·博士论坛·

基于细粒度语义实体的学术论文推荐研究

李晓敏[1,2],王　昊[1,2],李跃艳[1,2]

(1.南京大学 信息管理学院,江苏 南京 210023; 2.江苏省数据工程与知识服务重点实验室,江苏 南京 210093)

摘　要:【目的/意义】为帮助科研用户快速准确地找到与自身研究兴趣相关的学术论文,构建了基于细粒度语义实体的学术论文推荐模型。【方法/过程】将实验前期识别出的研究主题、研究对象和理论技术类语义实体作为学术论文和核心作者的内容特征,分别利用TF-IDF算法、TextRank算法和LDA模型得到学术论文和核心作者的特征词,利用Word2vec对特征词进行向量化,再计算核心作者和学术论文的余弦相似度,将余弦相似度值靠前的Top20推荐给作者。【结果/结论】利用准确率、召回率和F值对基于三种算法得到的特征词生成的推荐结果进行比较评价,结果表明,基于TF-IDF算法得到的特征词生成的推荐效果最佳,并对推荐结果进行了实例展示,可以看出本文提出的推荐模型能够更为全面地为科研用户推荐与其研究兴趣类似的学术论文,提高科研效率。【创新/局限】本文主要是从学术论文的内容特征入手,对类型细分后的关键词利用不同算法进行核心作者特征词筛选,进而实现学术论文推荐,但是对学术论文中包含的网络关系并未涉及。

关键词:特征词;核心作者;学术论文;个性化推荐;相似度

中图分类号:G252.8　　**DOI:**10.13833/j.issn.1007-7634.2022.04.019

1 引　言

互联网和大数据等科学技术的发展,为学术数据的获取和阅读带来了极大的便利。学术数据包括学术论文、专利、图书和著作者等,其中学术论文作为科学思想和方法技术传播的重要载体,能够帮助科研人员快速了解某一研究领域的研究前沿和研究热点,并对自己感兴趣的研究方向进行更进一步的论文追踪,加深对该研究方向的了解和掌握,以便确立自己的研究兴趣,顺利开展科学研究。但同时学术论文数量的激增也为科研用户带来一定的负担,科研用户在查找符合自身科研需求的论文时,会出现需求相关和需求不相关的结果,并且由于结果较多且杂,科研用户需要花费大量的时间和精力进行筛选,造成了科研效率的低下,不利于科研产出。而个性化推荐作为解决数据过载的有效手段之一,同样也可以应用到学术论文的推荐中,解决学术数据过载的问题。

目前,关于学术论文的推荐研究主要集中在三个方面,第一个方面是从内容特征的角度实现推荐。文献【1】利用TF-IDF从学术论文的摘要中提取关键词语并将其向量化,论文与论文以及论文与科研用户进行相似度计算,并将相似度融入到负例抽取和概率矩阵分解过程中生成推荐;文献【2】利用TF-IDF和Word2vec对论文的标题、关键词和摘要进行作者和论文的特征表示,并将时间权重和相似度融合实现推荐;文献[3]将LDA模型与Apriori算法结合挖掘出频繁主题集,考虑用户对频繁主题集的偏好进行学术论文的推荐;文献[4]针对科技论文文本内容,基于t-SNE和模糊聚类实现科技论文推荐;文献[5]将关键词的语义类型和学术论文时间价值进行学术论文对的推荐;文献[6]利用LSA将论文全文表示为向量空间,与用户兴趣模型计算得到的相似度作为相关性指标,并与引文分析法计算得到的多样性指标结合生成最终的推荐文献列表;文献【7】对论文文本利用GRU方法进行向量映射,进而实现学术论文推荐。第二个方面是从网络特征的角度实现推荐。文献【8】提出了一种基于作者、论文等信息构建的异质网络嵌入的学术论文推荐方法;文献【9-10】将引文网络图模型与重启随机游走算法结合实现论文推荐;文献[11]利用目标论文与引文之间的潜在关联实现推荐;文献[12]将作者和论文构成异构图,利用Doc2vec和元路径方法进行节点表示和相似性计算,进而实现论文推荐。第三个方面是将内容特征和网络特征混合进行学术论文推荐。文献【13】将用户偏好和论文引用关系结合实现科技论文推荐,并设计了一个原型系统;文献[14]在对论文进行特征提取的

收稿日期:2021-09-29

基金项目:国家社科基金重大项目"面向国家战略的情报学教育和发展研究"(20&ZD332);江苏省"六大人才高峰"高层次人才项目("Six Talent Peaks"Project in Jiangsu Province)"多粒度学术对象区分性测度和分析研究"(JY-001)。

作者简介:李晓敏(1996-),女,山西吕梁人,博士研究生,主要从事网络信息组织与检索研究;王昊(1981-),男,浙江义乌人,教授,博士生导师,主要从事语义网知识组织及应用研究,通讯作者:ywhaowang@nju.edu.cn;李跃艳(1991-),女,山西大同人,博士研究生,主要从事语义网知识组织及应用研究。

基础上，构建属性关系图，再利用PPR和SVD++图算法进行论文排序和推荐；文献【15】将论文语义间的相似性与图模型中的矩阵进行融合，并与重启随机游走算法结合实现论文推荐。由于本文重点基于内容特征进行研究，因此接下来主要论述目前研究中基于学术论文内容特征进行推荐的不足之处。在基于内容特征实现学术论文推荐的相关研究中，大多数都是直接利用学术论文的作者直接给定的标题、关键词或者摘要等可以作为文本类型进行挖掘的数据。在将关键词作为文本挖掘时，一般期刊对关键词的要求停留在数量上，比如关键词不能少于多少个，对于关键词的类型并没有明确的要求，这就导致学术论文的关键词参差不齐，有的作者只列出研究主题和研究对象类型的词语作为关键词，有的作者只列出研究主题和理论技术类型的词语作为关键词，有的作者可能三种类型的词语都作为关键词，而对于科研用户来说，在寻找满足自身需求的学术论文时，每种类型的关键词都是至关重要的。

基于以上论述，本文提出一种基于细粒度语义实体的学术论文推荐模型。本文将语义实体细分为研究对象、研究主题和理论技术三种类型，在实验前期利用深度学习模型从题名+摘要中识别出上述三种类型语义实体的基础上，将识别出的细粒度语义实体作为论文和作者的内容特征，分别利用TF-IDF算法、TextRank算法和LDA模型进行特征词识别，识别之后，利用Word2vec对三种不同算法识别出的特征词进行向量化，利用余弦相似度进行论文与作者的相似度计算，选择相似度靠前的Top20生成推荐，再利用准确率、召回率和F值对三种算法生成的推荐结果进行比较评价，并进行模型的有用性评价。

2 基于细粒度语义实体的学术论文推荐模型框架

本文的推荐模型框架包括数据收集、数据预处理、实验前期实体识别、核心作者特征词筛选、核心作者特征词向量化、学术论文特征词向量化、相似度计算和推荐结果生成八个部分，具体的框架模型如图1所示。

2.1 学术论文推荐模型框架

以图情档CSSCI来源期刊刊载的论文为数据来源，为避免作者发表论文数量太少导致关键词数量不足以支撑特征词筛选的问题以及保证足够的核心作者数量，因此本文选择发表论文数大于等于15的作者作为本文研究的核心作者，以实验前期利用深度学习模型BiLSTM+CRF对期刊论文的摘要+题名进行实体抽取得到的研究对象类、研究主题类和理论技术类细粒度实体作为每篇论文的特征词，将核心作者撰写的每篇学术论文的特征词组合，作为核心作者的研究兴趣，再分别利用TF-IDF算法、TextRank算法和LDA模型筛选核心作者的特征词，筛选完成之后，利用Word2vec模型将学术论文以及三种算法得到的核心作者的特征词向量化，分别得到学术论文的特征向量和核心作者的特征向量，再计算两者的余弦相似度，并将余弦相似度值从高到低进行排序，选择余弦相似度值靠前的Top20生成推荐结果。再将三种算法生成的推荐结果利用准确率、召回率和F值进行对比评价，并选择推荐效果最好的进行推荐结果的展示，以及进行模型有用性评价。

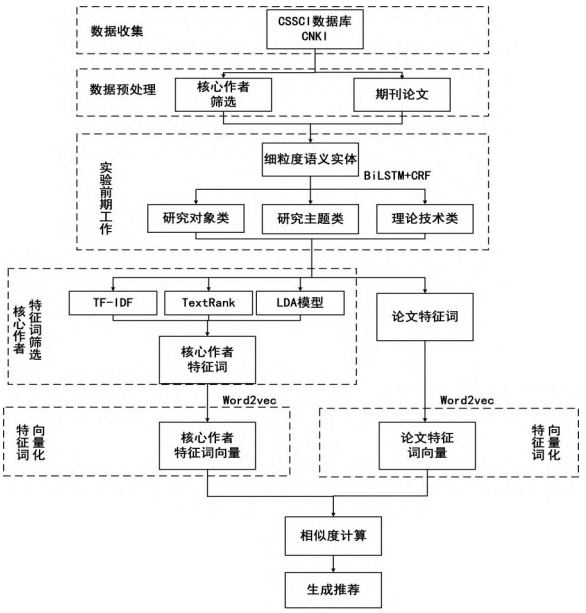

图1 推荐模型框架

Figure 1 Recommendation model framework

2.2 核心作者特征词筛选

作者撰写的学术论文能够较为直接和准确地表征作者的研究兴趣，因此学术论文内容的表征方式就显得尤为重要。一般地，学术论文的内容表征通常是用论文作者所提供的关键词。每本期刊对关键词的选择标准没有统一的界定，因此每位作者在选取论文关键词时也具有一定的随意性，可能会造成学术论文的内容揭示不全面，比如在选取关键词时，只选取了学术论文涉及的研究对象或研究主题，未选取学术论文所用到的理论模型或者技术方法类关键词。因此本文在实验前期将语义实体划分为研究主题、研究对象和理论技术三种类型，并利用深度学习模型BiLSTM+CRF对题名+摘要识别出这三种细粒度语义实体作为学术论文的内容表征，将作者撰写的学术论文形成集合，将学术论文集的研究主题、研究对象和理论技术类实体作为作者的研究兴趣，分别利用TF-IDF、TextRank和LDA模型对作者研究兴趣进行挖掘，筛选出作者研究兴趣的特征词。

2.2.1 TF-IDF算法

TF-IDF算法用来衡量某个词语在文档集中的重要程度，词语的重要性随着其在文档中出现次数的增加而增加，但同时会随着其在文档集中出现频率的增加而下降。TF-IDF算法可以很好地将具有类别区分能力的词语筛选出来。

因此,本文利用TF-IDF算法筛选出核心作者研究兴趣的特征词。TF-IDF的计算方式如公式(1)所示。

$$Wti = \mathrm{tf(ti,d)} * \log \frac{|D|}{df(ti)} \quad (1)$$

其中,t_i表示核心作者学术论文集中的语义实体,W_{t_i}表示语义实体t_i的权值,$tf(t_i,d)$表示语义实体t_i在学术论文d中的出现次数,$df(t_i)$表示语义实体t_i在学术论文集中的出现频率,$|D|$表示学术论文集数。

在完成核心作者的每个语义实体的TF-IDF值计算后,选取TF-IDF值靠前的Top20作为核心作者研究兴趣的特征词。

2.2.2 TextRank

TextRank算法与PageRank息息相关,PageRank是一种基于图的排序算法,被用来衡量网页的重要程度。TextRank算法借鉴PageRank算法思想,利用文本代替网页,将文本中的词语看作节点,将词语间的共现关系转换成节点间的共现关系,建立任意两个节点之间的图模型,再利用公式(2)迭代计算各个节点的权值,直至收敛,最后对节点的权值进行排序。本文将核心作者发表的学术论文中的语义实体作为文本,利用PageRank算法计算语义实体的TR值,并进行排序,选择其中TR值靠前的Top20的词语作为核心作者研究兴趣的特征词。上述中提到的图模型中的点即为学术论文中的语义实体。

$$TR(v_i) = (1-d) + d * \sum_{v_j \in In} \frac{w_{ji}}{\sum v_k \in Out(v_j) w_{jk}} TR(V_j) \quad (2)$$

其中,$TR(v_i)$表示词语v_i的权值,d为阻尼系数,取值范围为0到1,代表从图中某一特点节点指向其他任意点的概率,一般取值为0.85。w_{ji}表示任意两点v_i、v_j之间边的权重,$In(v_i)$表示指向点v_i的点的集合,$Out(v_j)$表示点v_j指向的点的集合。

2.2.3 LDA模型

LDA(Latent Dirichlet Allocation)模型是Blei等人在2003年提出的,是一个三层贝叶斯概率生成模型,其基本思想是假设每个文档都是由多个主题以一定的概率组合而成的,而每个主题又是由多个词语以一定的概率组合而成的。LDA模型的原理图如图2所示[16]。

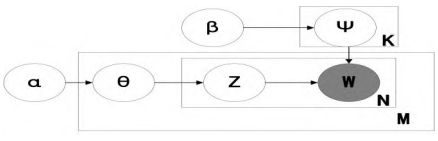

图2 LDA原理图

Figure 2 LDA schematic diagram

图2中,K表示主题个数,M表示文档总数,N_m表示第m个文档的单词总数,α表示每个文档下主题的多项分布的Dirichlet的先验分布,θ表示文档中的主题分布向量,β表示每个主题下词的多项分布的Dirichlet的先验分布,$Z_{m,n}$是第m个文档中第n个词的主题,$W_{m,n}$是m个文档中的第n个词。

本文将同一位核心作者的学术论文合在一起作为一篇文档,多位核心作者的学术论文组合成文档集,利用LDA模型对该文档集进行主题提取,得到主题—词汇分布,以及每篇文档的文档—主题分布,再将文档—主题分布矩阵与主题—词汇分布矩阵对应相乘,得到文档—词汇分布矩阵。这里的一篇文档就代表一位核心作者,选择文档—词汇分布矩阵中权值靠前的Top20个词汇作为每位核心作者研究兴趣的特征词。

2.3 特征词向量化

在分别利用TF-IDF算法、TextRank算法和LDA模型对每位核心作者筛选研究兴趣的特征词之后,需要对筛选出的特征词进行向量化。由于one-hot编码方式存在维数灾难和不能很好地特征词语与词语之间的语义关系的缺陷,本文选择分布式编码进行特征词向量化。Word2vec是分布式编码的典型代表,它是一种神经网络语言模型,通过模型训练,可以将语料库中的词语全部映射成一个低维实数向量,并且向量与向量之间的空间距离可以反映词语与词语之间的语义关系[17]。Word2vec包括CBOW和Skip-grow两种模型,训练方式包括负采样和层序softmax。两种模型中,CBOW是已知上下文,对当前单词进行预测,而Skip-grow与之相反,已知当前单词,对当前单词的上下文进行预测。本文利用Word2vec模型对利用不同方法筛选出的核心作者的特征词进行向量化,得到每位核心作者的向量表示,同时对学术论文也利用Word2vec进行向量化表示,以便于进行作者和学术论文相似度的计算。

2.4 相似度计算及推荐结果生成

在得到核心作者的特征词向量和学术论文的特征词向量之后,接着就是利用余弦相似度的计算方式来计算向量与向量之间的余弦相似度,由于是利用三种不同算法对核心作者特征词进行筛选,因此需要分别计算基于三种算法得到的特征词转换成的词向量与学术论文特征词的词向量之间的相似度,并对相似度从高到低进行排序,选择相似度值靠前的Top20篇学术论文推荐给核心作者,同时对三种推荐结果利用准确率、召回率和F值进行推荐效果比较,并选出三者中推荐效果较好的进行模型有用性评价和推荐效果展示。

3 基于细粒度语义实体的学术论文推荐实证研究

3.1 数据收集

本文的数据来源一部分是实验前期收集的2015—2019年18本图情档期刊论文,并利用深度学习模型BiLSTM+CRF对其中的摘要+题名数据进行实体抽取,得到每篇论文的研究对象类、研究主题类和理论技术类实体[18]。这里对这部分数据仅进行展示,具体过程不展开论述。一共对17992篇学术论文进行语义实体识别,部分数据如表1所示。

表1 学术论文的语义实体识别结果

Table 1 Semantic entity recognition results of academic papers

编号	object	subject	method
1	高校图书馆 双一流 大学建设 战略目标 科学数据服务 情报服务 智库建设 创客空间 产学研 协同创新 竞争情报服务	NULL	NULL
2	研究图书馆 智库 浙江大学 图书馆 高校 高等教育 大环境 智库职能 决策咨询 研究机构	NULL	NULL
3	馆配 中文 电子书 田田网 运营模式 图书馆需求 图书馆	阅读推广	NULL
4	中山大学 图书馆 阿拉伯语 馆藏建设 资源共享 文献保护 数字图书馆 图书出版 人才培养 馆藏	NULL	NULL
5	北京大学 期刊网 图书馆 法律 内容建设 平台建设 学术交流	开放获取	NULL
…	…	…	…
17992	中国 智库理论 热点主题 CSSCI 数据库 评价体系 智库建设 智库 新型智库建设 国际话语权 智库评价 功能定位 评价主体	NULL	利益相关者

表2 265位核心作者及其论文

Table 2 265 core authors and their papers

编号	姓名	发表论文数	论文篇名
1	李纲	83	公共事件意见领袖的认知相符与失调研究 基于类h混合中心性指标改进的作者影响力测度研究 突发传染病微博影响力的预测研究 … 国内突发事件预警研究评述
2	朱庆华	80	网络谣言话题传播与网民行为协调演进研究 情报学的创新与发展——第五届全国情报学博士生论坛会议综述 基于交互视角的O2O电子商务服务质量评价研究 … 知识图谱在数字人文中的应用研究
3	赵宇翔	78	国外社会化搜索引擎比较研究 移动互联环境下的跨屏行为研究综述 元分析方法在社会化媒体采纳和使用中的应用探索 … 基于层次分析法的公众科学项目游戏化设计的评价指标体系构建
…	…	…	…
265	朱光	15	社交网络环境下隐私保护投入的博弈策略分析——基于演化博弈的视角 大数据环境下社交网络隐私风险的模糊评估研究 网络环境下多媒体资源版权管理系统的设计与实现 … 隐私忧虑背景下的移动医疗APP使用意愿研究——基于三方博弈的视角

另一部分是作者及其发表的学术论文和学术论文对应的语义实体。

3.2 数据预处理

针对第二部分数据来源,即作者与其对应的学术论文和学术论文的语义实体,首先是对作者进行人工消歧,针对同名不同机构的作者,利用ORCID、邮箱、性别、出生年月、职称来判断其是否为同一个人。在只有职称信息进行判断时,将论文的发表时间作为辅助信息,若作者在距离当前时间较远的职称高于距离当前时间较近的职称,则可判断为不是同

表3 核心作者论文及其语义实体

Table 3 Core author papers and their semantic entities

编号	论文	论文实体
1	公共事件意见领袖的认知相符与失调研究	公共事件 意见领袖 社会心理学 认知失调理论 K-Means 聚类 评论主题
	基于类h混合中心性指标改进的作者影响力测度研究	类h混合中心性 影响力 合著网络 核心期刊 合著
	突发传染病微博影响力的预测研究	突发传染病 微博 影响力 传染病 风险 LDA 主题特征 时间特征 决策树
	...	...
	国内突发事件预警研究评述	突发事件 政府监测 危机 预警 工作流程 预警模型
2	网络谣言话题传播与网民行为协调演进研究	话题传播 网民 行为协调 网络谣言 传播系统 服务商 实验研究
	情报学的创新与发展——第五届全国情报学博士生论坛会议综述	情报学 博士生 论坛会议 学术交流 中国人民大学 互联网 大数据 信息资源管理
	基于交互视角的O2O电子商务服务质量评价研究	电子商务 服务质量评价 服务质量 影响因素 问卷调查 探索性因子分析
	...	...
	知识图谱在数字人文中的应用研究	知识图谱 数字人文 RDF 语义知识图谱 关联数据 图数据库 广义知识图谱 谷歌知识图谱 中国历代人物传记资料库 关联数据平台 推理规则 图运算
3	国外社会化搜索引擎比较研究	社会化 搜索引擎 Social Mention Social Searcher 社会化媒体 Yoono Wajam 社会化网络 盈利模式
	移动互联环境下的跨屏行为研究综述	移动互联 跨屏行为 用户信息行为 移动用户 复杂信息行为 情境 用户行为
	元分析方法在社会化媒体采纳和使用中的应用探索	元分析 社会化媒体 用户使用行为 影响因素 实证研究 隐私 风险
	...	...
	基于层次分析法的公众科学项目游戏化设计的评价指标体系构建	层次分析 公众科学项目 游戏化设计 评价指标体系 评价指标 科研 众包 游戏
...	...	...
265	社交网络环境下隐私保护投入的博弈策略分析——基于演化博弈的视角	隐私保护 演化博弈 社交网络平台 成本 社交网络 参与主体 行为策略 Matlab
	大数据环境下社交网络隐私风险的模糊评估研究	大数据环境 社交网络 隐私风险 大数据 隐私 风险评估 专家访谈
	网络环境下多媒体资源版权管理系统的设计与实现	网络环境 多媒体资源 版权管理系统 版权保护需求 版权管理 数字水印技术 版权保护
	...	...
	隐私忧虑背景下的移动医疗APP使用意愿研究——基于三方博弈的视角	隐私 移动医疗APP 使用意愿 三方博弈 隐私泄露 隐私保护 演化博弈 政府行为策略 演化稳定策略 保护行为 监管行为

表4 基于TF-IDF算法的特征词

Table 4 Feature words based on TF-IDF algorithm

编号	特征词
1	应急决策 突发事件 微博 微信群 网络结构 LDA 影响因素 信息传播 微信 智慧城市 大数据时代 情报学 情报服务 情报 应急管理 社会网络分析 情报体系 实证研究 网络舆情 可视化
2	影响因素 移动视觉搜索 实证研究 社会化搜索 数字人文 社会化媒体 移动互联 情境 互联网 数字图书馆 公众科学项目 信息素养 关联数据 众包 用户 问卷调查 用户行为 用户使用行为 扎根理论 社交媒体
3	公众科学项目 影响因素 实证研究 数字人文 互联网 社会化媒体 公众科学 移动视觉搜索 众包 移动互联 信息素养 情境 用户信息行为 转移行为 案例分析 运作流程 数字图书馆 用户体验 研究热点 用户
4	数字图书馆 移动图书馆 信息接受 知识发现 情境 信息接受情境 场景化 数字资源 数字资源聚合 微服务 信息需求 知识服务 影响因素 服务创新 数据驱动 可视化 用户 图书馆 场景 深度聚合
5	新媒体环境 信息生态 影响因素 实证研究 网络舆情 信息传播 微博 研究热点 新媒体 可视化 信息人 知识图谱 微信 网络社群 发展动态 移动互联网 社交媒体 用户 新浪 舆情
...	...
265	演化博弈 图博档 隐私 版权 隐私保护 社交网络 Matlab 大数据环境 大数据 版权保护 参与主体 行为策略 数值仿真 信息生命周期 图像 图书馆 社交网络平台 隐私风险 风险评估 专家访谈

一个人,反之则无法判断。在对作者进行消歧后,选择2015-2019五年内发文量大于等于15的265位作者作为本文研究的核心作者,部分数据如表2、表3所示。

3.3 核心作者特征词筛选

3.3.1 基于TF-IDF算法的特征词筛选

将表3中每位核心作者发表的每篇学术论文的语义实体组合在一起,通过公式(1)计算得到语义实体的TF-IDF权值,并将权值从高到低进行排序,选择权值靠前的Top20作为核心作者研究兴趣的特征词,部分数据如表4所示。

3.3.2 基于TextRank算法的特征词筛选

同理,对每位核心作者的每篇学术论文的语义实体组合在一起的文本,利用公式(2)计算得到每个语义实体的TextRank权值,并将权值从高到低进行排序,选择权值靠前

表5 基于TextRank的特征词

Table 5 Feature words based on TextRank

编号	特征词
1	政府 突发事件 情报学 情报服务 网络结构 生命周期 情报 语言学 影响力 智库 参考文献 引文 舆情 运行机制 应急 情境 高频词 用户 词表 预警
2	互联网 情境 服务 档案 游戏 服务平台 服务质量 供给 空间 方法论 档案馆 社会化 反应 跨学科 转译 老年人 政府 网民 图书馆 搜索引擎
3	互联网 游戏 用户 情境 政府 信任 档案 学生 可视化 信息管理 运行机制 期刊论文 澳大利亚 社会化 反应 转译 老年人 激励 报纸 搜索引擎
4	情境 标签 用户 图书 可视化 大学 虚拟社区 信息时代 本体 资源共享 边缘 政府 核心 场景 计算机科学 情报学 网络结构 图书馆学 图书馆 特征分析
5	企业 舆情 美国 专利 网络结构 影响力 用户 研究生 中美 清华同方 课程体系 行业 预警 系统 数据库 社交能力 消费者 满意度 突发事件 理论
...	...
265	图像 版权 图书馆 版权保护 博物馆 档案馆 成本 风险管理 企业 生命周期 水印 激励机制 彩色

表6 主题—词汇分布

Table 6 Topics - Vocabulary Distribution

编号	词1	概率	词2	概率	词3	概率	词4	概率	...	词20	概率
1	科学数据管理	0.000954	科学数据监管	0.000805	科研范式	0.000691	科学数据关联	0.000631	...	KRDS模型	0.000352
2	概念设计	0.002478	主题抽取	0.002352	知识流	0.002275	科学文献	0.001702	...	标题词	0.000522
3	用户感知	0.010329	公共档案馆	0.007760	服务质量	0.005443	知识元	0.004637	...	档案服务	0.001198
4	公共图书馆	0.014702	图书馆	0.006851	图书馆事业	0.006605	公共文化服务	0.006576	...	中华人民共和国公共图书馆法	0.003135
5	Altmetrics	0.010728	图书情报学	0.007895	引文	0.007071	图书情报领域	0.006700	...	被引频次	0.004086
...	...	...	..	...	...	...	...	...	...	...	...
28	智慧城市	0.008134	智库	0.007616	智库建设	0.006916	总体国家安全观	0.005413	...	新型智库建设	0.001752

表7 文档—主题分布

Table 7 Documents - Topic Distribution

编号	主题1	主题2	主题3	主题4	主题5	...	主题28
1	0.005429	0.005470	0.193729	0.007388	0.167948	...	0.006398
2	0.010020	0.010095	0.011047	0.013636	0.016226	...	0.011808
3	0.007460	0.007516	0.008225	0.368649	0.012080	...	0.008791
4	0.005544	0.005586	0.006113	0.038378	0.117036	...	0.006534
5	0.004558	0.004592	0.005025	0.006203	0.007381	...	0.005371
6	0.009089	0.009158	0.010021	0.012369	0.114729	...	0.010711
7	0.003981	0.004011	0.004390	0.005418	0.735117	...	0.004692
8	0.005157	0.005196	0.005686	0.007018	0.050265	...	0.006077
9	0.007776	0.007834	0.008573	0.143908	0.012592	...	0.009163
10	0.007306	0.445428	0.008055	0.009942	0.011831	...	0.008610
...	...	...	...	...	...	...	...
265	0.010169	0.010246	0.011212	0.013839	0.073709	...	0.011984

的Top20作为核心作者研究兴趣的特征词,部分数据如表5所示。

3.3.3 基于LDA模型的特征词筛选

将每位核心作者的每篇学术论文的语义实体组合在一起形成文本,利用LDA模型进行特征词的筛选。LDA模型作用于学术论文文档,生成文档—主题分布和主题—词汇分布,再将这两个分布对应相乘形成文档—词汇分布。LDA模型首先就是要确定主题数,本文利用困惑度来确定最优主题数,设置主题数K=10,20,30,绘制的困惑度曲线如图3所示。

表8 基于LDA模型的特征词

Table 8 Feature words based on LDA model

编号	特征词
1	移动图书馆 主题模型 信息传播 场景 时间序列 复杂网络 高校图书馆 自动分类 知识关联 网络结构 图书馆 共现分析 应急决策 科研合作 知识网络 知识交流 用户行为 阅读推广 影响因素 医学信息学
2	专利 图书馆 高校图书馆 竞争情报 数字图书馆 可视化 影响因素 研究热点 知识图谱 档案 LDA 企业 用户 阅读推广 微博 文献计量 技术领域 评价指标体系 情境 互联网
3	专利 图书馆 高校图书馆 竞争情报 数字图书馆 影响因素 研究热点 可视化 知识图谱 LDA 企业 用户 档案 阅读推广 微博 文献计量 技术领域 评价指标体系 情境 互联网
4	专利 竞争情报 研究热点 数字图书馆 可视化 知识图谱 LDA 企业 用户 影响因素 微博 文献计量 技术领域 评价指标体系 情境 互联网 学术影响力 主题词 术语 关联数据
5	高校图书馆 图书馆 阅读推广 影响因素 中国 CNKI 高校 美国 问卷调查 文献计量学 信息生态 评价指标 档案 科研人员 专利 知识服务 图书馆学 数据管理 知识库 国外
...	...
265	高校图书馆 档案 图书馆 阅读推广 影响因素 中国 CNKI 高校 美国 问卷调查 文献计量学 信息生态 评价指标 科研人员 互联网+ 知识服务 图书馆学 专利 数据管理 知识库

表9 三种算法的相似度计算结果

Table 9 Similarity calculation results of three algorithms

		1	2	3	4	5	...	17992
TF-IDF	1	0.620	0.564	0.501	0.393	0.418	...	0.692
	2	0.602	0.557	0.597	0.536	0.540	...	0.599
	3	0.615	0.587	0.604	0.561	0.583	...	0.602
	4	0.657	0.558	0.592	0.637	0.570	...	0.565
	5	0.545	0.494	0.516	0.391	0.460	...	0.624
	...	...	...	...	...	...	...	...
	265	0.537	0.503	0.580	0.589	0.526	...	0.473
TextRank	1	0.598	0.551	0.497	0.409	0.443	...	0.718
	2	0.669	0.621	0.620	0.629	0.630	...	0.604
	3	0.558	0.555	0.596	0.548	0.609	...	0.600
	4	0.599	0.602	0.579	0.579	0.580	...	0.627
	5	0.618	0.593	0.559	0.505	0.500	...	0.703
	...	...	...	...	...	...	...	...
	265	0.611	0.572	0.712	0.703	0.628	...	0.491
LDA	1	0.628	0.547	0.604	0.534	0.472	...	0.598
	2	0.676	0.605	0.635	0.548	0.533	...	0.686
	3	0.676	0.605	0.635	0.548	0.533	...	0.686
	4	0.571	0.508	0.525	0.441	0.458	...	0.660
	5	0.618	0.593	0.559	0.505	0.500	...	0.703
	...	...	...	...	...	...	...	...
	265	0.775	0.764	0.675	0.705	0.708	...	0.790

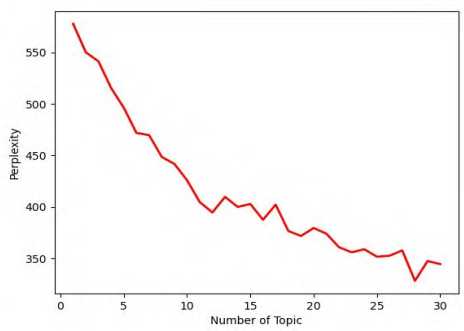

图3 困惑度曲线

Figure 3 Perplexity curve

从图3可以看出,当K=28时,困惑度值最小,模型效果最佳,因此,确定LDA模型的最佳主题数为28。确定了最佳主题数后,利用gensim中的LDA包进行主题提取,每个主题显示20个词,得到的主题—词汇分布以及文档—主题分布如表6、表7所示。

将文档—主题分布矩阵与主题—词汇分布矩阵对应相乘,得到的文档—词汇分布矩阵,这里的文档代指核心作者,即得到核心作者—词汇分布,再按照权值对词汇从高到低进行排序,选择权值靠前的Top20作为核心作者研究兴趣的特征词,部分结果如表8所示。

3.4 特征词向量化

在分别利用TF-IDF算法、TextRank算法和LDA模型筛

表10 三种算法的推荐结果

Table 10 Recommended results of three algorithms

	1	2	3	4	5	...	265
TF-IDF	2707	17801	14490	4114	17792	...	2974
	4119	12212	10982	4270	5212	...	2536
	6408	5435	16934	17772	3831	...	17914
	...	...	...	...	...	...	...
	5914	3127	12429	17682	15794	...	14831
TextRank	17170	8031	10998	6753	3595	....	7724
	15364	11686	14490	10197	6528	...	14745
	3421	14628	2965	1235	4245		14683
	...	...	....	...	...	...	...
	14329	17882	10327	4708	7395	...	15142
LDA	14617	3595	3595	15450	15450	..	15450
	10695	15460	15460	17529	17529	...	13045
	4570	3854	3854	13045	13045	...	17529
	...	...	...	...	...	...	...
	4180	4748	4748	15216	17495	...	15216

选出特征词后，利用Word2vec对特征词进行向量化。Word2vec模型包括CBOW和Skip-grow两种模型，CBOW适合规模较小的语料，而Skip-grow适合规模较大的语料。由于本文语料规模较小，因此选择CBOW进行模型训练。将表1中的17992条数据作为语料来训练词向量模型。模型训练时，设置维度size为100，窗口大小window为5，迭代次数iter为100次。训练好模型之后，分别将基于TF-IDF算法、TextRank算法和LDA模型得到的特征词的向量相加求和取平均，得到核心作者的向量表示。同理，学术论文的向量也是将语义实体相加求和取平均。

3.5 相似度计算以及推荐结果生成

3.5.1 相似度计算

在得到核心作者向量和学术论文向量之后，利用余弦相似度进行相似度计算，基于三种不同算法的相似度结果分别如表9所示。

3.5.2 推荐结果生成

将上文计算得到的相似度从高到低进行排序，选择相似度值靠前的Top20推荐给核心作者，基于三种不同算法的推荐结果如表10所示。

4 推荐模型评价

4.1 推荐效果评价

生成推荐结果之后，本文借助准确率、召回率和F值对推荐效果进行评价。准确率的计算方式为在推荐的20篇学术论文中，是否有核心作者自己撰写的学术论文，若有，则视为推荐成功，将推荐成功的数量与总数量的比值作为准确率。召回率的计算方式为在推荐的20篇学术论文中，包含的核心作者自己撰写的学术论文的数量与核心作者自己撰写的学术论文的总量的比值。F值的计算方式为准确率和召回率的调和平均值。三种算法的准确率、召回率和F值如表11所示。

从表11中可以看出，基于TF-IDF算法得到的推荐结果是效果最好的。以核心作者“邓胜利”为例，生成的推荐结果如表12所示。

表11 三种算法的准确率、召回率和F值

Table 11 Precision, recall and F value of three algorithms

算法	准确率	召回率	F值
TF-IDF	0.917	0.165	0.280
TextRank	0.713	0.069	0.088
LDA	0.106	0.007	0.009

在为核心作者“邓胜利”推荐的20篇学术论文中，13931、2533、2634和16417是核心作者自己发表的文献，其余的16篇文献是利用TF-IDF算法生成的推荐结果。从表12中可以看出，推荐效果是较为理想的，查阅核心作者“邓胜利”的相关文献可知，该作者主要的研究兴趣集中在“网络用户信息行为研究”“信息服务”“医疗健康信息”等方面。结合表12可知，为核心作者推荐的学术论文是与该核心作者研究兴趣密切相关内容的，无论是从研究对象还是理论技术上都能为该核心作者提供一定程度上的科研帮助，并且通过这些学术论文，作者可以进一步追踪到这些学术论文的作者，可以知晓本学科领域哪些作者与自己有着相同研究兴趣，为潜在的科研合作提供可能性。

4.2 模型有用性评价

在保证推荐效果较好的前提下，对推荐模型的有用性进行评价。在推荐的20篇学术论文中，如果大部分都是作者自己撰写过的论文，那么该推荐模型的有用性就极低。因此

表12 推荐结果举例

Table 12 Examples of recommended results

编号	相似度	学术论文
14105	0.934	社会化问答社区不同用户行为影响因素的实证研究
5034	0.932	微信学术检索用户行为分析与实证研究
13931	0.920	社会化问答平台用户体验影响因素实证分析——以知乎为例
4819	0.915	年龄梯度视角下网络用户健康信息甄别能力研究
8659	0.915	逃离还是回归? ——用户社交网络间歇性中辍行为实证研究的影响因素综述
2533	0.914	社会化问答社区用户信息需求对信息搜寻的影响研究——基于问答社区卷入度的中介作用分析
2634	0.914	国外用户音乐信息行为研究述评
16417	0.912	社会化问答社区用户信息行为的转化研究——从信息采纳到持续性信息搜寻的理论模型构建
2178	0.910	社交网络用户信息贡献行为影响因素分析
9253	0.909	广州中大布匹商圈外来务工人员日常生活信息行为实证研究
12716	0.908	社交网络中个人信息安全行为影响因素的实证研究
12518	0.906	虚拟社区用户知识付费意愿实证研究
13584	0.905	网络信息偶遇影响因素个性特征的调查实验研究
6979	0.904	知识众包社区中用户参与意愿的实证研究:基于虚拟社区归属感的视角
4366	0.903	付费知识问答社区中提问者的答主选择行为研究
12536	0.902	使用与满足视角下社交网络用户行为研究综述:基于国外54篇实证研究文献的内容分析
10200	0.901	基于社会化媒体的适应性信息分享影响因素研究
1949	0.900	移动阅读服务用户行为影响因素元分析模型构建
10133	0.900	微信用户学术信息交流行为影响因素研究
12883	0.899	社交媒体使用动机与功能使用的关系研究:以微信为例

有用性的计算方式为推荐的20篇学术论文中除去核心作者自己撰写的学术论文的比例,该值应该介于0-1之间,且该值越高,模型有用性越高。将每位核心作者计算的有用性指标求和取平均得到推荐模型的有用性值,最终计算得到有用性值为0.816,表明本文提出的推荐模型是能够为核心作者推荐满足其研究兴趣的学术论文。

5 结 语

随着大数据和互联网时代的发展进步,学术数据变得越来越丰富和繁杂,本文主要以较为典型的学术数据—学术论文及其作者作为研究对象,目的是实现为作者推荐与之研究兴趣相关的学术论文。为更准确以及全面地实现推荐,未直接采用学术论文中给定的关键词,而是更细粒度地,在实验前期将语义实体细分为研究对象、研究主题和理论技术类实体并利用深度学习算法进行实体抽取的基础上,将抽取出的细粒度语义实体作为学术论文和核心作者的内容表征,再分别借助三种不同的特征词筛选算法进行学术论文和核心作者的特征词提取,再用Word2vec生成词向量并计算余弦相似度,选择相似度值靠前的Top20生成推荐,同时对推荐结果利用准确率、召回率和F值进行比较评价,结果表明,基于TF-IDF提取的特征词产生的推荐效果最佳,之后利用有用性指标对模型的有用性进行了评价,以及对推荐结果进行了实例展示,可以看出,本文提出的推荐模型能够为科研用户推荐与其研究兴趣相关的学术论文,减少科研用户时间和精力的浪费,且推荐的论文在研究对象、研究主题和理论技术三个方面为科研用户提供更为全面的参考。然而,本文的研究也存在一定的局限性,侧重学术论文内容特征,忽略了学术论文包含的网络关系,以及没有考虑到不同关键词类型对于内容揭示的重要性是不同的,今后将这些因素考虑在内,得到更为精准的结果。

参考文献

1 吴磊,岳峰,王含茹,王刚.一种融合科研人员标签的学术论文推荐方法[J].计算机科学,2020,47(2):51-57.

2 陈长华,李小涛,邹小筑,叶志锋.融合Word2vec与时间因素的馆藏学术论文推荐算法[J].图书馆论坛,2019,39(5):110-117.

3 李冉,林泓.基于频繁主题集偏好的学术论文推荐算法[J].计算机应用研究,2019,36(9):2675-2678.

4 白金源.基于t-SNE和模糊聚类的科技论文推荐方法研究[D].保定:河北大学,2018.

5 熊回香,孟璇,叶佳鑫.基于关键词语义类型和文献老化的学术论文推荐[J].现代情报,2021,41(1):13-23.

6 李响,谭静.融合相关性与多样性的学术论文推荐方法研究[J].情报理论与实践,2017,40(6):99-103.

7 Bansal T, Belanger D, Mccallum A.Ask the GRU:Multi-task Learning for Deep Text Recommendations[J]. Recsys16,2017,arXiv:1609.02116.

8 许侃,刘瑞鑫,林鸿飞,刘海峰,冯娇娇,李家平,林原,徐博.基

于异质网络嵌入的学术论文推荐方法[J].山东大学学报(理学版),2020,55(11):35-45.
9 孙婧.基于引文网络图模型的论文推荐系统研究与应用[D].昆明:云南师范大学,2020.
10 丁芳媛.基于新颖性和影响力的论文推荐方法研究[D].广州:华南理工大学,2020.
11 Haruna K, Ismail M A, Bichi A B, et al.A Citation-Based Recommender System for Scholarly Paper Recommendation [C]//Computational Science and Its Applications - ICCSA, 2018:514-525.
12 Ma X, Wang R.Personalized Scientific Paper Recommendation Based on Heterogeneous Graph Representation[J]. IEEE Access,2019,7:79887-79894.
13 戴大文.基于用户偏好和引文关系的科技论文推荐算法研究[D].重庆:重庆大学,2018.
14 潘峰,怀丽波,崔荣一.基于分布式图计算的学术论文推荐算法[J].计算机应用研究,2019,36(6):1629-1632,1642.
15 孟伟龙.基于图模型的论文推荐系统设计与实现[D].杨凌:西北农林科技大学,2019.
16 David M Blei, Andrew Y Ng, Michael I Jordan. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003(3):993-1022.
17 Mikolov T,Sutskever I,Chen K,etal.Distributed Representations of Words and Phrases and Their Compositionality [DB/OL] .arXiv Preprint,arXiv:1310.4546.
18 李晓敏.面向图情档期刊的学术知识图谱构建及应用研究[D].武汉:华中师范大学,2021.

（责任编辑:毛秀梅）

Recommendation of Academic Papers Based on Fine-Gained Semantic Entities

LI Xiao-min[1,2], *WANG Hao*[1,2], *LI Yue-yan*[1,2]

*(*1.*School of Information Management, Nanjing University, Nanjing* 210023*, China;*
2.*Jiangsu Key Laboratory of Data Engineering and Knowledge Service (Nanjing University), Nanjing* 210093*, China)*

Abstract:【Purpose/significance】In order to help scientific research users find academic papers related to their research interests quickly and accurately, an academic paper recommendation model based on fine-grained semantic entities is constructed.【Method/process】The research topics, research objects, and theoretical and technical semantic entities identified in the early stage of the experiment are used as the content features of academic papers and core authors, and academic papers and core authors are obtained by using TF-IDF algorithm, TextRank algorithm and LDA model, respectively Use Word2vec to vectorize the feature words, and then calculate the cosine similarity between the core author and the academic paper, and recommend the top 20 cosine similarity values to the author.【Result/conclusion】Using accuracy, recall and F-values to compare and evaluate the recommendation results generated by the feature words based on the three algorithms, the results show that the feature words generated based on the TF-IDF algorithm have the best recommendation effect.The results of the recommendation are shown with examples, and it can be seen that the recommendation model proposed in this paper can more comprehensively recommend academic papers with similar research interests to scientific research users and improve the efficiency of scientific research.【Innovation/limitation】It mainly starts with the content characteristics of academic papers, and does not involve the network relationships contained in academic papers.

Keywords:feature words; core author; academic papers; personalized recommendation; similarity