文章编号:1007-1423(2019)05-0013-05 DOI:10.3969/j.issn.1007-1423.2019.05.003

基于 LDA 模型和文本聚类的水族文献主题挖掘研究

杨秀璋

(贵州财经大学信息学院,贵阳 550025)

摘要:

针对传统民族文献主题识别不精准,缺乏深层次语义理解等问题,提出一种基于 LDA 模型和文本聚类的水族文献主题挖掘算法。通过 Python 抓取中国知网水族文献 990 篇,利用 LDA 模型挖掘水族文献主题分布特征,融合水族特征词典进行文本聚类,并挖掘出水族文化、体育、音乐、医学和水书五大主题的关键词,通过准确率、召回率和 F 特征值进行实验评估。实验结果表明,该方法有效地挖掘出水族文献主题关键词及热门研究领域,使得水族文献的主题脉络更加清晰,为下一步水族引文分析和数字化保护民族文献提供帮助,具有一定的应用前景和实用价值。

关键词:

LDA 模型; 文本聚类; 水族文献; 主题挖掘; 民族研究

基金项目:

贵州省教育厅青年科技人才成长项目(黔教合 KY 字[2016]172)

0 引言

随着科学技术迅速发展,学术成果呈爆炸式增长,互联网上存在着数量庞大、实时更新的学术文献,它们能够通过学术文献数据库进行阅读及下载,如中国知网、万方、维普等。如何从这些文献数据中精准地挖掘出用户所需的信息,获取文献的主题,已经成为了当今研究的热点内容。目前国内外很少利用数据挖掘或机器学习算法深层次分析民族文献,也没有针对水族文献的主题挖掘研究。

本文针对文献数据存在噪声,传统的中文文本分析模型主题识别不精准,数据维度过高,缺乏深层次语义理解等问题,提出了基于 LDA 模型和 K-means 文本聚类的主题挖掘算法。该算法通过 Selenium 和 XPath 技术抓取中国知网 1953-2018 年间 990 篇水族文献信息,再提取文献标题、摘要、关键词,经过中文分词、数据清洗、特征提取等步骤,将文本数据集转换为向量矩阵,最后利用 LDA 主题模型和 K-means 文本聚类算法进行实验分析。实验结果表明,本文提出的算法有效地挖掘出水族文献的主题关键词及研究领域,使文本的主题脉络更加清晰。本文的研究成果具有重要的理论研究意义和实际应用价值,该模型可以广泛应用于文本挖掘、文献分析、民族研究等领域,为水族文化的研究和进一步发展提供相关启示,为后续的水族引文分析和水族文化传承提供有效支撑。

1 相关研究进展

1.1 水族文献

水族是一个具有悠久历史和古朴文化的民族,主要聚居在黔桂滇交界的龙江、都柳江上游地带,长期为世界各国学者所关注。水族地区被誉为“像凤凰羽毛一样美丽的地方”,他们崇拜自然,信仰万物有灵[1]。本文采集了中国知网 1953 年至 2018 年间的 990 篇水族文献,拟通过机器学习算法挖掘出水族文献的主题,数字化保护水族文献,以揭示中国水学的轨迹、内涵、特点、趋势及影响,助促国内外各界相关人士客观地认识中国在世界水学研究体系中的地位,同时提升文本主题挖掘的准确性。

1.2 主题挖掘

主题挖掘是数据挖掘尤其是文本挖掘和舆情分析领域的重要知识,其旨在通过主题模型挖掘与识别出不

同来源文本的主题、关键词、情感分数、聚类类标等[2]。主题模型(Topic Model)通过计算概率来挖掘文本主题,常见的算法包括 LSA 和 LDA,目前主要应用于引文文献挖掘、情感倾向分析、自然语言处理、社交网络短文本分析等领域。

随着机器学习和文本分析的飞速发展,国内外学者对主题挖掘做了大量的研究和实践。在算法创新上,Xu 等[3]提出了一种将非结构化文本数据存储至向量空间模型中,再进行文本聚类的主题挖掘方法;Deerewster 等[4]提出了基于线性代数的主题挖掘算法(Latent Semantic Analysis,LSA),通过数学手段在低维语义空间里对文本进行相关性分析;Blei 等[5]研究出了 LDA(Latent Dirichlet Allocation)主题模型,并被广泛应用于各个领域;王振振等[6]研究了 LDA 主题模型的文本相似度计算,利用 Gibbs 算法进行抽样,挖掘潜在的文本主题与词之间的关系;张晨逸等[7]提出了 MB-LDA 模型方法并挖掘微博主题与人物间的关系。在应用领域上,李霄野等[8]通过 LDA 模型研究文本聚类检索;王树义等[9]通过主题模型挖掘企业新闻文本及情感分析;Shi 等[10]通过 LDA 主题建模分析了企业非结构化业务数据,量化企业在产品、市场和科技空间中的位置;王婷婷等[11]优化了 LDA 模型及其主题数量选择,并通过科技文献进行实验研究。

尽管主题挖掘在算法创新和应用领域都有一些研究,但是国内外很少有利用主题挖掘算法分析民族文献,并且传统的民族文献研究需要消耗大量的资金、人力和时间,无法获取深层次的主题信息,也不能进行精准的文本主题挖掘,处理海量文献效果不理想。本文针对上述问题,提出了基于 LDA 和文本聚类的水族文献主题挖掘算法。

1.3 LDA 模型

LDA 是一种文档主题生成模型,由 Blei 等[5]在 2003 年首次提出,是一种三层贝叶斯结构,包括主题、文档和主题词三层结构,其中文档到主题、主题到词都服从多项分布。LDA 模型将一篇文本的每个词都按照一定概率分布到某个主题上,并从这个主题中选择相关的词语集,如图 1 所示,将 d 篇文档映射到 k 个主题中,每个主题包括一定量的主题词。

LDA 模型表示法称为“盘子表示法”,其模型生成过程如图 2 所示。数据集中每篇文档 D 都与 T 个主题的多项式分布相对应,记为多项分布θ;每个主题都与特征词表中 n 个单词的多项式分布对应,记为多项分布φ,并且θ和φ均存在一个带超参数的α和β的狄利克雷先验分布。图中单圆圈表示潜在变量,双圆圈表示可测变量,箭头表示两个变量之间的依赖关系,矩形框表示重复抽样,对应的重复次数在矩形框的右下角显示。

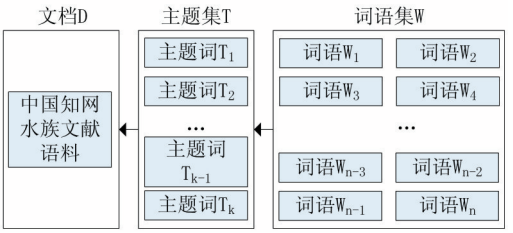

图 1 文档-主题-词映射模型

LDA 模型的具体实现步骤如下:

(1)从每篇文档 D 对应的多项分布θ中抽取每个单词对应的一个主题 z。

(2)从主题 z 中抽取一个单词 w,其主题对应的多项分布为φ。

(3)重复步骤(1)(2),共计 N_d 次,直至遍历文档中每一个单词。

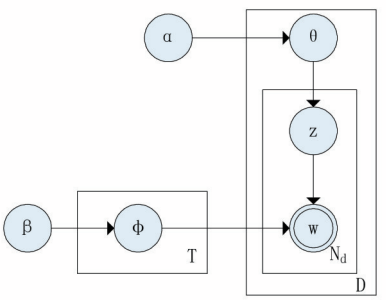

图 2 LDA 主题模型

本文主要研究水族文献的主题挖掘,经过 LDA 主题分布后,得到各个文档的不同主题所占比例,各主题的关键词,从而识别数据集中潜藏的主题信息,计算文献语料相关的主题和每篇文档所涵盖的主题比例。

2 基于 LDA 和文本聚类的主题挖掘算法

本节主要介绍基于 LDA 和文本聚类的主题挖掘算法,重点阐述了本文提出方法的流程,经过数据抓

取、数据预处理、特征提取、权重计算之后，利用 LDA 模型、K-means 算法进行分析。

2.1 基本思路与流程

本文的核心思想是引入了 LDA 主题模型和文本聚类方法分析水族文献信息，识别深层次的关联主题，更好地进行文献挖掘。该算法的框架图如图 3 所示。

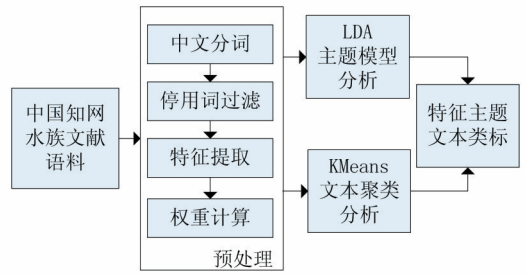

图3 基于LDA和文本聚类的主题挖掘算法框架图

(1)调用 Python、Selenium 和 XPath 技术自动抓取水族文献，并存储至本地保存。

(2)对抓取的文本语料进行数据预处理，包括中文分词、停用词过滤、数据清洗、特征提取、降维等处理，这是文本数据分析的重要处理环节。

(3)将预处理之后的网页文本转换为特征词矩阵，采用 TF-IDF 表示，并进行 LDA 主题模型分析和 K-means 文本聚类分析。LDA 模块可以对语料进行深层次语义挖掘，得到“主题-词”和“文档-主题”的概率分布矩阵；文本聚类分析可以将相似主题的文本聚集在一起，得出关联信息。

2.2 主题挖掘算法

水族文献主题挖掘包括基于 LDA 模型的主题分析和基于 K-means 算法的文本主题聚类。采用 Python 中的 Jieba 工具进行中文分词和数据清洗，通过 TF-IDF 计算特征词权重，该技术计算特征词的重要程度是依据特征词在文本中出现的次数和在整个数据集中出现的文档频率实现的，它能尽可能多地保留影响程度更高的特征词，并过滤掉一些常见却无关紧要的词语。利用 LDA 主题模型分析不同主题的 Top-N 个关键词及“文档-主题”分布，利用 K-means 算法进行文本聚类，旨在根据文档内容的相似性，将无标签的文档自动归类，尽可能地使得同类文档的内容相似性较大，不同类文档的内容相似性较小；最后对输出的“文档-主题-关键词”、文本聚类类标、情感分数进行实验结果评估。

3 实证分析

(1)数据抓取与预处理

本文采集了中国知网 990 篇水族文献，其中学术期刊论文 662 篇，会议论文 106 篇，博士论文 7 篇，硕士论文 91 篇，中国专利 5 篇，科技成果 3 个，报纸 116 篇，所抓取的字段包括文献标题、摘要、关键词、发布时间、文献类型等。

紧接着进行中文分词和数据清洗，将不常用的词语和特色符号进行过滤，并导入关键词字典构建关键词信息，主要利用 Jieba 分词工具进行数据预处理。数据清洗为后面的 LDA 主题模型分析、K-means 文本聚类分析提供良好的数据基础。

(2)评价指标

本实验采用信息检索和机器学习领域常用的性能评估指标对所得的文本数据进行评估，即准确率(Precision)、召回率(Recall)和 F 值(F-measure)。准确率 P(i, j)定义如公式(1)所示，召回率 R(i,j)定义如公式(2)所示。

$$Precision = P(i,j) = \frac{n_{ij}}{n_j} \tag{1}$$

$$Recall = R(i,j) = \frac{n_{ij}}{n_i} \tag{2}$$

其中，n_i 表示类别为 i 的文本数目，n_j 表示聚类 j 的文本数目，n_{ij} 表示聚类 j 中属于 i 的数目。

F 值是准确率和召回率的调和平均值，它平衡了准确率和召回率在特定环境下的制约问题，可用来评估整个实验的最终结果。F 值指的计算公式如公式(3)所示。

$$F-score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

(3)主题挖掘实验

在基于 LDA 模型的主题挖掘实验中，设置的主题数(n_topic)为 5，迭代次数(iterations)为 500，调用 Python 环境下的 LDA 主题模型进行模拟训练，每个主题内的主题词根据其概率大小排序，核心主题词如表 1 所示。由下表可以看出，每个主题之间的区分非常明显。主题 0 中的特征词主要是“水族文化及民俗遗产”方面，包括“水族文化”、“马尾绣”、“文化遗产”、“农耕文化”、“三都”等核心词汇；主题 1 中的特征词主要是

“水族音乐舞蹈”方面，包括“民歌”、“铜鼓音乐”、“古歌”、“酒歌”、“祭祀”等核心词汇；主题 2 中的特征词主要是“水族医学”方面，包括“水族医药”、“遗传”、“中医体质”、“群体遗传学”、“ABO”等核心词汇；主题 3 中的特征词主要是“水书及水族语言”方面，包括“水书”、“水字”、“水族语言”、“祝词”、“水书学”等核心词汇；主题 4 中的特征词主要是“水族体育”方面，包括“传统体育”、“赛马”、“吞口舞”、“水族武术”、“村落集体活动”等核心词汇。

表 1 LDA 模型主题-词识别结果

主题	核心主题词
Topic0 水族文化及民俗遗产	水族文化；民族；民俗；马尾绣；文化遗产；信仰；三都；水书； 传承保护；文化产业；原生态；民族认同；演变；农耕文化； 变迁；铜鼓；图腾；族群；特色旅游；文化变迁；火耕水耨；帮扶；
Topic1 水族音乐舞蹈	水族；音乐；民歌；舞蹈；铜鼓音乐；酒歌；音乐教育；音调特征；丧葬仪式；祭祀；音 乐文化；流变；谬跳圆；古歌；乐器；赛马；三都；音乐语境；铜鼓；传统音乐；四音音 组；端节；敬霞节；
Topic2 水族医学	水族医学；血型；水族医药；遗传；ABO；中医体质；人类学； 冠心病；民间习俗；基因频率；群体遗传学；少数民族；Rh 血型； 遗传性状；特征；染色体；地中海贫血；水族人群；DNA；流行病；
Topic3 水书及水族语言	水书；水字；水语；水族语言；水书教育；水族源流；文字；祝词； 水书学；文献；象形；古籍；民族文化；马尾绣；水书文化；押韵； 文化遗产；社会记忆；古文明；传承；本土传承；易经；银饰；
Topic4 水族体育	水族；传统体育；体育文化；赛马；吞口舞；水族武术；舞蹈；　文化变迁；农耕民族；民 族传统体育；棋类活动；棋类活动； 非物质文化遗产；进化机理；少数民族；村落集体活动；传承；

图 4 展示了中国知网 1953 年至 2018 年水族文献五大主题的词云分布情况，该图清晰地展示了不同主题的热点关键词。

图4 中国知网水族文献五大主题的词云分布图

文本聚类旨在根据文档内容的相似性，将无标签的文档集进行自动归类，水族文献 LDA 模型文本聚类的实验结果如表 2 所示。

表 2 基于 LDA 模型的文本聚类实验结果

主题	准确率	召回率	F 值
水族文化及民俗遗产	0.935	0.941	0.938
水族音乐舞蹈	0.840	0.958	0.895
水族医学	0.830	0.800	0.815
水书及水族语言	0.899	0.810	0.852
水族体育	0.882	0.918	0.900

从表 2 可见，“水族文化及民俗遗产”主题实验结果最好，准确率为 0.935，召回率为 0.941，F 值为 0.938；“水族医学”主题实验效果不太理想，准确率为 0.830，召回率为 0.800，F 值为 0.815。由于部分水族文献融合了多个主题，呈跨主题分布，所以会常出现主题识别不精准的情况。各主题文本聚类的实验结果对比图如图 5 所示。

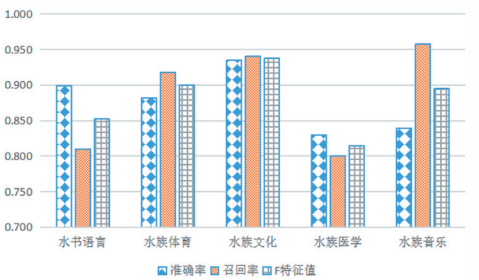

图5 各主题文本聚类实验结果对比

水族文献的 K-means 文本聚类效果如图 6 所示，共将文本聚集成五类主题。其中方块表示“水族文化及民俗遗产”主题，圆形表示“水书及水族语言”主题，六边形表示“水族医学”主题，五角星形表示“水族体育”主题，菱形表示“水族音乐舞蹈”主题。从图中可以看到五个类簇有效地分隔开来，相似主题的文献聚集在一起，文本聚类效果良好。

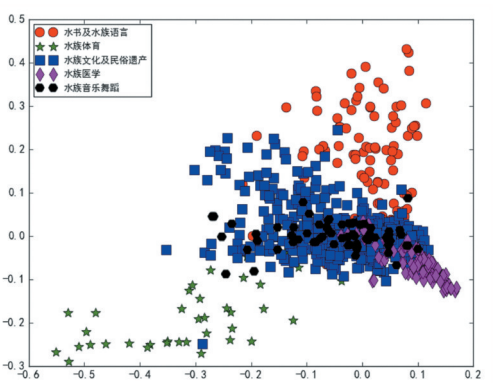

图6 K-means 文本聚类效果图

4 结语

本文的研究成果主要应用于民族文献的文本挖掘分析，以水族文献为例进行主题分布和文本聚类的研究。实验结果表明，本文提出的算法有效地挖掘出水族文献的主题关键词及研究领域，提取了五大主题的

核心主题词，使得文本的主题脉络更加清晰，同时计算出各水族文献的聚类类标，当给出一篇新的水族文献时，我们能及时实现主题分类和关键词词挖掘。当然，本研究也存在不足，一方面没有深层次理解文本语义信息，另一方面没有进行情感分类打分等，后续将继续深入研究。

总之，本文的研究成果具有重要的理论研究意义和实际应用价值，该模型可以广泛应用于民族文献研究、主题分析和文本挖掘等领域，更好地帮助科研工作者、高校师生和民族研究者进行相关研究，实现文本知识主题挖掘，为下一步水族引文分析和数字化保护民族文献提供帮助。

参考文献：

[1]潘朝霖. 水族鱼图腾析[J]. 广西民族研究,2001(3):65-69.

[2]王树义,廖桦涛,吴查科. 基于情感分类的竞争企业新闻文本主题挖掘[J]. 数据分析与知识发现,2018(3):70-78.

[3]Xu R,Wunsch D. Survey of Clustering algorithms[J]. IEEE Trans on Neural Networks,2005,16(3):645-7678.

[4]Deerwester S,Dumais S,Landauer T,et al. Indexing by Latent Semantic Analysis[J]. Journal of the American Society of Information Science,1999,41(6):391-407.

[5]Blei D M,Ng A Y,Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research,2003,3:993-1022.

[6]王振振，何明，杜永萍. 基于 LDA 主题模型的文本相似度计算[J]. 计算机科学，2013,40(12)：229-232.

[7]张晨逸，孙建伶，丁轶群. 基于 MB-LDA 模型的微博主题挖掘[J]. 计算机研究与发展，2011，48(10)：1795-1802.

[8]李霄野,李春生,李龙,张可佳. 基于 LDA 模型的文本聚类检索[J]. 计算机与现代化,2018,6:7-11.

[9]王树义,廖桦涛,吴查科. 基于情感分类的竞争企业新闻文本主题挖掘[J]. 数据分析与知识发现,2018,3:70-78.

[10]Shi Z M,Lee G,Whinston A B. Toward a Better Measure of Business Proximity: Topic Modeling for Industry Intelligence[J]. MIS Quarterly,2016,40(4):1035-1056.

[11]王婷婷,韩满,王宇. LDA 模型的优化及其主题数量选择研究——以科技文献为例[J]. 数据分析与知识发现,2018,1:29-39.

作者简介：

杨秀璋(1991-)，男，贵州凯里人，硕士，，研究方向为 Web 数据挖掘、知识图谱、文献分析

收稿日期：2019-01-15　　修稿日期：2019-01-22

Research on the Shui Literature Topic Mining Based on LDA Model and Text Clustering

YANG Xiu-zhang

(School of Information, Guizhou University of Finance and Economics, Guiyang 550025)

Abstract：

Aiming at the inaccurate recognition of traditional national literature topics and the lack of deep semantic understanding, proposes a Shui literature mining algorithm based on LDA model and text clustering. Grabs 990 Shui literature from CNKI by Python, uses the LDA model to explore the distribution characteristics of Shui literature, integrates the feature dictionary for text clustering, and excavates five key themes of Shui culture, sports, music, medicine and Shui word. Carries out experimental evaluation by precision, recall and F-measure. The experimental results show that the method proposed effectively mines the topic keywords and popular research fields of Shui literature, which makes the theme of Shui literature more clear, and provides help for the next step of citation analysis and digital protection of national literature. It has certain application prospects and practical value.

Keywords：

LDA Model; Text Clustering; Shui Literature; Topic Mining; Ethnic Studies