DOI:10.16644/j.cnki.cn33-1094/tp.2023.05.026

数智赋能视域下侗族大歌主题挖掘及演化分析研究*

杨秀璋[1]，武　帅[1,2]，项美玉[3]，廖文婧[1]，任天舒[1]，刘建义[1]

(1. 贵州财经大学信息学院，贵州 贵阳 550025；2. 南京农业大学信息管理学院；
3. 贵州财经大学大数据应用与经济学院(贵阳大数据金融学院))

摘　要：针对新范式对传统人文社科造成学科边界模糊的问题，提出一种基于LDA模型和关系图谱的领域主题演化算法。首先运用综合指数确认领域核心研究人员；再利用LDA模型和层次聚类确认领域划分的主题数，构建关系图谱，探究主题词关联性；最后创新性的提出演化各主题间的关联性。实验将侗族大歌领域分为三个主题，并演化出他们之间的联系。该研究可为数智赋能语音音频的深入研究提供基础。

关键词：侗族大歌；数智赋能；领域主题挖掘；关系图谱；主题演化

中图分类号：TP391　　**文献标识码：**A　　**文章编号：**1006-8228(2023)05-118-05

Research on the theme mining and evolution analysis of the Dong Chorus from the perspective of data intelligence empowerment

Yang Xiuzhang[1], Wu Shuai[1,2], Xiang Meiyu[3], Liao Wenjing[1], Ren Tianshu[1], Liu Jianyi[1]

(*1. School of Information, Guizhou University of Finance and Economics, Guiyang, Guizhou 550025, China;*
2. School of Information Management of Nanjing Agricultural University;
3. Guiyang School of Big Data and Finance, School of Big Data Application and Economics, Guizhou University of Finance and Economics)

Abstract: Aiming at the problem that the new paradigm has blurred the boundaries of traditional humanities and social sciences, a domain theme evolution algorithm based on the LDA model and relationship map is proposed. Firstly, the composite index is used to identify core researchers in the domain. Secondly, the LDA model and hierarchical clustering are used to confirm the number of themes divided by the domain, and a relationship map is constructed to explore the relevance of theme words. Finally, the innovative proposal is made to evolve the relevance among the themes. In the experiments, the domain of the Dong Chorus is divided into three themes and the relationship between them is evolved. The research can provide a basis for the in-depth study of data intelligence empowering voice audio.

Key words: Dong Chorus; data intelligence empowerment; domain theme mining; relationship map; theme evolution

0 引言

侗族大歌作为一种起源于春秋战国时期的演唱方式，距今已有2500年历史，是我国最早且被国际所认可的一种无指挥、无伴奏、自然多声部的民间复调音乐艺术类型[1]。侗族大歌演奏环节无指挥、无伴奏，民族特征性强，研究侗族大歌的意义一定程度高于其他民族歌曲。

随着数智赋能研究热度不断增加，国内学者高度重视文化载体研究，尝试对中国传统文化进行创新性发展和转化，探索人文定性与数字定量分析的融合过程，贡献了大量研究研究成果，开拓出一系列研究领域和热点[2]。但由于新范式一定程度冲击了传统人文

收稿日期：2022-11-08

***基金项目：**贵州省科技计划项目(黔科合基础[2020]1Y279，黔科合基础[2019]1041)；贵州省教育厅青年科技人才成长项目(No.黔教合KY字[2016]175，No.黔教合KY字[2021]135)；贵州财经大学2021年度校级项目(No.2021KYQN03)

作者简介：杨秀璋(1991-)，男，苗族，贵州凯里人，硕士，助教，主要研究方向：数据挖掘、知识图谱、数据分析。

通讯作者：武帅(1994-)，男，江苏淮安人，硕士，工程师，主要研究方向：数智赋能、自然语言处理。

社科的研究认知体系，存在学科边界模糊、共识标准无序、评价体系欠缺等问题，一定程度影响数智赋能研究的发展。本文以“侗族大歌”的学术文献作为研究对象，研究其主题变化趋势，一定程度能反映相关领域的研究变化，为数智赋能语音音频领域研究的深入提供理论基础，同时也为民族文化保护提供了指导性意见。

1 相关研究

1.1 主题挖掘

主题挖掘(Topic Mining)[3]作为文本挖掘领域的一个重要研究问题，旨在发掘文本隐含的主题信息，实现解决“一词多义”和“一义多词”的语言现象。最初运用于信息检索领域，后推广至网络舆情、电商推荐、文献挖掘以及语料提升。

1.2 主题演化

主题演化分析旨在通过共词分析、主题模型、文本聚类等方法挖掘各个阶段主题的发展历程及演变趋势[4]。主题演化的构建需要应用到多方面信息处理技术，其分析过程可分为三步。①实体识别，是构建主题演化的基础，对文献结构化数据进行实体抽取，并存储于图数据库中。②共词网络，是提升主题演化的关键，将抽取的实体映射到共词网络中，发现共现关键词组，并赋予共现词组权重系数。③图谱可视化，是展现主题演化的结果，将已确认的各实体间的关系按照时间段进行可视化呈现。

2 总体框架

本文的主题演化分析框架大致可分为四个模块，如图1所示，分别是数据采集和预处理；核心科研群体发现；侗族大歌主题挖掘；侗族大歌主题演化模块。

⑴ 数据采集和预处理：以“侗族大歌”为关键词，截至2022年10月，检索筛选中国知网(CNKI)平台上论文1378篇，存于csv文件中，并对其进行中文分词、去停用词等数据预处理，编码UTF-8。

⑵ 核心科研群体发现：第一作者发文量和总引用量结合普莱斯定律、综合指数确认核心作者。

⑶ 侗族大歌主题挖掘：对目标数据的主题及相似度进行LDA主题挖掘和层次聚类挖掘。

⑷ 侗族大歌主题演化：结合目标数据的主题和时间数据来构建关系图谱，进行主题演化分析。

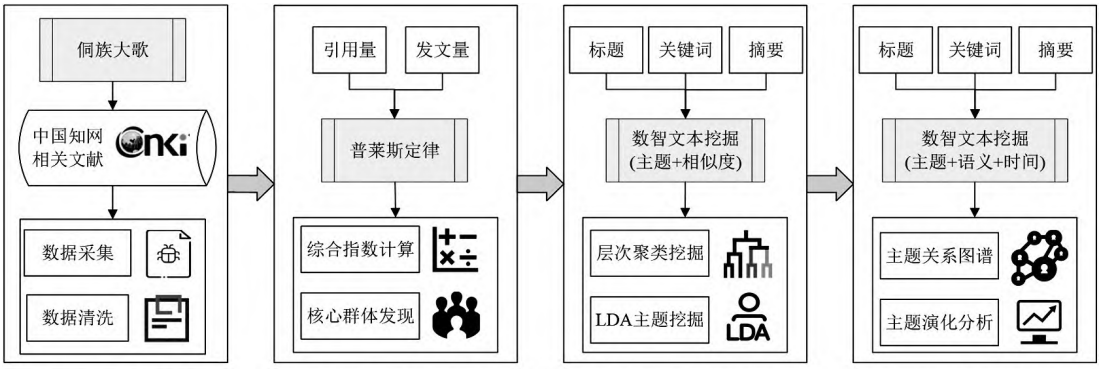

图1 侗族大歌领域主题演化分析整体框架图

3 核心作者群体发现

文献数字化提升学术成果传播速度的同时，一定程度造成学术信息过载的现象，研究专业团队识别程度降低，不利于专业学术的发展。核心作者作为学术研究的基准，决定研究方向和质量。然而，单维度通过发文量，忽略论文质量来确定核心作者一定程度存在片面性。针对上述不足，本文尝试采用一种结合普莱斯定律和综合指数的核心作者确认方法来确认“侗族大歌”核心研究人员。首先，依据普莱斯定理确认核心作者候选人，再结合综合指数遴选“侗族大歌”核心研究人员。

3.1 普莱斯定律分析

本文通过使用普莱斯定律初步确认“侗族大歌”核心研究人员，具体确认方式如下：

⑴ 确认最低被引量

侗族大歌文献被引用次数最高的文献是彭兆荣[5]于1999年发表《中国音乐学》的“族性的认同与音乐的发生”，被引用84次，记C_{max}。结合普莱斯定律确认最低被引用量Mc，详见公式⑴。文献引用量超过七次的作者可视为核心作者的候选人。

$$M_c = 0.749 \times \sqrt{C_{max}} = 0.749 \times \sqrt{84} \approx 6.86 \quad (1)$$

(2) 确认最低发文量

“侗族大歌”发文量最多的作者为吴媛姣，共计发文13篇，记P_{max}。结合普莱斯定律确认最低发文量Mp，详见公式(2)。发表3篇及以上的作者可入选核心作者候选人。

$$M_p = 0.749 \times \sqrt{P_{max}} = 0.749 \times \sqrt{13} \approx 2.70 \quad (2)$$

(3) 核心作者候选人确认

步骤(1)和步骤(2)初步筛选以吴媛姣为代表的178位核心作者候选人，共发文486篇，被引4617次。

3.2 综合指数遴选核心作者

结合综合指数从上述筛选的178位核心作者候选人中遴选出10位“侗族大歌”研究核心作者。具体步骤如下。

(1) 确认平均发文量

178位核心作者候选人总发文量，记$X_{总}$，核心作者候选人数，记n，确认核心作者候选人的平均发文量$\bar{x}$，详见公式(3)。

$$\bar{x} = \frac{X_{总}}{n} = \frac{486}{178} \approx 2.73 \quad (3)$$

(2) 确认平均被引量

178位核心作者候选人所发文献的总被引用量，记$Y_{总}$，核心作者候选人数，记n，确认核心作者候选人的平均被引量$\bar{y}$，详见公式(4)。

$$\bar{y} = \frac{Y_{总}}{n} = \frac{4617}{178} \approx 25.96 \quad (4)$$

(3) 综合指数遴选

结合公式(5)计算各核心作者候选人的综合指数$score_i$。第i位核心作者候选人的发文量记x_i；第i位核心作者候选人的总被引量记y_i。

$$Score_i = \frac{x_i}{\bar{x}} \times 0.5 + \frac{y_i}{\bar{y}} \times 0.5 \quad (5)$$

设置综合指数阈值为2.2，遴选出10位“侗族大歌”的核心作者，详见表1。第一位核心作者为四川音乐学院音乐学系的杨晓[6]，发文10篇，篇均被引19.30次，综合指数5.55，单篇最高被引文献为“南侗‘歌师’述论——小黄侗寨的民族音乐学个案研究”，被引用47次。第二位核心作者为中央音乐学院音乐学系的樊祖荫[7]，发文9篇，篇均被引19.22次，综合指数4.98，单篇最高被引文献为“侗族大歌在中国多声部民歌中的独特地位”，被引用52次。篇均被引次数最高的是桂林理工大学旅游学院的陈炜[15]，发文3篇，篇均被引29.00次，综合指数2.23，单篇最高被引文献为“旅游开发对少数民族非物质文化遗产保护的影响研究——以广西三江侗族自治县为例”，被引用63次。

表1 “侗族大歌”文献核心作者

第一作者	科研群体单位	发文数量	被引数量	篇均被引	综合指数	单篇最高被引文献
杨晓	四川师范大学	10	47	19.30	5.55	南侗“歌师”述论——小黄侗寨的民族音乐学个案研究[6]
樊祖荫	中央音乐学院	9	52	19.22	4.98	侗族大歌在中国多声部民歌中的独特地位[7]
张中笑	贵州省群众艺术馆	7	29	18.57	3.79	侗族大歌研究50年(上)[8]
徐新建	四川大学	6	32	18.00	3.18	无字传承“歌”与“唱”:关于侗歌的音乐人类学研究[9]
吴媛姣	贵州财经大学	13	7	2.69	3.06	侗族音乐文化生态:研究综述及意义[10]
普虹	贵州省榕江县文化馆	7	25	11.29	2.80	侗族大歌——民族的瑰宝[11]
甘明	贵州凯里学院	8	23	8.50	2.78	论非物质文化遗产保护法权利主体制度的构建——以黔东南苗族侗族自治州为例[12]
李延红	中央音乐学院	10	14	4.50	2.70	“国家在场”与侗族嘎老的乡村传承——以贵州省黎平县“十洞”地区两个侗寨为例[13]
乔馨	东北师范大学	8	25	6.13	2.41	论侗族大歌传统音乐文化的传承[14]
陈炜	桂林理工大学	3	63	29.00	2.23	旅游开发对少数民族非物质文化遗产保护的影响研究——以广西三江侗族自治县为例[15]

4 侗族大歌领域主题挖掘

特征主题词作为领域知识的重要知识元，能较好反映该领域研究主题，本文首先尝试使用LDA模型对筛选的“侗族大歌”文献进行主题挖掘研究，之后结合Hierarch算法进行层次聚类主题挖掘。

4.1 基于LDA模型的主题挖掘

首先借助Jieba分词系统，对目标语料进行中文分词处理，结合去停用词表剔除以虚词、数词等语义相

对较低语料，形成特征主题词相对集中的语料数据集。通过困惑度计算，确认“侗族大歌”领域文献的主题包括三类，分别是：区域旅游文化(Topic 1)、侗族音乐民歌(Topic 2)、文化传承保护(Topic 3)。

4.2 基于层次聚类的主题挖掘

层次聚类旨在通过聚类算法将侗族大歌文献中具有相似属性的特征进行聚集，绘制树状主题聚类图。本文采用Sk–leran环境下的Hierarch算法对特征主题进行层次聚类计算，绘制出图2所示层次聚类图，共划分为三个类别：

⑴ T1：以地域文化和旅游地点为主的主题，关键特征包括“肇兴侗寨”、“风雨桥”、“民族文化”等；

⑵ T2：以侗族音乐和音乐特征为主的主题，关键特征包括“多声部民歌”、“侗族音乐”、“民间音乐”等；

⑶ T3：以文化传承和文化保护为主的主题，关键特征包括“非物质文化遗产”、“文化传承”、“侗歌”等。

对比发现层次聚类与LDA主题挖掘效果基本一致，两者均将“侗族大歌”文献分成三类主题。

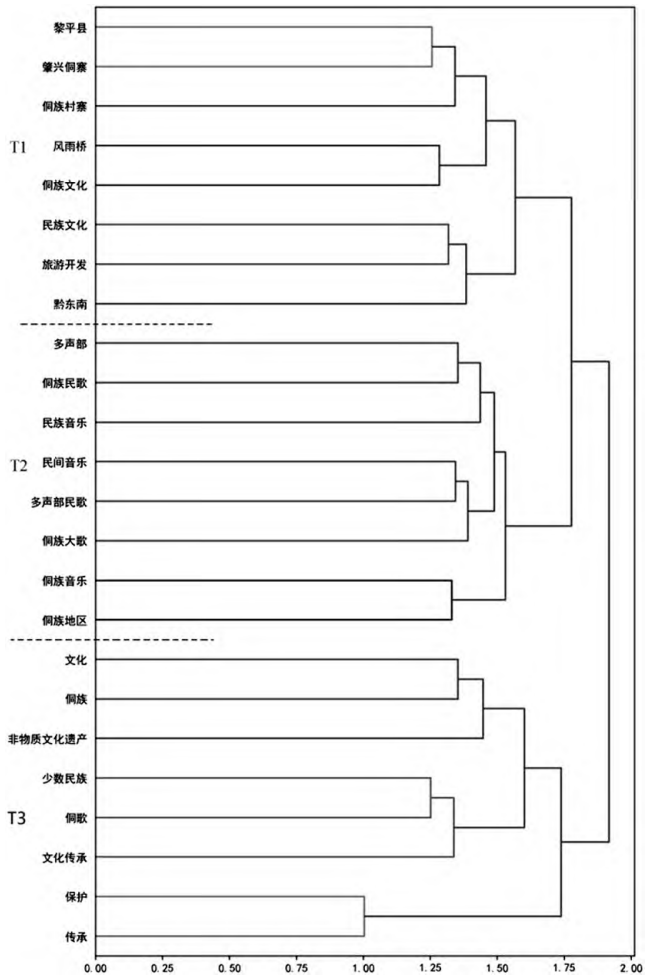

图2 侗族大歌领域主题特征层次聚类效果图

5 侗族大歌领域主题演化分析

侗族大歌领域主题演化分析主要包括构建主题关系图谱和探究各主题间演化关联性。

5.1 主题关系图谱

本文对筛选后的主题特征词构建共词网络，发现主题词间关联性，构建“主题词A—主题词B—关系—权重”四元关系组。由于存在大量低关联性四元关系组，在此利用普莱斯定律设置阈值筛选目标四元关系组，提升主题关系图谱整体聚类效果。

统计发现，主题词“侗族大歌”与主题词“黎平县”共现次数最多，共计出现35次，结合普莱斯定律设置阈值为5，计算公式详见公式(6)。其中F_{max}表现权重系数最高的共现四元组，M_f表示筛选为主题关系图谱的共现四元组权重最低系数。

$$M_f = 0.749 \times \sqrt{F_{max}} = 0.749 \times \sqrt{35} \approx 4.43 \quad (6)$$

最终筛选出88组符合要求的共现四元组，利用Gephi工具绘制图3所示侗族大歌领域主题关系图谱。该图谱分为两大研究领域，分别是以“侗族大歌”为主的民族文化理论研究和以“侗族”为主的民族文化传承与保护实践研究。两大研究领域就“旅游”、“民族文化”、“非物质文化遗产”等主题词将两大主题交叠融合，从侧面反映民间民族文化的保护需重点传承其特有的非物质文化遗产。

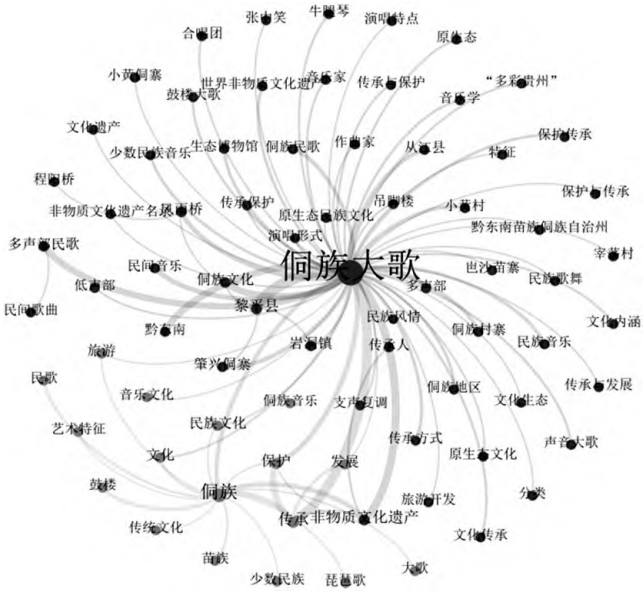

图3 侗族大歌主题关系图谱

5.2 主题演化分析

相比于传统主题演化分析旨在挖掘不同阶段主题的发展历程及变化趋势，本文创新性地结合主题挖掘

的分类结果探究各个主题间的演变关系。

首先，前文已确认将侗族大歌文献主题分为三类，通过采用主题词频和主题贡献度评估主题词贡献度，计算方法详见公式(7)。其中 W_{ij} 表示主题词 i 在 Topic n的贡献度，该值由词频 C_{ij} 与 Topic n下主题词总数的比值来决定。

$$W_{ij} = \frac{C_{ij}}{S_j} \tag{7}$$

其次，按照主题贡献度筛选各主题下贡献前20主题词，能够一定程度代表该主题的活跃度和关联性，结合WordCloud绘制主题词云图。最后，对三大主题词云图深入挖掘，探究各主题间的演化关系，绘制图4所示侗族大歌领域主题演化趋势图，深层次语义挖掘，发现如下结论：

通过对各主题下的高贡献度主题词分析可知，区域旅游文化主题包括“非物质文化遗产”、“侗族文化”、“黎平县”、“肇兴侗寨”等特征；侗族音乐民歌主题包括“传承”、“民族文化”、“多声部民歌”、“民族音乐”等特征；文化传承保护主题包括“侗族民歌”、“黔东南”、“文化传承”、“音乐教育”等特征。

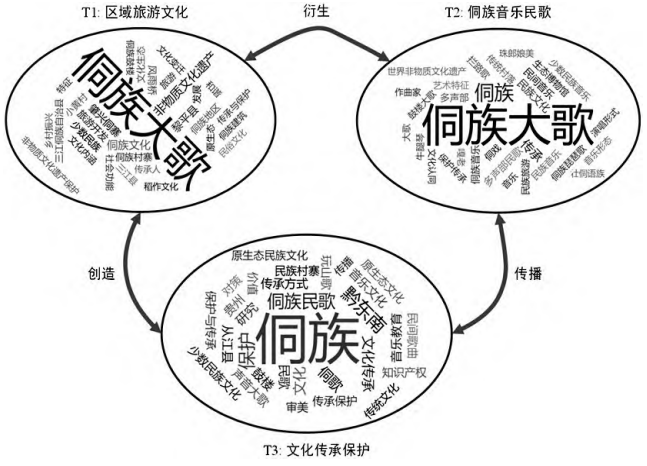

图4 侗族大歌领域主题演化趋势图

通过对各主题之间关联性分析可知，民族区域文化(Topic 1)和侗族音乐民歌(Topic 2)的关联性是“衍生”；侗族音乐民歌(Topic 2)和文化传承保护(Topic 3)的关联性是“传播”；民族区域文化(Topic 1)和文化传承保护(Topic 3)的关联性是“创造”。

6 结束语

针对新范式对传统人文社科的研究认知体系的影响，造成学科边界模糊、共识标准无序、评价体系欠缺等问题，一定程度影响数智赋能研究的发展。本文以“侗族大歌”学术文献作为研究对象，运用计量统计与文本挖掘相结合的研究框架体系探究其主题变化的趋势。

首先，运用普莱斯定律和综合指数的核心作者确认方法，探究“侗族大歌”核心研究人员。研究表明“侗族大歌”的核心研究人员以西南地区为主，多数集中于音乐学院，符合学科研究发展趋势。其次，分别运用LDA模型和层次聚类模型对侗族大歌的领域主题数进行了对比计算，对比结果表明两个模型均将侗族大歌文献内容分成三大类。之后，构建侗族大歌领域主题关系图谱，发现侗族大歌文献的研究呈现以“侗族大歌”和“侗族”两大主题为主，“民族文化”、“非物质文化遗产”等主题词将两大主题交叠融合的研究态势。最后，创新性提出探究各主题间的关联性演化，发现三大主题以“衍生”、“传播”、“创造”三种关联性有效形成一个研究整体。总体而言，本文的研究结论一定程度反映民间民族文化领域的研究变化趋势，为数智赋能语音音频领域研究的深入提供理论基础，同时为民间民族文化保护提供指导性意见。

参考文献(References):

[1] 余艳．中国传统音乐研究——以侗族大歌为例[J]．艺术品鉴，2022(6):55-57
[2] 武帅，任天舒，刘建义，等．基于数据计量和社交网络的图书情报学科发展探究[J]．情报探索，2022(1):28-40
[3] 武帅．科学研究前沿主题识别[D]．硕士，贵州财经大学，2021
[4] 杨秀璋，武帅，宋籍文，廖文婧，任天舒，刘建义．基于LDA和关系图谱的数据治理文献主题演化研究[J]．信息技术与信息化，2022(8):6-12
[5] 彭兆荣．族性的认同与音乐的发生[J]．中国音乐学，1999(3):47-55
[6] 杨晓．南侗“歌师”述论——小黄侗寨的民族音乐学个案研究[J]．中央音乐学院学报，2003(1):89-98
[7] 粟祖荫．侗族大歌在中国多声部民歌中的独特地位[J]．贵州大学学报(艺术版)，2003(2):1-5
[8] 张中笑．侗族大歌研究50年(上)[J]．贵州大学学报(艺术版)，2003(2):33-37
[9] 徐新建．无字传承“歌”与“唱”：关于侗歌的音乐人类学研究[J]．民族艺术研究，2006(1):61-70
[10] 吴媛妓，吐尔洪·司拉吉丁．侗族音乐文化生态：研究综述及意义[J]．贵州民族研究，2014,35(11):95-99
[11] 普虹．侗族大歌——民族的瑰宝[J]．贵州大学学报(艺术版)，2003(2):11-16

(下转第126页)

3.5 算法举例

例句:“这个方案的目的是可以高效准确地实现中文文档的主题词条抽取和词频统计”。经过预处理后句子为:“**这个**/方案/的/**目的**/**是**/**可以**/高效准确地实现/**中文**/文档/**的**/主题词条抽取/**和**/词频统计”(加粗部分为本算法匹配出的高频词)。

4 实验结果及分析

本文的实验是基于Apache Jakarta家族中的开源项目Lucene,实验数据来自搜狗实验室的全网新闻数据(SogouCA)的精简版(一个月数据, 437MB),其数据来自若干新闻站点2020年5月–6月期间奥运、体育、IT、国内、国际等18个频道的新闻数据,提供URL和正文信息。本实验针对正向最大匹配算法,在相同实验环境下,选取不同的数据集,进行三次数据测试,其实验结果见表1。

表1　实验结果数据对比表

		未经过预处理分词	经过预处理分词
第一次测试	正确率(%)	97.623	97.635
	速度(字/min)	117,630	120,528
第二次测试	正确率(%)	96.759	96.750
	速度(字/min)	110,243	116,333
第三次测试	正确率(%)	97.023	97.030
	速度(字/min)	120,335	127,667

从表1可以看出,经过预处理后的变化:①分词速度有明显的提高,证明了此预处理技术的可行性;②分词正确率没有降低,因为此预处理过程同样是基于词典的匹配过程。这说明该方法具有一定的实用性。切分错误原因主要有两个方面:一是未登录到字典中的词;二是含有错别字的字串。

5 结论

随着中文信息处理技术的发展和互联网信息数据的日益增加,对中文分词的速率要求越来越高,作为中文分词基础的词典机制研究已成熟。本文研究现有的基于词典的最大匹配算法的机制,根据高频词的特点,通过提前匹配出所有高频词进而把整个文本分成更多的段,从而提高分词的速度,并且高频词出现次数越多,该算法的性能越好。当然此算法只是在分词速度上有所提高,而对于正向最大匹配算法的分词准确率及未登陆词的识别等没有改善。

参考文献(References):

[1] LI X, MENG Y, SUN X, et al. Is word segmentation necessary for deeplearning of chinese representations?[c]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,2019:359–401

[2] 宋成庆.统计自然语言处理[M].北京:清华大学出版社,2008

[3] 王佳楠,梁永全.中文分词研究综述[J].软件导刊,2021,20(4):247–252

[4] 吴育良.百度中文分词技术浅析[J].河南图书馆学刊,2008,28(4):115–117

[5] 化柏林.知识抽取中的停用词处理技术[J].知识组织与知识管理,2007(8):48–51

[6] 孙茂松,左正平,黄昌宁.汉语自动分词词典机制的实验研究[J].中文信息学报,1999,14(1):1–6

(上接第122页)

[12] 甘明,刻光辉.论非物质文化遗产保护法权利主体制度的构建——以黔东南苗族侗族自治州为例[J].广西民族研究,2009(1):172–176

[13] 李延红.“国家在场”与侗族唱老的乡村传承——以贵州省黎平县“十洞”地区两个侗寨为例[J].中央音乐学院学报,2015(1):35–45

[14] 乔馨.论侗族大歌传统音乐文化的传承[J].东北师大学报(哲学社会科学版),2007(4):109–114

[15] 陈炜,唐景薇.旅游开发对少数民族非物质文化遗产保护的影响研究——以广西三江侗族自治县为例[J].前沿,2010(15):142–149