	文明标准	文明、标准、人有、伤风、宣传、底线、败俗、丑陋、有悖、道德

此语义网络图(图 2)中的所展示的占据较多的为减肥、运动一系列有助于健康的词语；而比例第二高的“恐怖”一类具备真实色彩的极限纪录片为主，可以说明当下人们对自身知识的局限性以及对未知事物的好奇。

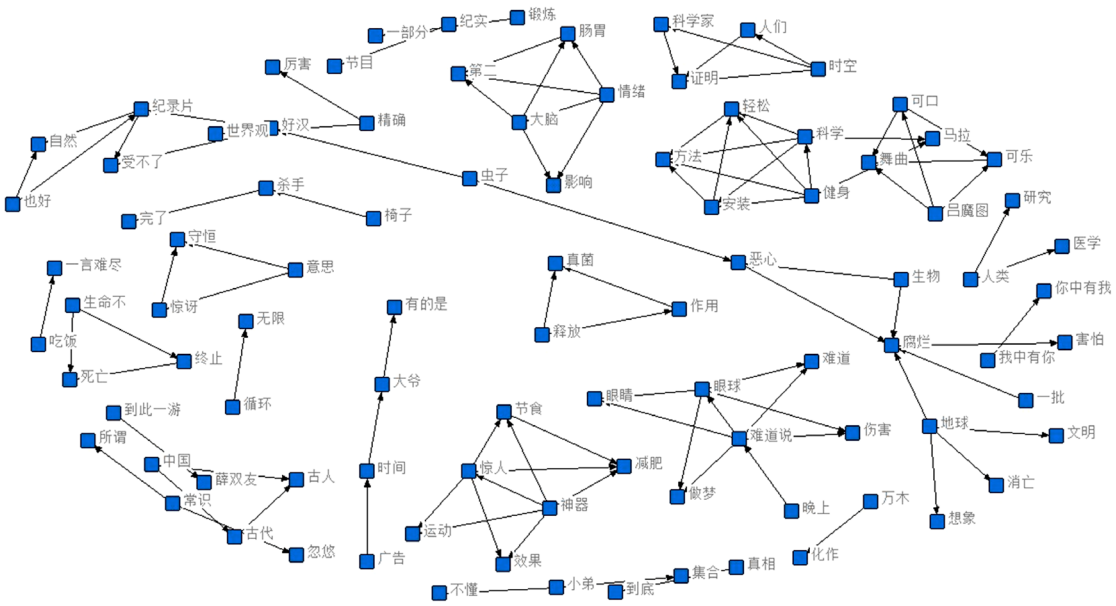

Figure 2. Semantic network graph
图 2. 网络语义图

根据词频所绘制的词云图(图 3)，可以明显看出大部分人对于通过纪录片来记录生活中的美好呈乐观状态。回溯评论内容较具备代表性的纪录片可以发现，能流传时间较长的纪录片一般都是类似于中国历史以及人物传记类具备较强民族特点以及时代意义的影片。

Figure 3. Reality wordcloud graph
图 3. 真人秀词云图

4. 结论

本文通过建立 LDA 模型来进行分析得到影响纪录片传播的因素，通过 Gibbs 算法推进 LDA 模型的实现，并通过困惑度的选择获得较为适合的主题数量，并进行主题分析。通过主题分析可知当下较为受欢迎的纪录片类型是具备科学性、真实性以及可以实现经久不衰特点的这一类影片，此类纪录片不仅与当下的时代形态相关，更与纪录片所包含的历史背景和所输出的价值观相关，是否能做到抓住人们的眼球并分泌多巴胺促进大脑皮层活跃也是影响纪录片能否收到广泛关注的关键因素。而网民对纪录片的态度超过一半呈积极赞同的态度，三分之一数量呈中性态度，而剩余不足 10%的人表现出较强烈的不满情绪，这说明人们的价值观与纪录片的传播可展现双向的影响作用。总体来看，纪录片的影响因素受群众的价值观影响较大，同时也与传播展现形式有着密不可分的关系。

参考文献

[1] 何莹莹. 电视纪录片故事化叙事研究[D]: [硕士学位论文]. 济南: 山东师范大学, 2015.

[2] 孟庆龙. 纪录片“故事化”叙事手法渊源及演进研究[D]: [硕士学位论文]. 扬州: 扬州大学, 2013.

[3] 王鹏飞. 纪录片题材选择研究[D]: [硕士学位论文]. 南京: 南京航空航天大学, 2013.

[4] 万燕蓉. 受众视角下的中国纪录片发展研究[D]: [硕士学位论文]. 济南: 山东大学, 2018.

[5] 张芳. 现实题材纪录片创作研究[D]: [硕士学位论文]. 济南: 山东师范大学, 2017.

[6] 李琰. 从归化和异化角度探讨华语纪录片字幕翻译[D]: [硕士学位论文]. 北京: 首都经济贸易大学, 2017.

[7] 董悦, 王梦. 基于情感分析与 LDA 模型的网络舆情案例研究[J]. 价值工程, 2019, 38(34): 169-172.

[8] 张磊. 基于 C-LDA 的微博推荐算法[D]: [硕士学位论文]. 乌鲁木齐: 新疆大学, 2016.

[9] Masood, M.A., Abbasi, R.A., Maqbool, O., *et al.* (2017) MFS-LDA: A Multi-Feature Space Tag Recommendation Model for Cold Start Problem. *Program*: *Automated Library and Information Systems*, **51**, 218-234. https://doi.org/10.1108/PROG-01-2017-0002

[10] 刘亚姝, 王志海, 侯跃然, 等. 一种基于概率主题模型的恶意代码特征提取方法[J]. 计算机研究与发展, 2019, 56(11): 2339-2348.

[11] 薛佳奇, 杨凡. 基于交叉熵与困惑度的 LDA-SVM 主题研究[J]. 智能计算机与应用, 2019, 9(4): 45-50.