基于大数据挖掘的赛珍珠文化元素提取与应用

蒋驷驹，卢章平，李明珠
（江苏大学，镇江 212013）

摘要：目的 在大数据环境下，运用大数据技术提取赛珍珠文化元素，探究大数据挖掘理念在文创产品设计中应用的可行性。**方法** 首先，采集赛珍珠相关数据资料，借助网络爬虫工具采集网络媒体中赛珍珠相关的文本信息，同时人工搜集赛珍珠相关学术研究以及社会访谈资料，然后将数据保存为可编辑的文本形式。其次，运用中文分词工具对采集的文本信息进行处理，将语言字符串切分成词语，滤除中文停用词、低频词及干扰词，形成精炼的赛珍珠数据集合。之后，采用 LDA 主题模型算法对数据集合进行降维、聚类，形成初步的主题模型，然后经过人工筛选构建赛珍珠文化元素主题模型。最后，根据文化元素主题模型内容，选择赛珍珠文化元素进行赛珍珠文创产品设计实践。**结论** 依照大数据挖掘理念，通过对网络爬虫技术、中文分词工具以及 LDA 主题模型算法等大数据处理工具的综合应用，能够科学高效地从庞大的社会网络媒体中提炼赛珍珠文化元素，从而达到促进整个文创产品设计流程的效果。

关键词：大数据挖掘；网络爬虫；中文分词；主题模型；赛珍珠；文创产品
中图分类号：TB472 **文献标识码：**A **文章编号：**1001-3563(2021)22-0337-10
DOI：10.19554/j.cnki.1001-3563.2021.22.044

Extraction and Application of Pearl S Buck's Cultural Elements Based on Big Data Mining

JIANG Si-ju, LU Zhang-ping, LI Ming-zhu
(Jiangsu University, Zhenjiang 212013, China)

ABSTRACT: This paper aims to use big data technology to extract Pearl S Buck's cultural elements and explore the feasibility of applying big data mining concepts in the design of cultural and creative products in the big data environment. First, collect data about Pearl S Buck, use web crawler tool to collect text information about Pearl S Buck in online media, and manually collect related academic research and social interviews about Pearl S Buck at the meantime, and then save the data in editable text form. Second, use the Chinese word segmentation tool to process the collected text information, divide the language string into words, filter out Chinese stop words, low-frequency words and noise words, and form a refined Pearl S Buck data set. After that, the LDA theme model algorithm is used to reduce the dimensionality and clustering of the data set to form a preliminary theme model, and then construct the theme model of Pearl S Buck's cultural elements through manual screening. Finally, according to the content of the cultural element theme model, the Pearl S Buck's cultural elements were selected for the practice of Pearl S Buck's cultural and creative product design. According to the concept of big data mining, through the comprehensive application of big data processing tools such as web crawler technology, Chinese word segmentation tools and LDA theme model algorithm, it can scientifically and efficiently extract Pearl S Buck's cultural elements from the huge social network media, so as to achieve the effect of promoting the entire cultural and creative product design process.

KEY WORDS: big data mining; web crawler; Chinese word segmentation; theme model; Pearl S Buck; cultural and creative products

收稿日期：2021-06-09
基金项目：镇江市 2020 年度社科应用研究项目
作者简介：蒋驷驹（1995—），男，浙江人，江苏大学硕士生，主攻工业设计工程。
通信作者：卢章平（1958—），男，江苏人，博士，江苏大学教授，主要研究方向为图书馆情报。

近年来，随着互联网技术的迅猛发展，网络数据的快速增长成为了各行各业需要共同面对的机遇与挑战。庞大的数据体量不仅影响着互联网世界的运转，也使人们的工作与生活产生了前所未有的变化，人类社会正式迈入了大数据时代[1]。一般而言，大数据是指无法应用传统 IT 技术或者软硬件工具在可接受时间范围内进行获取、管理、分析和应用的数据集合，其特征是：体量大（Volume）、种类多（Variety）、生成快（Velocity）以及价值高（Value）[2]。基于大数据的特点，注定了传统 IT 技术无法高效地处理海量的数据信息，由此应运而生了大数据处理技术的研究，成为了互联网技术发展的新突破点。

大数据挖掘是大数据处理技术的重要内容之一，其具体理念是面对海量的数据采用科学方法高效地提取有用信息[3]。依据大数据挖掘的理念，本研究将从社会网络媒体中搜集赛珍珠相关数据资源，然后应用科学的方法进行数据分析，从中提炼赛珍珠文化元素，为赛珍珠文创产品开发提供创作源泉。赛珍珠是一位在中国成长、生活了数十年的美国女作家，她通过创作中国题材小说向西方世界展现了近代中国的真实现状，由此获得了 1938 年的诺贝尔文学奖。她的一生与中国结下了不解之缘，直至生命的最后时刻她仍想回到中国，被美国总统尼克松称为“沟通东西方文明的桥梁”[4]。

1 研究框架

根据赛珍珠文化元素挖掘的研究流程，首先需要采集赛珍珠相关的数据资源。为保证数据采集的全面性，本研究将从网络媒体、学术研究、社会访谈 3 个角度进行赛珍珠数据的采集，力求研究结果的准确性。同时在网络媒体数据采集阶段，将运用网络爬虫工具[5]自动采集主流网络媒体中的赛珍珠相关数据资源，以此提高采集效率。然后通过人工采集的方式搜集知网内赛珍珠文化相关的研究文献，并将之转换为可编辑的文本形式。最后设计访谈提纲，以电话访谈的形式获取社会人士对赛珍珠相关内容的认知与想法，将之记录并转换为文本形式。

将上述数据进行整合归纳后，需要对其进行数据预处理，即中文分词。利用中文分词工具对采集的文本数据进行语言字符串切分，同时滤除停用词、低频词及干扰词，提高后续研究的精确性。之后采用 LDA 主题模型算法进行文本数据降维、聚类，将精炼的赛珍珠数据集合通过机器语言处理以一定主题进行初步聚类，同时计算特征词权重比，从而形成有序的数据集合，然后经过人工筛选构建赛珍珠文化元素主题模型。至此，大数据挖掘工作基本完成。最后进入元素应用阶段，根据文化元素主题模型内容，选择赛珍珠文化元素转换为设计元素进行赛珍珠文创产品设计实践。依照以上研究流程及内容，可将研究步骤分为数据采集、数据预处理、主题模型构建、元素应用 4 个部分，研究框架见图 1。

2 数据采集

赛珍珠数据采集主要应用两种方法：一是应用网络爬虫工具进行互联网数据的高效搜集。网络爬虫是针对网络数据搜索、采集而开发的一种工具，在网络数据采集阶段具有至关重要的作用；二是人工搜集赛珍珠学术研究文献以及社会访谈资料。对于难以用网络爬虫进行快速搜集的数据信息，采用人工搜集的方式进行补全，力求数据获取的全面性。

2.1 网络媒体数据采集

网络爬虫又称 Web 信息采集器，是互联网搜索引擎获取信息的重要组件，它能够以特定的规则自动提取互联网网页信息与数据，被广泛地应用于大规模网络数据的搜索与采集，是当前大数据研究的重要工具之一。一般认为，网络爬虫的运行原理是将原始的网络信息地址（又称 URL）集合按照一定顺序全部列入到一个待执行队列中，然后以特定的规律从队列中取出网络信息地址并下载其指向的页面，之后根据所下载页面的内容，从中提取新的网络信息地址，并将之存入上述队列中。以此为基础流程重复运作，直至待执行队列中信息地址为空或者运行工作满足某个终止条件后结束网页爬取，最终达到遍历 Web 并获取有效的网络信息数据的效果[6]。

随着大数据研究的热度不断提升，网络爬虫工具的发展也越发成熟，当前互联网中已有多款面向用户的网络爬虫工具，如八爪鱼采集器、GooSeeker、HTTrack 等，用户仅需学习其操作方法，便可自行挖掘网页中所需要的数据信息。同时国内外众多科研团队研发了集数据采集、处理、分析于一体的信息处理软件，壮大了大数据分析工具的队伍，如北京理工大学张华平团队开发的 NLPIR 大数据语义智能分析平台[7]、马萨诸塞大学 Andrew McCallum 团队开发的 Mallet、武汉大学开发的 ROST 内容挖掘系统[8]等，为大数据研究提供了高效的途径。本研究首先使用 NLPIR 大数据语义智能分析平台的网络文本精准采集功能进行主流网站数据信息的精准采集，以“赛珍

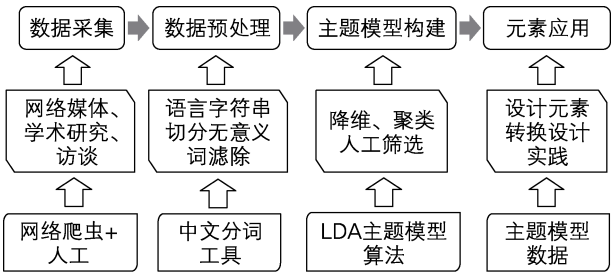

图 1 研究框架

Fig.1 Research framework

珠”为主题词，检索 1990 年至今包括人民网、中国新闻网、凤凰网、天涯社区等在内的主流新闻站点、网络论坛的相关主题新闻、报道及文章，经过人工清洗筛选后，共计搜集新闻报道、公告、论坛话题等网络文本 47 篇（约 11 万余字）。随后使用站点采集功能，对镇江赛珍珠研究会官网进行站点采集，采集的文本数据经人工清洗筛选后，共计搜集赛珍珠相关报道、公告及文章 132 篇（约 30 万余字）。

此外，采用八爪鱼采集器和简数数据采集平台，分别采集以微博、微信为代表的主流社交媒体中发布的赛珍珠相关短讯与文章。以“赛珍珠”为主题关键词进行数据搜集，经人工清洗筛选后，共计采集微博短文本 500 余条（约 62000 余字），微信公众号文章 78 篇（约 26 万余字），时间跨度为 2010 年至今。最后，将上述文本数据进行归类存档，粗略分析其文本内容，主要包含赛珍珠生平事迹、赛珍珠著作评论、赛珍珠与中国、赛珍珠人物评价等方面。

2.2　学术文献及社会访谈数据采集

在完成网络媒体方面的数据采集后，人工搜集学术界对赛珍珠的相关研究。在知网中以“赛珍珠”为主题关键字进行检索，搜集检索结果中的高被引学术期刊论文共 30 篇（约 23 万余字），内容涉及赛珍珠与中西文化、赛珍珠小说解读、赛珍珠意识形态辨析等方面。同时，以电话访谈的形式从社会认知角度采集赛珍珠相关资料。在访谈对象选取上，选择对赛珍珠具有一定认知度的人员，以确保采集过程成功进行并获得有价值的访谈数据。访谈问题的设计主要围绕赛珍珠浅层资料进行问询，如赛珍珠著作及其中的人物解析等，以期后续提取的赛珍珠文化元素具有表征性，便于理解与应用。在上述准备的基础上，共对 10 位受访者进行每位 10 到 20 分钟的访谈并录音，然后将录音转换为文本形式存储。至此，数据采集基本完成。

3　数据预处理

3.1　中文分词

在中英文网络数据的处理中，由于英文文本的单词之间是用空格分隔的，因此只需对其后缀进行处理，即可进行数据分析。而中文是连续的字符串语言，词与词之间没有间隔，因此中文文本在数据统计分析之前需要进行分词操作[9]。

当前，中文分词工具种类繁多，但主要分词原理分为 3 种：一是匹配字符串分词，这类方法是在已有字典的基础上，使文本按照指定的规则实行字符串“最大化”匹配，符合则识别出一个词，以此推进，最终完成分词操作；二是理解分词，即模拟人对自然语言的理解，结合语法、语义、语用等因素进行分词操作；三是统计分词，根据计算机统计语料库中字符串出现的频次来判断是否对其进行分词操作[10]。

通过测试文件对多款分词工具进行对比分析，最终选用 Jieba 分词对赛珍珠文本数据进行分词操作[11]。Jieba 分词是基于 Python 语言开发的分词模块，是目前国内常用的中文分词工具之一，其包含多种分词模式，适用于用户不同的需求。本研究主要使用 Jieba 分词的精确分词模式进行赛珍珠数据集合的分词操作。

3.2　自定义词典添加及停用词滤除

在分词操作前需要添加用户自定义词典以及停用词表。用户自定义词典是指根据用户自主意识规定分词工具避免切分的词语集合。通过 Jieba 分词源码可以发现，Jieba 分词本身包含一个 2 万多词条的词典，用来实现中文语句中词语的快速扫描及切分[12]，因此 Jieba 分词支持在现有词典的基础上添加用户自定义词典，来满足用户的个性化需求。具体操作是在用户构建自定义词典的基础上，通过 Python 在 Jieba 分词模块中调用用户自定义词典文件，然后在分词过程中 Jieba 分词就能识别语料库中的用户自定义词语，提高用户语料切分的准确性。

停用词起源于信息检索领域，具体是指文本数据中出现频次较多，但又没有过多检索意义的词，如“的、是、太、of”等[13]。对比当前主流的几款中文停用词表，最后选用哈工大停用词表进行停用词滤除[14]。哈工大停用词表是当前常用的几款停用词表之一，一般包含 700 余个常见的停用词，能够有效地提高中文分词的精确性。停用词表的使用是在 Jieba 分词模块中加入停用词模块，然后调用停用词表文件，实现分词过程中停用词的自动滤除。

3.3　低频词及干扰词滤除

低频词及干扰词滤除主要是在词频统计以及词性标注的基础上，筛选出低频词以及无意义干扰词，然后将其纳入停用词表进行分词结果的二次遍历。通过调用 Jieba 分词模块的词性标注功能，对一次遍历的分词结果进行词性标注，然后利用 AntConc 工具[15]对标注结果进行词频统计。将统计结果导入 Excel 进行词频及词性筛选，对筛选出的无意义低频词以及副词、代词、连词等无意义干扰词进行整合并输入停用词表中，然后再对一次遍历的分词结果进行二次遍历，最终形成更为精炼的分词结果。综合上述数据预处理的过程，构成数据预处理流程图，见图 2。

4　主题模型构建

面对大规模的文本数据集合，应用主题模型进行数据分析是从中挖掘有用信息的一种高效方式。主题模型是一种概率生成模型，定义一个文档是由多个潜在主题以一定概率分布组成，而每个潜在主题是由其特征词按一定概率的隐含内涵聚类形成[16]，因此应用

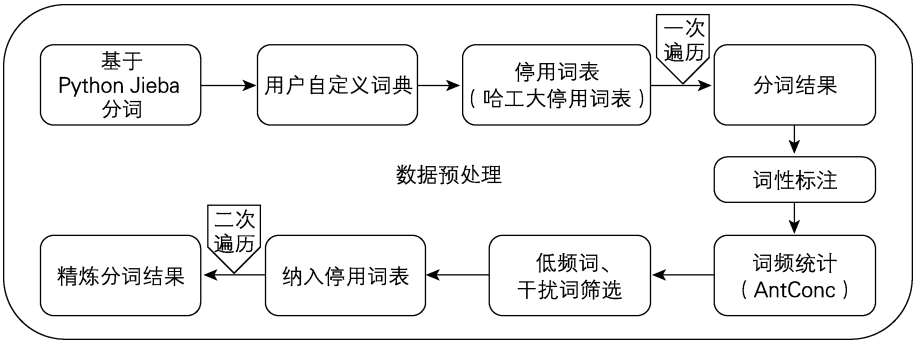

图 2 数据预处理流程图

Fig.2 Flow chart of data preprocessing

主题模型能够有效地对数据集合进行降维、聚类，被广泛地应用于大规模文本数据的研究中。经过中文分词处理的赛珍珠数据集合虽已精炼了文本中所蕴含的文化元素信息，但仍处于高维、无序的分散状态，如按传统方式进行提取需要耗费大量的时间与精力，因此采用主题模型进行初步地降维与聚类，可形成赛珍珠文化元素主题模型的雏形，为后续构建完整的文化元素主题模型建立基础。

本研究所采用的 LDA（Latent Dirichlet Allocation，隐含狄利克雷分布）主题模型算法是一种常见的主题模型算法，其以“词、主题、文档”三层贝叶斯概率为核心结构，能够有效发现文本中隐含的主题，具有优秀的降维能力，适用于文本分类、主题识别以及关联性判断等数据挖掘任务[17]。

4.1 参数设置

运用本模型算法进行主题建模前需要进行参数设置，参数包含用于生成主题向量的 α、单词概率分布矩阵 $\boldsymbol{\beta}$、主题个数 K、迭代次数以及主题显示词项个数 W 等。由于本研究对参数量化的要求并不苛刻，故依照默认设置：α 取 50/K，$\boldsymbol{\beta}$ 取 0.01，迭代次数为 1000 次[18]，迭代次数过小会导致模型尚未收敛，过大则会浪费计算资源；K 值对主题模型的影响较大，K 值太小容易导致主题泛化，太大则会导致主题过度细分，故按照 K={25,50,75,100}顺序依次扫描，选取主题模型结果最符合预期的 K 值。且由于赛珍珠文化元素主要来源于主题概率分布选择的特征词，因此主题显示词项个数 W 默认选择为 100，尽量多地显示高权重词语。

明确以上参数设置规则后，将精炼的赛珍珠数据集合导入 LDA 主题模型算法进行主题聚类，按主题个数 K={25,50,75,100}顺序依次测试后发现，当主题个数 K 为 50 时，所得主题模型内容最符合预期效果，因此确定设置主题个数 K 为 50、α 为 1，$\boldsymbol{\beta}$ 取 0.01，迭代次数为 1000 次，显示词项个数 W 为 100。最后将所得主题模型数据进行整理归纳，以便下一步操作。

4.2 构建赛珍珠文化元素主题模型

由于计算机处理自然语言的局限性，运算后获取的主题模型需要人为调整才能构建完整的赛珍珠文化元素主题模型。计算机自然语言处理是一个复杂的研究方向，其不仅需要研究计算机学方面的内容，还要运用语言学以及其他客观世界的知识，因此现阶段计算机自然语言处理虽取得了一定的研究成果，但仍有很多难题尚未克服，要达到人主观认知的精确自然语言处理效果仍需要走很长一段路。且汉语作为一门没有形态变化的意合分析型语言，其语义会因词语搭配、语用场景、人为表达等因素的影响而改变，导致计算机对中文词语的语义识别及分类一直是个难以攻克的问题[19]，因此初步聚类的文化元素主题模型需要人工干预才能构建完整的赛珍珠文化元素主题模型。

为确定赛珍珠文化元素主题模型的主题类型，需要从赛珍珠文化元素的提取角度进行思考。赛珍珠作为一位文化名人，即可从她的名人身份角度进行发散。一般认为，名人是指在历史上取得过巨大成就或者因造成重大影响而被人熟知的个人或群体。赛珍珠一生辗转于中国和美国两个国家，既是一位高产作家，也是一位慈善活动家，她创办“欢迎之家”和赛珍珠基金会，救助美国人与亚洲妇女非婚所生的弃儿；创办“东西方协会”鼓励亚洲作家为自己国家发声，同时她创作诸多中国题材小说，为当时的西方世界了解中国做出了不可磨灭的贡献，被尼克松总统亲切地称为“沟通东西方文化的桥梁”[20]。因此从名人角度思考，可从其相关人物、地域、著作、生平事迹及物品、相关机构、后世评价、经历时代以及相关思想等方面考虑文化元素的提取，即主题模型类型。

在明确主题模型类型后，将原始主题模型中的特征词按主题模型类型重新划分，再根据权重判定特征词在语料中的重要性及印象性进行降序排序，最终构建完整的赛珍珠文化元素主题模型。赛珍珠人物主题模型（部分）见表 1，赛珍珠地域主题模型（部分）见表 2，赛珍珠著作主题模型（部分）见表 3，赛珍珠事物主题模型（部分）见表 4，赛珍珠机构主题模型（部分）见表 5，赛珍珠评价主题模型（部分）见表 6，赛珍珠时代主题模型（部分）见表 7，赛珍珠思想主题模型（部分）见表 8。

表 1　赛珍珠人物主题模型（部分）
Tab.1　Character theme model of Pearl S Buck (Part)

人物主题					
赛珍珠	0.64198	凯丽	0.033241	胡适	0.012224
林语堂	0.1093	王源	0.029278	毛泽东	0.010107
王龙	0.107284	刘龙	0.027639	泰戈尔	0.009778
鲁迅	0.058061	尼克松	0.023238	王虎	0.009167
徐志摩	0.052586	王妈	0.020271	齐美尔	0.009143
赛兆祥	0.052202	孔先生	0.019244	龙墨芗	0.009107
阿兰	0.047623	吴太太	0.018528	胡风	0.009043
王莹	0.039984	蒋介石	0.016089	茅盾	0.008863
布克	0.035306	桂兰	0.015141	戴德生	0.008541
老舍	0.033278	安德鲁	0.014712	邵德馨	0.008093

表 2　赛珍珠地域主题模型（部分）
Tab.2　Regional theme model of Pearl S Buck (Part)

地域主题					
中国	0.321756	宾夕法尼亚州	0.009752	满洲	0.0056
美国	0.271726	德国	0.008649	宿迁	0.005531
镇江	0.186695	北京	0.008042	金陵	0.005241
上海	0.114269	纽约	0.007426	云台山	0.005025
庐山	0.061448	金山	0.007277	斯德哥尔摩	0.004925
南京	0.048432	清江浦	0.007186	法国	0.004666
宿州	0.02897	青山农场	0.006926	坦佩市	0.004564
亚洲	0.014714	润州	0.006404	香港	0.004416
弗吉尼亚州	0.010045	皖北	0.006113	扬子江	0.004367
英国	0.009844	淮安	0.005903	梁山	0.003903

表 3　赛珍珠著作主题模型（部分）
Tab.3　Works' theme model of Pearl S Buck (Part)

著作主著作主题					
《大地》	0.04351	《爱国者》	0.011127	《赛珍珠》	0.00589
《我的几个世界》	0.023134	《东风・西风》	0.010241	《骆驼祥子》	0.005754
《亚洲》	0.02205	《红楼梦》	0.009974	《三国演义》	0.005497
《龙种》	0.017328	《战斗的天使》	0.009818	《王龙》	0.004098
《流亡者》	0.016733	《春江花月夜・赛珍珠》	0.009605	《赛珍珠文化传记》	0.004041
《分家》	0.014852	《母亲》	0.009135	《过路的桥》	0.004024
《儿子们》	0.011882	《圣经》	0.008479	《四世同堂》	0.003842
《四海之内皆兄弟》	0.011706	《中国小说》	0.006311	《也说中国》	0.003553
《吾国与吾民》	0.011521	《西行漫记》	0.006233	《中国之美》	0.003448
《大地》三部曲	0.011307	《北京来信》	0.006035	《纽约时报》	0.003446

表 4 赛珍珠事物主题模型（部分）

Tab.4 Objects' theme model of Pearl S Buck (Part)

事物主题					
诺贝尔文学奖	0.145419	任教	0.00996	《排华法案》	0.005275
土地	0.106385	坟墓	0.009546	颁奖典礼	0.005238
翻译	0.09271	大地	0.009526	纪念碑	0.005025
创作	0.067499	结婚	0.008828	歌曲	0.004676
普利策奖	0.028326	演讲	0.008642	授奖仪式	0.004479
墓碑	0.027879	版税	0.007426	离婚	0.004302
收养	0.014714	畅销书	0.006645	手稿	0.004107
奥斯卡	0.012269	水牛	0.006193	旗袍	0.003879
教堂	0.01191	援华	0.005994	大学毕业	0.003849
布道	0.010743	科举制	0.005598	慈善事业	0.003753

表 5 赛珍珠机构主题模型（部分）

Tab.5 Mechanism theme model of Pearl S Buck (Part)

机构主题					
基督教	0.037361	米高梅公司	0.006933	金陵神学院	0.002955
纪念馆	0.032615	白宫	0.006493	南京政府	0.002668
赛珍珠基金会	0.024405	出版社	0.005122	美国之音	0.002492
南京大学	0.020144	伦道夫·梅肯女子学院	0.005048	联合会	0.002399
赛珍珠故居	0.018703	农学院	0.004755	瑞典文学院	0.002382
长老会	0.016347	东南大学	0.004528	西南联大	0.00236
委员会	0.01534	每月图书俱乐部	0.004354	寺庙	0.002325
好莱坞	0.013596	文化公园	0.004107	慈善机构	0.002311
东西方协会	0.011026	赛珍珠故居	0.003171	福利院	0.002311
欢迎之家	0.00981	崇实女子中学	0.002967	美国国家广播公司	0.002236

表 6 赛珍珠评价主题模型（部分）

Tab.6 Evaluation theme model of Pearl S Buck (Part)

评价主题					
传教士	0.075654	大爱	0.003464	急先锋	0.002334
异乡人	0.034342	社会活动家	0.003357	翻译家	0.002199
艺术家	0.006482	汉学家	0.003112	傲慢	0.002104
农学家	0.005886	美国作家	0.00298	慈善家	0.001979
博爱	0.005139	文化边缘人	0.002795	仁爱	0.00187
著名作家	0.004924	同情心	0.002683	女权活动家	0.001668
评论家	0.004738	桥梁	0.002501	剧作家	0.001557
人桥	0.004247	慷慨解囊	0.002447	镇江人	0.001522
国际友人	0.004158	社会活动家	0.002371	爱心	0.001445
生命力	0.003863	赛中国通夫人	0.002334	传播者	0.000736

表 7　赛珍珠时代主题模型（部分）
Tab.7　Time theme model of Pearl S Buck (Part)

时代主题					
抗日战争	0.030803	第二次世界大战	0.004158	少年时代	0.002568
1932 年	0.018328	圣诞节	0.00405	1938 年	0.002382
1934 年	0.015105	1933 年	0.003966	尼克松访华	0.001973
冷战	0.006893	辛亥革命	0.003609	1949 年	0.001734
1991 年	0.006793	改革开放	0.003597	文化大革命	0.001557
义和团运动	0.006325	乒乓外交	0.00313	太平天国	0.001542
青年时代	0.005903	1935 年	0.002718	民国时期	0.001363
1936 年	0.004925	1914 年	0.002718	20 世纪 20 年代	0.000974
南京大屠杀	0.004484	1926 年	0.002718	1973 年	0.000746
20 世纪五六十年代	0.004307	新文化运动	0.002615	世界反法西斯战争	0.000736

表 8　赛珍珠思想主题模型（部分）
Tab.8　Thought theme model of Pearl S Buck (Part)

思想主题					
中西文化	0.022937	文化差异	0.006617	西方中心主义	0.003141
人道主义	0.019906	文化相对主义	0.006176	历史主义	0.003103
跨文化	0.019408	后殖民主义	0.006083	资本主义	0.002944
中国文化	0.016909	东方文化	0.005925	儒家思想	0.002897
西方文化	0.014518	麦卡锡主义	0.004247	四海之内皆兄弟	0.002756
共产主义	0.011844	反法西斯	0.004157	封建主义	0.002561
殖民主义	0.010987	现实主义	0.004125	马克思主义	0.002337
文化交流	0.01044	女性主义	0.004039	封建礼教	0.002336
帝国主义	0.009688	天下一家	0.003951	文化冲突	0.002207
全球化	0.006911	理想主义	0.003283	后殖民主义	0.001999

5　元素应用

根据赛珍珠文化元素主题模型，进行赛珍珠文创产品设计实践，以验证大数据技术在文创产品设计中应用的可行性。文创产品是人为创意化设计的产物，是将文化与功能产品有机融合形成的特殊产品[21]。一般认为，文创产品的文化表现源自设计对象的文化特征，即具有特殊性文化内容。其可凝练为包涵文化内容的文化元素，作为文创产品设计流程中文化特性的简化表达。但文化元素因不具备设计的艺术性而无法直接应用于设计实践，需要将其艺术化地转换为设计元素才能与功能产品有机融合形成文创产品[22]。以同样是名人作家的鲁迅为例，其选择了与鲁迅相关的人物、地域、事迹等文化元素，解读了其文化内涵，并将之转换成了可应用于设计实践的设计元素，然后再与功能产品融合形成了种类丰富的鲁迅文创产品，鲁迅文创产品示例见表 9。

因此根据鲁迅文创产品的案例可知，赛珍珠文创产品设计同样需要先将文化元素转换为设计元素才能与功能产品进行融合形成文创产品。在此基础上，选取赛珍珠文化元素主题模型中的特征词进行引用，解读其内涵，将之转换为可应用于设计实践的设计元素，然后再将其与功能产品有机融合，形成包涵赛珍珠文化特征的文创产品，赛珍珠文创产品设计流程表见表 10。

经上述设计实践论证，赛珍珠主题模型中的文化元素经解读后可形成赛珍珠文化内涵，将之作为文化特征以艺术化手段转换为设计元素后，可与功能产品有效融合形成具有赛珍珠文化表现的文创产品。从研究全局来看，赛珍珠文创产品设计验证了大数据挖掘所获的赛珍珠文化元素能够被有效地应用于设计实践，证明了大数据研究技术在文创产品设计领域应用的可行性。

表 9 鲁迅文创产品示例

Tab.9 Examples of Lu Xun's cultural and creative products

文化元素	解读	设计元素	文创产品
鲁迅	鲁迅是近代中国最伟大的文学家之一，他创作了诸多脍炙人口的作品，是中国现代文学的奠基人。同时他也是一位思想家、教育家、革命家，引领了破除封建礼教五四新文化运动。		 鲁迅头像钥匙扣
绍兴	绍兴是著名的水乡城市，是大文豪鲁迅的故乡。这里水多、桥多、美食多，青瓦白墙构建了鲁迅的童年。		 “古城绍兴”冰箱贴
“早”字	少年鲁迅因迟到而被老师责骂，为了督促自己，他在桌子上刻下“早”字，并许下誓言不再迟到。		 “早”字钥匙扣

表 10 赛珍珠文创产品设计流程表

Tab.10 Design flow chart of Pearl S Buck's cultural and creative products

主题类型	文化元素	解读	设计元素	文创产品
人物	赛珍珠	赛珍珠通过创作中国题材小说获得了 1938 年的诺贝尔文学奖，她用文字向世界展现了近代中国农民的真实状况，被称为沟通中西方文化的“人桥”。		赛珍珠头像钥匙扣见图 3
地域	镇江	镇江被称为“天下第一江山”，拥有著名的旅游景区——三山风景区，其市徽标志来源于此。赛珍珠一出生便被父母带到了镇江，在此度过了人生宝贵的童年和青年时期，奠定了她对中国的认知与情感。		赛珍珠“魅力镇江”胸针见图 4
著作	《大地》	《大地》是赛珍珠著名的长篇小说，是诺贝尔奖和普利策奖双获奖作品，描述了中国农民与土地的不解之情。		赛珍珠《大地》书签见图 5
思想	四海之内皆兄弟	“四海之内皆兄弟”是赛珍珠硕士论文卷首语，源自孔子《论语》中的对话，表达了赛珍珠对种族平等、人人平等的向往。		赛珍珠“四海之内皆兄弟”开瓶器见图 6

图 3　赛珍珠头像钥匙扣
Fig.3　Key chain of Pearl S Buck's head

图 4　赛珍珠“魅力镇江”胸针
Fig.4　“Charming Zhenjiang” brooch of Pearl S Buck

图 5　赛珍珠《大地》书签
Fig.5　*The Good Earth* bookmark of Pearl S Buck

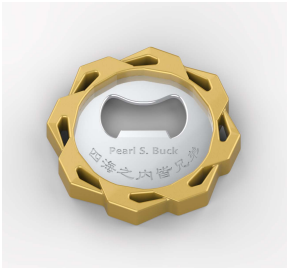

图 6　赛珍珠“四海之内皆兄弟”开瓶器
Fig.6　“All men are brothers” bottle opener of Pearl S Buck

6　结语

随着互联网数据的快速增长，大数据时代已经成为了社会发展不可逆转的趋势。作为现实世界在虚拟世界中投射的数据集合，大数据逐渐成为了连接现实与虚拟的重要纽带，被越来越多地应用于促进现实社会的运行与发展。本研究植根于大数据研究技术，经过采集、处理、筛选等层层工序，从社会网络媒体中提取了赛珍珠文化元素，并将之应用于赛珍珠文创产品的开发，不仅为文创产品设计提供了新的方法思路参考，而且为大数据应用研究开拓了一条前所未有的发展道路。

参考文献：

[1]　王元卓. 网络大数据：现状与展望[J]. 计算机学报, 2013, 36(6): 1125-1138.
WANG Yuan-zhuo. Network Big Data: Current Situation and Prospect[J]. Journal of Computer Science, 2013, 36(6): 1125-1138

[2]　李国杰, 程学旗. 大数据研究：未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J]. 中国科学院院刊, 2012, 27(6): 647-657.
LI Guo-jie, CHENG Xue-qi. Big Data Research: a Major Strategic Area of Future Science, Technology and Economic and Social Development: Research Status and Scientific Thinking of Big Data[J] Journal of the Chinese Academy of Sciences, 2012, 27(6): 647-657.

[3]　童星强. 基于数据挖掘的网络新闻热点发现系统设计与实现[D]. 北京：北京邮电大学, 2019.
TONG Xing-qiang. Design and Implementation of Web News Hotspot Discovery System Based on Data Mining[D]. Beijing: Beijing University of Posts and Telecommunications, 2019

[4]　赛珍珠. 大地[M]. 北京：人民文学出版社, 2019.
Pearl S Buck. Earth[M]. Beijing: People's Literature Publishing House, 2019.

[5]　孙立伟, 何国辉, 吴礼发. 网络爬虫技术的研究[J]. 电脑知识与技术, 2010, 6(15): 4112-4115.
SUN Li-wei, HE Guo-hui, WU Li-fa. Research on Web Crawler Technology[J]. Computer Knowledge and Technology, 2010, 6 (15): 4112-4115.

[6]　郭丽蓉. 基于 Python 的网络爬虫程序设计[J]. 电子技术与软件工程, 2017(23): 248-249.
GUO Li-rong. Design of Web Crawler Program Based on Python[J]. Electronic Technology and Software Engineering, 2017(23): 248-249.

[7]　张华平, 商健云. NLPIR Parser: 大数据语义智能分析平台[J]. 语料库语言学, 2019, 6(1): 87-104.
ZHANG Hua-ping, SHANG Jian-yun. NLPIR Parser: Corpus Linguistics of Big Data Semantic Intelligence Analysis Platform[J]. 2019, 6(1): 87-104.

[8]　张雯雯. 文本挖掘工具述评[J]. 图书情报工作, 2012(8): 26-31.
ZHANG Wen-wen. Review of Text Mining Tools[J]. Library and Information Work, 2012(8): 26-31.

[9]　马玉春, 宋瀚涛. Web 中文文本中文分词技术研究[J]. 计算机应用, 2004, 24(4): 134-135.
MA Yu-chun, SONG Han-tao. Research on Chinese Word

Segmentation Technology of Web Chinese Text[J]. Computer Application, 2004, 24(4): 134-135.

[10] 于游, 付钰, 吴晓平. 中文文本分类方法综述[J]. 网络与信息安全学报, 2019, 5(5): 1-8.
YU You, FU Yu, WU Xiao-ping. A Review of Chinese Text Classification Methods[J]. Journal of Network and Information Security, 2019, 5(5): 1-8.

[11] 黄翼彪. 开源中文分词器的比较研究[D]. 郑州: 郑州大学, 2013.
HUANG Yi-biao. A Comparative Study of Open Source Chinese Word Breakers[D]. Zhengzhou: Zhengzhou University, 2013.

[12] 邢彪, 根绒切机多吉. 基于 jieba 分词搜索与 SSM 框架的电子商城购物系统[J]. 信息与电脑, 2018(7): 104-108.
XING Biao, Genrongqiejiduoji. Information and Computer of E-mall Shopping System[J]. Based on Jieba Segmentation Search and SSM Framework, 2018(7): 104-108.

[13] 江兆中. 基于语境和停用词驱动的中文自动分词研究[D]. 合肥: 合肥工业大学, 2010.
JIANG Zhao-zhong. Research on Chinese Automatic Segmentation Driven by Context and Stop Words[D]. Hefei: Hefei University of Technology, 2010

[14] 官琴, 邓三鸿, 王昊. 中文文本聚类常用停用词表对比研究[J]. 数据分析与知识发现, 2017(3): 72-80.
GUAN Qin, DENG San-hong, WANG Hao. A Comparative Study of Frequently Used Stop Lists in Chinese Text Clustering[J]. Data Analysis and Knowledge Discovery, 2017(3): 72-80.

[15] 王春艳. 免费绿色软件 AntConc 在外语教学和研究中的应用[J]. 外语电化教学, 2009, 125: 45-48.
WANG Chun-yan. Application of Free Green Software AntConc in Foreign Language Teaching and Research[J]. Foreign Language Audio Visual Teaching, 2009, 125: 45-48.

[16] 徐晨飞, 孙静. “江海文化”资源知识聚合策略与模型设计研究[J]. 情报探索, 2019(12): 10-14.
XU Chen-fei, SUN Jing. Research on the Aggregation Strategy and Model Design of “Jianghai Culture” Resource Knowledge[J]. Information Exploration, 2019(12): 10-14.

[17] 王丽, 邹丽雪, 刘细文. 基于 LDA 主题模型的文献关联分析及可视化研究[J]. 数据分析与知识发现, 2018(3): 98-106.
WANG Li, ZOU Li-xue, LIU Xi-wen. Literature Association Analysis and Visualization Research Based on LDA Theme Model[J]. Data Analysis and Knowledge Discovery, 2018(3): 98-106.

[18] 高璐. 基于主题模型的藏汉跨语言信息检索查询扩展研究[D]. 北京: 中央民族大学, 2017.
GAO Lu. Research on Query Expansion of Tibetan Chinese Cross Language Information Retrieval Based on Theme Model[D]. Beijing: Central University for Nationalities, 2017.

[19] 郭艳华, 周昌乐. 自然语言理解研究综述[J]. 杭州电子工业学院学报, 2000(1): 58-65.
GUO Yan-hua, ZHOU Chang-le. A Review of Natural Language Understanding[J]. Journal of Hangzhou Institute of Electronic Technology, 2000(1): 58-65.

[20] [美]保罗 A 多伊尔. 张晓胜, 等. 译. 赛珍珠[M]. 沈阳: 春风文艺出版社, 1991.
[US]PAUL A Doyle. Translated by ZHANG Xiao-sheng, et al. Pearl S Buck[M]. Shenyang: Chunfeng Literature and Art Publishing House, 1991.

[21] 鲁志伟. 现代文化创意产品设计的现状研究[J]. 大众文艺, 2018(7): 83.
LU Zhi-wei. Research on the Current Situation of Modern Cultural Creative Product Design[J]. Popular Literature and Art, 2018(7): 83.

[22] 蒋驷驹, 卢章平, 李明珠. 文化创意产品多元化设计研究与应用[J]. 包装工程, 2020(1): 1.
JIANG Si-ju, LU Zhang-ping, LI Ming-zhu. Research and Application of Diversified Design of Cultural and Creative Products[J]. Packaging Engineering, 2020(1): 10.