基于模型融合的男频和女频网络小说分析

刘喜平 李艳丽 熊丽媛

（江西财经大学 信息管理学院，江西 南昌 330013）

摘要：网络小说根据目标读者的性别可以分为男频小说和女频小说。由于目标群体不同，男频和女频小说在很多方面具有不同的特征。目前对于男频和女频小说的分析大都停留在定性分析，定量分析很少，总体来说缺乏客观性和数据支撑。本文利用机器学习的手段来对男频、女频小说文本进行分析。我们构造了两个数据集，分别对应男频、女频网络小说。对于每一部作品分别抽取计量风格学特征、小说标题特征和小说文本 LDA 主题特征，从三个方面分别建立分类模型，最后进行模型融合，由此来发现男频、女频网络小说的的差异。研究结果表明男频小说和女频小说可以被自动区分，它们的差异主要体现在一些关键特征上，这些特征也可以解释围绕网络小说的一些社会现象。

关键词：男频小说;女频小说；模型融合；文本分类;网络文学

Analysis of Male and Female Online Novels Based on Model Fusion

Abstract：Online novels can be divided into male novels and female novels according to the gender of target readers. Because of the different target groups, male and female novels have different characteristics in many aspects. Existing study of male and female novels is mostly focused on qualitative analysis, and lacks objectivity and support from real data in general. This paper uses machine learning techniques to analyze the text of male and female novels. Two data sets are built, consisting of male and female novels respectively. For each novel, we extract the stylometric features of the novel, features about the title of the novel, and features about the LDA topics of the novel, and build three classification models with different sets of features, respectively. Finally, we fuse the models to classify male and female novels, and reveal their difference. The results show that male and female novels can be automatically classified, and their differences mainly lie in some key features, which can also explain some social phenomena about online novels.

Key words： Male novels; Female novels; Model fusion; Text classification; Online novels

1 引言

根据中国互联网络信息中心发布的第 43 次《中国互联网络发展状况统计报告》[1]，截止到 2018 年 12 月，网络文学受众用户达到 4.32 亿，网民使用率超过 50%，各项指标均逼近网络游戏。而随着网络文学的爆炸式增长，各类读者的阅读偏好愈加明显，继而出现了很多类别的网络文学作品，如根据性别划分的男频小说和女频小说。通俗来说，“女频”就是女生频道，写作风格和视角偏向女性，多为言情类，读者七成为女性；所谓男频就是以男性为目标读者群的网络小说，如修真、玄幻类。

目前，除了综合类网络文学网站有男频、女频类别和标签外，还出现了很多专注于某一类读者群的网站，如男频网站排名第一的起点中文网、女频热门网站起点女生网，以及人气女频网站晋江文学城。然而，在熟悉网络文学的人看来，同样是文学网站，它们之间却有着显著的差异，如“晋江出 IP，

起点出大神”，起点女频在同类网站中名不见经传，而专注于男频的起点中文网却是如日中天。

除此之外，近几年掀起的 IP 热更是让网络文学以强势的姿态入侵影视荧屏。然而从票房收视率等数据来看，男频和女频网络小说改编的影视作品收视存在较大差异，女频 IP 频频爆热，男频 IP 却屡屡爆冷。近日，最新公布的 2018 网络作家富豪榜[①]也引起了广泛关注，上榜者均为男频作家，而且历年网络作家富豪榜皆是如此，这一现象令很多人疑惑。

不止如此，随着中国网文的出海，中国网络小说揭开了新篇章。据调查，目前在海外流行的中国网络小说类型差异明显，在欧美主要是仙侠、玄幻和魔幻，在东南亚则是言情和都市[2]。由此可见，男频小说在欧美更受欢迎，而东南亚则更偏爱女频小说。

当然，这些网络文学带来的差异不排除有平台本身、地域文化和传播渠道等因素的影响。王浩从“传统美学”的角度对此进行了专门的探讨，他在文献[3]中指出男性在自我崇拜的同时，似乎显得很“传统”，主要表现为大男子式的英雄意识，强者意识。女性所创作的网络文学，一个明显的特征是个体性别意识空前高涨。陈晓华以“晋江文学城”为首的言情网络社区的女性群体作为研究对象，做了一次全面和深入的观照。研究发现，这些女性群体在传播与接受罗曼史小说方面呈现出一种跨媒介、多元化的特征，形成了女性自我展示、自我创造、自我传播的文化现象[4]。二十世纪法国最有影响力的女性主义理论家西蒙娜·德·波伏瓦在其著作《第二性》中提出[5]：“女人不是天生的，而是后天形成的。”换言之，性别并不仅仅是一种生理的属性，同时也是具有社会性的。陈熙熙[6]有类似的观点，她提出性别是在特定社会文化的影响下所形成的，标志着人在文化和心理上的明显差异。而且她强调这种差异性在文学创作领域鲜明地表现为女性创作主体对语言有意识地选择和使用。

在网络小说的挖掘方面，姜崇等[7]从隐藏价值量和热度两方面对网络小说的价值进行预测分析，为企业出版商正确判断网络小说商业价值提供依据。林钊生等[8]研究了网络小说推荐的问题，提出了将基于内容的推荐算法与结合标签排行的协同过滤融合算法相结合的混合推荐算法。李艳丽等人[9]分析了小说的质量，研究发现，一些简单的特征可以区分优秀小说和一般小说。

从上述学者的分析可以看出，大多数研究者都将重心放在对文本文学性的定性分析上，定量研究较少，特别是对文本内容的分析较少。本研究属于文学和数据挖掘的交叉研究，利用自然语言处理和机器学习的方法和手段，分别对男频和女频小说从总体统计量、标题和内容等方面进行深入分析，从而揭示男频和女频小说的关键差异。本文的研究成果可以为文学研究提供参考和依据，也可以用于网络小说社区管理，有助于实现小说的自动分类和标签，并进一步用于小说的自动推荐。

2 数据集

本文使用的语料库由两类网络小说组成：男频网络小说和女频网络小说。语料库中包括具有代表性的 250 本完结小说，其中 200 本小说（男频、女频各 100 本）作为训练集，50 本小说（男频、女频各 25 本）作为测试集，具体采集情况如下：

男频网络小说：搜集速途研究院发布的 2017[②]、2018[③]最受欢迎的男频网络小说排名前 50 的网络小说作品以及中国主流网络文学网站如起点、纵横等热门男频网站点击、收藏、下载各大榜单排名靠前的网络小说作品，还包括被影视化、动漫改编、海外广泛传播的热度高的男频网络小说作品。

女频网络小说：搜集速途研究院发布的 2017、2018 最受欢迎的女频网络小说排名前 50 的网络小说作品以及晋江、潇湘等热门女频网站点击、收藏、下载各大榜单排名靠前的网络小说作品，还包括热门影视 IP、动漫改编的女频网络小说作品。

[①] http://www.sohu.com/a/233257779_119746

[②] http://www.sootoo.com/content/673962.shtml

[③] http://city.huanqiu.com/csxx/2018-12/13831486.html?qq-pf-to=pcqq.group&agt=15438

上述语料库大小近 10 亿字，作品门类涵盖了古言、现言、穿越、重生、青春、都市、悬疑、科幻、游戏、军事、修真、仙侠等各个门类，且在各个门类上分布比较均匀，关键统计数据如表 1 所示。

可以发现，男频网络小说的平均字符数、平均词数以及平均章节数都要远远大于女频网络小说，包含的门类也更广。

表 1 统计数据

	男频小说	女频小说
样本数/本	125	125
平均字符数	5,098,305	2,242,238
平均词数	3,351,553	1,513,486
平均章节数	1,848	1,428
包含门类	9	6

3 特征抽取

本文从多个角度提取了男频、女频网络小说的相关特征，并进行分析。

首先是计量风格学特征。从计量风格学的角度考虑，可以参考的典型特征有：词法特征，如单词长度、句子长度、词汇丰富度、词频等；句法特征，如词性、句子结构等；结构特征，如词汇分布等。这些特征在某种程度上反映出了一个作品的风格，而男频和女频小说由于目标群体不同，可能具有不同的风格，因此我们考虑了这些计量风格学特征，如表 2.1 所示。

其次是小说标题。对于广大读者来说，了解一本网络小说一般都是从书名开始的，从文献[10]的研究可以看出，小说的标题也具有较大的信息量。但是小说标题很短，无法用常规的文本分析手段来进行分析，为此，本文仅从词或字的角度进行分析，即基于词向量或者字向量将标题文本转化为向量表示，再对其进行词频或 TF-IDF 加权处理，以该向量表示作为标题特征。

最后是小说文本。小说的独特性最终还是体现在小说文本内容上，因此对文本内容的分析必不可少。在以前的研究中，对于文本内容的分析主要是定性分析，较为主观。本文认为网络小说作为一种文学样式，具备文学作品的共性：主题，于是考虑从小说文本中提取主题信息。在提取主题信息时，进一步考虑了词性和词频信息，对比了使用不同词汇表生成的主题信息。

表 2 计量风格学特征

特征名称	特征类型	特征名称	特征类型
章节数	连续	地名	连续
字数	连续	机构团体	连续
词数	连续	平均段长度	连续
句数	连续	平均句长度	连续
段落数	连续	平均词长度	连续
不重复词	连续	均章形容词	连续
形容词	连续	均章动词	连续
动词	连续	均章名词	连续
名词	连续	均章副词	连续
副词	连续	人名	连续
方位词	连续		

4 数据分析

本文的研究目的是揭示男频小说和女频小说的关键差异，为此，我们构造男频小说和女频小说的分类模型，根据分类模型来确定对分类有贡献的关键特征，从而对它们进行对比分析。

4.1 基于计量风格学特征的网络小说分类

这一节试图回答这一问题：简单的计量风格学统计特征是否足以区分男频小说和女频小说？我们先分析这些特征，然后基于这些特征建立分类模型，以检验这些特征对于分类的区分能力。

表 2 已经列出了所考虑的计量风格学特征，如章节数、字数、词数等，很明显，这些特征之间存在很多关联。为此，需要进行特征选择。我们用 Pearson 相关系数做相关性分析，然后根据特征重要度排序。

图 1 显示了各个特征的相关性热力图。从图中

可以看出，章节数、字数、词数、句数等汇总量之间呈强相关，相关性系数在 0.8 以上，而这些特征与机构团体数、不重复词呈中等强度相关，与各平均量呈弱相关，而各平均量之间也呈弱相关。由此可以看出，各网络小说作品分布特征鲜明，差异化较大。根据 Pearson 相关系数做特征重要性排序，发现在汇总量中，重要性排名第一的字段是字数。确定字数作为第一特征之后，再验证各特征与字数的相关性，剔除强相关特征。由此，初步选取的特征有：字数、不重复词、机构团体数、平均段长度、平均句长度、平均词长度、均章形容词、均章动词、均章名词、均章副词。

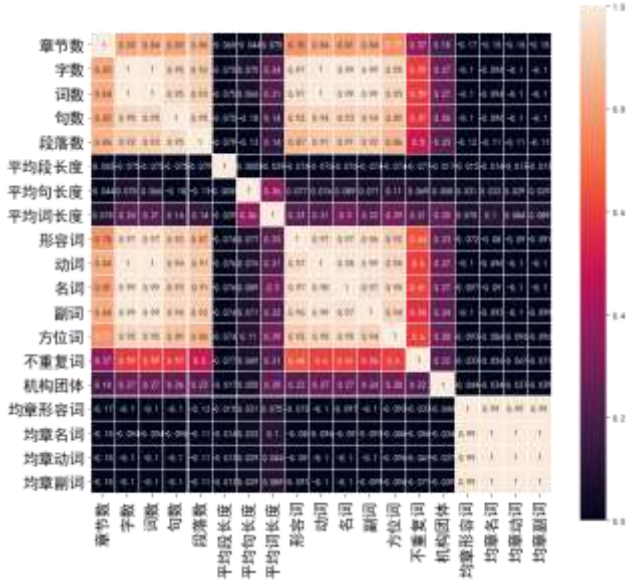

图 1 相关性热力图

下面使用上述特征来进行建模。我们使用贝叶斯模型、CART 模型和 XGBoost 来进行建模。模型以根据相关性筛选出来的特征为输入，以小说的类别标记为输出。实验过程中，决策树采用 10 折交叉验证，10 次 Boosting 试验来增强模型准确性。结果如表 3 所示。可以看出，基于这些简单的频率特征，模型的准确率就可以达到 0.96。与贝叶斯模型、CART 模型相比，XGBoost 模型的表现最好。我们进一步显示了在分类过程中各个特征的重要性，图 2 显示的是按照重要性排序的前 6 个特征，可以看出，在分类中贡献较大的特征有：平均词长度、字数、均章动词、机构团体数、平均句长度、平均段长度、不重复词。

表 3 小说统计特征建模效果评估

	贝叶斯网络	CART	XGBoost
准确率	0.89	0.94	0.96
精准率	0.80	0.89	0.93
召回率	1.00	1.00	1.00
F1 分值	0.88	0.94	0.96

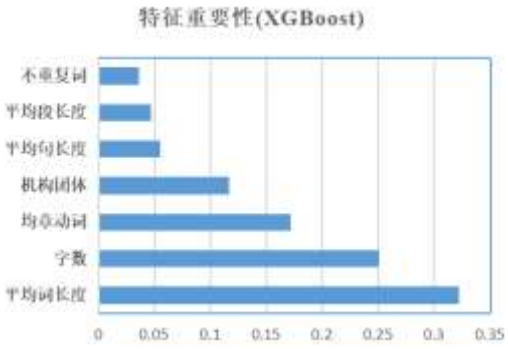

图 2 各模型特征重要性对比

4.2 基于标题的网络小说分类

不同类型的网络小说的标题本身就具有一定的风格差异，因此，本节考虑基于网络小说标题来进行文本分类。网络小说标题非常简短，因此其分类本质是短文本分类。与一般的短文本相比较，网络小说标题比较特殊，表现在网络小说标题用词却变化不一，与传统文本的遣词造句有较大差异。如果用传统的向量空间模型来对标题进行建模，会导致标题向量过于稀疏，无法揭示它们之间的相似性。为此，本文考虑以词向量或者字向量为基础来进行标题的分析。

我们进行了网络小说标题分类的实验。分类的目的是区分男频小说和女频小说的标题，分类使用的特征是词向量或者字向量。本文使用的词向量和字向量来自于预训练好的词向量集合，所有的词向量由 ngram2vec[11]工具包训练。ngram2vec 工具包是 word2vec[12]和 fastText[13]工具包的超集合，支持

抽象上下文特征和模型。训练使用的语料库是中文文学作品语料库[14]，其中包含了 8599 部现代中文小说。为了突出部分特征，我们还对词向量或者字向量进行加权。考虑了两种加权方案：根据词频加权，以及根据 TF-IDF 加权。我们进行了分类实验以比较两种方案。具体来说，我们对小说标题进行分类，用加权后的词向量或者字向量作为特征，分类目标是小说的类别：男频小说或者女频小说。我们将词向量和字向量，以及两种加强方案进行了对比，实验结果如表 4 所示，其中分类模型采用了 SVM[15]。由实验结果可以看出，使用字向量的词频加权方案时，小说标题分类效果最好。

表 4 小说标题 SVM 建模效果对比

方案	准确率	精准率	召回率	F1 值
词向量 (词频加权)	0.76	0.69	0.96	0.80
字向量 (词频加权)	**0.86**	0.82	0.92	**0.87**
词向量 (TF-IDF 加权)	0.82	0.77	0.92	0.84
字向量 (TF-IDF 加权)	0.84	0.90	0.76	0.83

4.3 基于小说文本的网络小说分类

文本分类具有很长的研究历史，目前已经提出了针对各类文本的分类模型和方法，但是小说文本分类比较特殊，主要体现在小说文本的篇幅巨大。例如一个典型的网络小说可以有数百到数千章，字数达到数百万字，内容非常庞杂。基于这样的大型文本构造的向量空间维数过大，使得空间特别稀疏，从而影响分类效果。基于上述考虑，本文用 gensim 包[16]来进行 LDA 主题模型分类。

LDA 主题模型[17]是一种文档生成模型，它认为一篇文档是有多个主题的，而每个主题又对应着不同的词。一篇文档的构造过程，首先是以一定的概率选择某个主题，然后再在这个主题下以一定的概率选出某一个词。不断重复这个过程，就生成了整篇文章。在 LDA 模型中，需要先假定一个主题数目 K，所有的分布就都基于 K 个主题展开。LDA 模型会产生两个分布：一个是每个文档属于不同主题的分布，另一个是每个主题下词的分布。本文主要使用第一个分布。由于不同的男频小说的类型、故事和人物都不相同，每部小说都有自己的主题，因此很难确定主题的数量。我们认为，虽然不同的男频小说的主题各不相同，但是所有的男频小说必然有些相似之处，比如大部分男频小说中都离不开成功、英雄等情节；同理，不同的女频小说虽然具体主题有较大差异，但是也有很多相似之处，例如感情、家庭等。为此，本节将主题数量确定为 2，分别对应男频主题和女频主题，记为 T_{male} 和 T_{female}。最终希望男频小说能够尽可能归到 T_{male}，女频小说能够归到 T_{female}。

模型训练完成后，计算每个文档属于每个主题的概率，如果文档 D 属于 T_{male} 的概率大于属于 T_{female}，则将 D 归入 T_{male} 主题，否则归入 T_{female} 主题。根据分类的结果可以计算分类的准确率。

我们注意到，小说文本特别冗长，词汇表非常庞大，其中包含的大量修饰类和不重要的词汇会干扰主题的抽取和分类。为此，在进行 LDA 建模之前，首先对词汇表进行压缩。考虑依据两个指标来对词汇表进行压缩：

(1) 词性。不同词性的特征项在文本分类中所做的贡献是不一样的。从文学角度来看，一篇文章的核心在于谁、做了什么，而这些要素通常体现在名词和动词上。另一方面，形容词和副词主要是其修饰作用，如“友好”“轻松”“忧伤”等，对揭示小说的主题并没有太大的贡献。因此主要考虑用名词和动词来建模。我们在实验中对比使用了不同词性的结果。

(2) 词频。小说文本中的词语数量过多，使得模型的训练代价太大，为此，可以利用词频来对词语进行筛选。

为了确定词性和词频对于主题分类的影响，我们选择了几种不同的词汇表进行了比较。第一种方

案(V_{all})包含了所有的词汇；第二种方案(V_{infreq})只包含了词频较低的词汇，其出发点是高频词往往是常用词，对主题区分能力不足，因此将高频词过滤掉；第三种方案($V_{n+v+adj}$)只包含了名词、动词和形容词；第四种方案(V_{n+v})只包含了名词和动词。比较结果如表 5 所示。

表 5 基于不同词汇表的 LDA 分类结果

词汇表	词汇表构造方案	准确率
V_{all}	所有词汇	48.8%
V_{infreq}	词频低于 0.5	32.8%
$V_{n+v+adj}$	选择名词、动词、形容词	52.6%
V_{n+v}	选择名词、动词	74%

由表 5 可知，用全部词汇来进行主题提取只能达到 48.8%的准确率，这说明词汇中的噪音太多。选取低频词并不能提高模型准确度，这是因为高频词中有很多区分度高的词，如果完全过滤掉高频词则会很大程度上过滤掉重要的关键词。实验结果表明，根据词性进行筛选可以提高主题分类的准确性，其中仅选择名词和动词的情况下，分类准确率达到了 74%，分类效果最好。

在选择名词和动词的基础上，运行 LDA 模型来进行主题提取。将各主题中的主题词按照概率排序，排名前十的主题词具体如表 6 所示。

表 6 不同主题的主题词

主题	top-10 主题词
T_{male} (男频)	实力 力量 修士 时间 弟子 感觉 声音 世界 长老 强者
T_{female} (女频)	声音 男人 喜欢 女人 说话 孩子 眼睛 感觉 回来 电话

由上述主题词可以看出，男频和女频网络小说主题色彩有明显的区别：男频小说中的主题词更多围绕力量，比较宏大，女频小说则比较细腻和具体。这说明了 LDA 主题模型提取的小说主题词对分析男频和女频网络小说的差异有重要作用。

4.4 模型融合

前面提到的三个模型分别从不同的角度，基于不同的特征集来进行分类，由于特征集差异较大，因此考虑将三个模型的分类结果进行融合。最终分类结果确定方式为：

Predict=α*lda_pred+β*title_pred+γ*stat_pred

其中，*lda_pred*、*title_pred* 和 *stat_pred* 分别是基于 LDA 的小说文本分类结果、小说标题短文本分类结果和基于统计信息的小说分类结果。每个分类都是二分类问题，因此用 0 和 1 表示两个类别。α、β和γ分别是三个模型的权重，它们满足：α+β+γ=1。最终分类结果根据 Predict 的值和某一阈值（如 0.5）的比较结果来确定。

在实验中，我们尝试了不同的模型权重取值，最终确定α=0.4，β=0.2，γ=0.4，此时融合模型的准确率达到了 96%。通过观察发现，融合后的结果主要依赖于基于计量风格学特征的分类结果和基于 LDA 的小说文本分类结果。这一结果说明：大部分的男频小说和女频小说在总体统计特征上面有着较大的差异，通过统计特征即可较好地区分两类小说。对于那些统计特征类似的小说，可以进一步通过分析文本的主题来进行分类。总之，这一结果证实了男频小说和女频小说具有较为明显的差异，而且这种差异是可以捕捉并且量化的。

5 特征分析

这一节讨论网络小说分类中具体的特征，分析两类小说在具体特征上的差异。

(1) 平均词长度与均章形容词

平均词长度、平均段长度、平均每章名词、平均每章形容词等特征在基于频率特征的分类中起到了重要的作用，它们在一定程度上反映了小说的用词习惯和结构。图 3 是决策树给出的准确率较高的规则之一。由图 3 和图 4 可以看出，女频网络小说的平均词长度一般不超过 1.52，且普遍低于男频，也就是说，女频网络小说使用的形容词多为简短型，且用词较为单一，查看实际分词过程中产生的

形容词，发现“大”、“老”、“好”、“冷”、“黑”这类的形容词出现较为频繁，不难看出这类词多偏向感官评价，且短小精悍；而男频网络小说分出的却大多是“巨大”、“最好”这类极端化、夸张化表达的形容词，体现了在语言表达风格上的性别差异。

```
均章名词 <= 149.267 [ 模式: 0 ] => 女频
    均章名词 > 149.267 [ 模式: 1 ]
        平均词长度 <= 1.526 [ 模式: 0 ]
            字数 <= 2,132,290 [ 模式: 0 ] => 女频
            字数 > 2,132,290 [ 模式: 1 ]
                平均段长度 <= 67.325 [ 模式: 1 ]
                    均章名词 <= 176.719 [ 模式: 0 ] => 女频
                    均章名词 > 176.719 [ 模式: 1 ]
                        机构团体 <= 1.773 [ 模式: 1 ] => 男频
                        机构团体 > 1.773 [ 模式: 0 ] => 女频
                平均段长度 > 67.325 [ 模式: 0 ] => 女频
        平均词长度 > 1.526 [ 模式: 1 ] => 男频
```

图 3 决策树部分分类规则

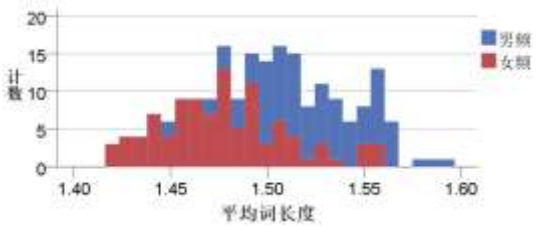

图 4 平均词长度分布

(2) 字数

字数通常彰显一本小说的篇幅。对于网络小说而言每一章的篇幅相差不大，段落又多以对话的方式呈现，因而字数、章节数在区分小说篇幅方面扮演着重要的角色。从图 5 可以看出，男频和女频的分界线大概在 2,448,859，男频小说篇幅字数基本都大于这个数。而根据图 3 可以看出，在均章名词大于 149.267 的情况下，男频、女频的篇幅差异以 2,132,290 为分界线，并且可以看出，这两个数字与语料库女频平均字符数都非常接近，而从图 5 可以看出，男频网络小说普遍比女频网络小说篇幅长。多数女频小说为满足广大女性读者的情感需求而存在，故事主线一般从男女主相识、相知到相爱展开，最多再加上结婚生子[18]，这样的故事即便是在现实生活中，跨越的时间维度也不会很大，何况是单纯以男女主角的情感为主线的网络小说。相比之下，男频小说则更偏向于一种英雄主义，打怪升级、解救苍生，至于谈情说爱，不过是其中的一条副线，场面铺的太大，维度自然就多了，这么多的内容不是寥寥几语可以说得清楚的。

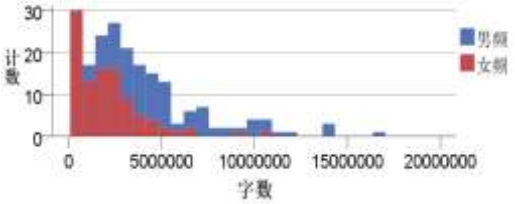

图 5 字数分布

(3) 机构团体数

机构团体词大部分都是一些颇具背景色彩的词汇，比如某都市生活类小说的机构团体词为：商业中心、XX 银行、XX 大学、公安局、高中同学等。这一类词的词频反映的是故事背景、年代还有阶层，具有很强的生活气息。根据分类规则，机构团体词无论多少，大部分为女频，在篇幅较长、均章名词较多的网络小说中数量较多。换句话说，大多数女频偏向生活化，家常里短、生活琐碎类的细节描写比较多，而只有篇幅达到一定长度的男频小说才会注意这些生活细节的描写。显然，这与男性、女性普遍关注点的不同关系紧密。

(4) 主题词

根据 LDA 主题模型提取的主题词来看，男频、女频的差异与传统性别文化有分不开的关系。在传统性别文化中，男性是家庭的顶梁柱，是责任的代表；而女性就应该相夫教子。由此引发了男性、女性显著的心理和情感需求差异，男频小说中传递的英雄情结和救世精神能满足男性被需要的心理诉求，而女性在风花雪月的女频小说中寄托自己的情感诉求。模型结果中，男频网络小说提取的关键词：实力、力量、修士、时间、弟子、感觉、声音、世界、长老、强者，彰显了一种向上的力量和宏大的气势，同时也带有男频特有的神秘色彩，而女频网络小说关键词：声音、男人、喜欢、女人、说话、孩子、眼睛、感觉、回来、电话，则显得非常细腻质朴，透着浓浓的生活气息。这些也与男性和女性的普遍生活、心理状态相契合，彰显了男频和女频

网络小说在主题和词汇分布方面的差异。

6 总结

本文借助自然语言处理和机器学习的方法从文本计量风格学特征、文本内容和文本标题三个方面对网络小说男频、女频的差异展开了研究，从而得出结论：计量风格学特征以及主题特征可以显著区分男频、女频网络小说。具体表现为：男频主题偏热血、大而壮阔，女频主题偏细腻，小而感性；男频小说篇幅、平均词长度、平均每章名词数普遍大于女频，而最能体现生活背景和细节的机构团体词，男频小说明显少于女频。

本文的研究结果可以应用于文学和文化研究领域，在一定程度上可以解释一些社会和文化现象。如男频小说篇幅普遍比女频长，而网络小说平台大多是按字数收费，因此男频网络小说作者收入普遍比女频网络小说高；而欧美之所以更偏爱男频，离不开男频新奇的故事背景和视角，而且大多数男频网络小说已经脱离了中国传统文化背景，其情节和人物设置与西方中世纪相近，便于西方读者理解；男频网络小说借由“打斗升级”等方式传递了一种“唯我独尊”的理念，迎合了西方读者的心理诉求，这些都是男频在西方备受追捧的重要原因。

本研究目前还存在许多不足，下一步计划进一步加大语料库，对文本内容做进一步的分析，提取更丰富的特征来进行建模。

参考文献

[1] 中国互联网络信息中心(CNNIC). 中国互联网络发展状况统计报告[EB/OL], http://www.cac.gov.cn/2019-02/28/c_1124175677.htm

[2]邵燕君, 吉云飞, 肖映萱. 媒介革命视野下的中国网络文学海外传播[J]. 文艺理论与批评, 2018(02):119-129.

[3] 王浩. 论当下网络文学的性别倾向[J]. 广西师范学院学报, 2009(04).

[4] 陈晓华. 跨媒介使用中的女性文化传播[D]. 复旦大学, 2013.

[5] 西蒙娜·德·波伏娃. 第二性[M]. 中国书籍出版社, 2004.4

[6] 陈熙熙, 性别视域下的网络小说语言[J]. 小说评论, 2013(05).

[7] 姜崇. 基于数据挖掘的网络小说价值预测分析[D]. 沈阳航空航天大学, 2018.

[8] 林钊生. 基于混合推荐算法的网络小说推荐系统设计与实现[D]. 华南理工大学, 2017.

[9] 李艳丽, 李宛蓉, 廖欣, 李静娟, 汤露, 刘喜平. 基于计量风格学的小说质量分析[J], 计算机与现代化, 2019, 05: 19-24+107

[10] 谢娜娜. 网络小说标题的语言综观[D]. 安徽大学, 2017.

[11] Z. Zhao, T. Liu, S. Li, et al. Ngram2vec: Learning Improved Word Representations from Ngram Co-occurrence Statistics[C], Proceedings of EMNLP, 2017, 244–253.

[12] Goldberg, Yoav; Levy, Omer. word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method [J/OL]. arXiv:1402.3722 [cs.CL], 2014

[13] fastText[EB/OL]. https://github.com/facebookresearch/fastText

[14] Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, Xiaoyong Du. Analogical Reasoning on Chinese Morphological and Semantic Relations[C], Proceedings of ACL, 2018.

[15] Support Vector Machines[EB/OL]. https://scikit-learn.org/stable/modules/svm.html

[16] gensim[EB/OL]. https://radimrehurek.com/gensim/

[17] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research.2003, 3(4–5): 993–1022.

[18] 王玉焦. 网络“女频”小说研究[D]. 贵州民族大学, 2018.