文章编号:1000-2367(2023)04-0034-09 DOI:10.16366/j.cnki.1000-2367.2023.04.005

基于 Doc2Vec 和 LDA 模型融合文献质量的学术论文推荐研究

王大阜,邓志文,贾志勇,王静

(中国矿业大学 图书馆,江苏 徐州 221116)

摘 要:为解决海量的电子资源给读者带来“信息过载”的困扰,采用基于内容的推荐算法为读者推荐内容适配、质量优良的学术论文.考虑论文文本的上下文语义、词序及全局主题信息,首先采用 Doc2Vec 和 LDA(Latent Dirichlet Allocation)混合语义模型训练候选论文集摘要语料库,学习得到每篇论文的文本向量,其次利用 K-Means 算法对候选论文集进行聚类,然后探寻目标论文所属簇的类群成员作为待推荐论文,最后融合文献质量权重进行相似度计算并排序,从而得到 TOP-N 近邻推荐结果.以 CNKI 图书情报类期刊论文作为语料库,通过实证分析,采用的混合模型与传统的 TF-IDF(Term Frequency-Inverse Document Frequency)、Word2Vec、LDA 3 种模型相比,推荐结果的精确率较高、排序差异度低,达到良好的推荐效果.

关键词:学术论文;混合语义模型;文献质量;推荐

中图分类号:TP391.3;G250 **文献标志码**:A

图书馆是文化、资源的聚集地与传播中心,为了满足读者泛在化科研、学习的需求,更好地推进高校“双一流”建设,促进学科建设发展,电子资源已经成为图书馆主要引进的文献结构类型.科研大数据时代,面对海量的学术资源,学者难以就其感兴趣的学科领域,进行相关文献的择优选取,从而对学者造成“信息过载”的困扰[1].学者通常借助 OPAC 系统、中外文数据库搜索引擎,从中检索、选取相关的论文文献.根据美国科学基金会统计,学术人员在开展学术活动的过程中,花费在资料收集上的时间占全部科研时间的 51%,科研效率低下[2].因此,用户更倾向于智慧化的个性化服务,希望系统自动、高效地向读者推荐、呈现感兴趣的优质资源.以满足读者的需求为导向,图书馆亟需利用大数据、推荐算法、机器学习等技术,从用户阅读行为信息、学术成果等信息中,挖掘读者阅读兴趣偏好,为其提供与研究兴趣较匹配、文献质量优越的资源推荐服务,帮助读者从耗时耗力的检索、挑选论文的事务中解脱出来,同时精准的推荐效果亦能够激发读者的学术活跃度和科研创作潜能.

近年来,学界围绕纸本资源、电子资源的个性化推荐开展了大量研究,采用的技术方法主要分为 3 种:(1)基于协同过滤推荐:刘岩[3]基于传统的协同过滤(Collaborative Filtering,CF)算法实现纸本文献的推荐.协同过滤算法依赖读者对物品的评分数据,大多数读者在使用 OPAC 系统检索、借阅文献时不关注评分事项,导致用户-项目评分矩阵数据稀疏,推荐精度不理想,为此,有学者研究利用借阅次数、借阅时长计算用户-项目的隐性评分[4].(2)基于内容推荐:论文、专利等学术资源属于典型的文本数据,学者们主要使用信息检索的理论和技术,围绕“语义相关性”展开,结合自然语言处理(NLP)技术对文本进行特征提取和挖掘分析,同时能够缓解物品冷启动问题[1].张戈一等[5]采用 TF-IDF 算法分别对地质文献进行特征提取、相似度计算及推荐.阮光册[6]构造读者借阅行为共现矩阵,利用 Word2Vec 的潜在语义分析特性提供多样性的推荐结果.耿立校等[7]针对向量空间模型存在的文本向量维度灾难问题,采用 TF-IDF 算法提取关键词,再结合 Word2Vec 模型实现文献推荐.熊回香等[8]从关键词语义类型和文献老化两个维度出发,为用户推荐符合其研究方向且时间价值高的学术论文.陈长华等[9]结合 Word2Vec 与时间因素进行论文推荐.杜永萍等[10]使用

收稿日期:2022-08-11;**修回日期**:2023-03-21.

基金项目:江苏省高校哲学社会科学研究项目(2022SJYB1129);国家社科基金(22BTQ023).

作者简介(通信作者):王大阜(1981—),男,江苏盐城人,中国矿业大学图书馆馆员,研究方向为推荐系统、知识图谱,E-mail:wdf@cumt.edu.cn.

LDA 模型对候选文献和用户发表的文献进行建模，根据两者相似度值进行推荐.张卫卫等[11]融合 LDA 和 Doc2Vec 模型，利用语料库的全局语义信息和上下文语义信息，分别进行学术摘要聚类、挖掘学者研究兴趣标签，对本文的研究提供了一定的参考借鉴.WANG 等[12]基于 TF-IDF 算法得出学术论文的关键词分布为用户推荐学术论文.KANAKIA 等[13]将 Word2Vec 模型和共被引方法相结合对微软学术论文进行推荐.(3)基于社交网络推荐：该方法是基于学术社交网络进行社区划分，并对社区内用户进行学术论文推荐[14-15].此外，有学者根据用户基本信息、阅读习惯、阅读行为及情景数据进行用户画像，将用户标签化，并向其提供个性化推荐服务[2,16].随着深度学习技术迅猛发展，CNN、RNN 模型应用于论文推荐系统，实现深层次地挖掘文本隐式特征[17-18].

综上所述，学术论文推荐方法主要是采用 TF-IDF、Word2Vec、LDA 不同模型对论文摘要语料库进行训练，学习得到论文的文本特征向量，再将目标论文与候选论文集进行相似度计算，选择相似度较高的候选论文集作为推荐结果.以上 3 种模型的推荐结果的精度和效果不理想，原因是存在以下局限性：(1)特征提取信息不完善，TF-IDF 仅考虑词权重，存在“词汇鸿沟”现象，认为词语间是相互独立的；Word2Vec 获取到上下文语义信息，但缺失词序、全局主题信息；LDA 从全局语境挖掘隐性的主题信息，但忽略了局部上下文语义关系.(2)缺乏文献质量因素的度量，仅从文本内容角度作相似度对比，导致向读者推荐过时的、学术质量相对较低的论文.本文研究内容旨在设计学术论文推荐模型，向用户推荐内容相似度高的学术论文，并在此基础之上，引入文献质量权重对相似度进行加权修正，从而使用户获取到高质量的学术论文.

1　相关模型

1.1　词向量模型

文本向量化是从文本中提取特征，将文本表示成可量化、可运算的数字形式.词嵌入(Word Embedding)技术诞生之前，文本向量化通常采用词袋模型(Bags of Words，BoW)中的独热编码(One-Hot)、词频(Term Frequency，TF)、词频-逆文档频率(TF-IDF)3 种表示方法.One-Hot 方法根据词在字典中的索引位置将文本转化为 0 或 1 二值向量，该方法会造成严重的维度灾难和数据稀疏问题.TF 和 TF-IDF 是基于向量空间(VSM)模型，将一篇文本投射成高维空间中的一个点，该点的坐标对应文本的多个特征词向量.词袋模型的缺点在于：假定词与词间相互独立，不考虑词间的语义关系，而且对于大规模语料库而言，仍然存在维度灾难问题.

2013 年 Google 公司 Tomas Mikolov 团队发布了 Word2Vec 词嵌入模型，Word2Vec 认为相似语境的词语语义相近，通过三层神经网络模型训练，将多维词向量映射成稠密的低维向量，从而实现了词的分布式表示[9].2014 年 Tomas Mikolov 提出改进模型 Doc2Vec，也称为段落向量(Paragraph Vector)，增加了词序语义的分析，用于创建文档向量，文档可以是句子、段落或文章.Doc2Vec 模型分为分布式内存模型(PV-DM)和分布式词袋模型(PV-DBOW)两种，架构如图 1 所示，PV-DM 模型是在输入层增加了段落 ID 作为样本用于预测目标词概率，段落 ID 类似一个特殊的上下文单词，存储着段落信息，段落向量被该文档所有上下文窗口共享[11].PV-DBOW 模型将段落 ID 作为输入，从文档中预测随机采样的词概率.Doc2Vec 模型能够同时训练学习到词向量和文档向量，适用于论文文本向量化高效处理的需求，而且它考虑了词序信息，对词预测也更为准确、灵活.

1.2　LDA 主题模型

潜在狄利克雷分布(LDA)主题模型是通过语义分析技术，对上下文理解后，挖掘出隐含的抽象主题[11].LDA 模型的基本思想是：一篇文档隐含了多个主题，一个主题由多个词语构成，通过迭代模拟文档生成过程，识别文档和文档集中潜在的主题信息.鉴于此，LDA 模型由文档、主题、词组成的三层贝叶斯概率分布生成，α 和 β 是两个 Dirichlet 先验超参数，θ 表示文档到主题之间的多项分布，ψ 表示主题和词之间的多项分布.α、β 与其他变量之间的服从分布关系为：$\theta \sim \text{Dirichlet}(\alpha)$，$z \sim \text{Multinomial}(\theta)$，$\psi \sim \text{Dirichlet}(\beta)$，$w \sim \text{Multinomial}(\psi)$.

对于语料库中每篇文档 d 生成的过程如图 2 所示：(1)为文档 d 选择一个由 T 个主题混合的概率分布 θ，从超参数 α 吉布斯采样生成；(2)对于文档 d 每个单词从 Multinomial 分布中取样生成主题 z；(3)从超参

数 β 吉布斯采样生成 ψ,以 ψ 为参数的 Multinomial 分布中采样生成词 w;(4)上述 3 个步骤重复 N 次,产生文档 d.

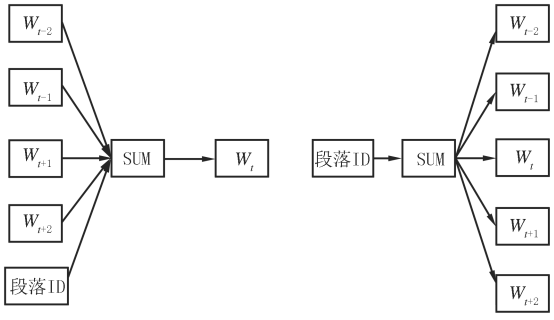

图1 Doc2Vec PV-DM模型与PV-DBOW模型

Fig. 1 Doc2Vec PV-DM model and PV-DBOW model

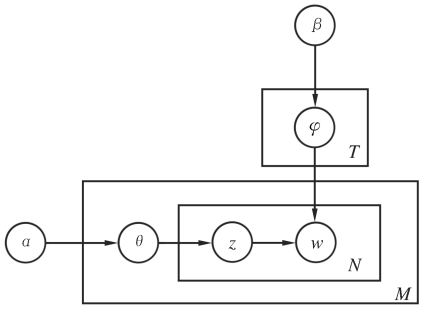

图2 LDA模型结构

Fig. 2 LDA model structure

LDA 模型的作者 Blei 采用困惑度(Perplexity)评估 LDA 主题模型好坏,确定最优主题数.困惑度小,说明模型具有更好的泛化能力[19].困惑度(P_e)的计算公式如下:

$$P_e = \exp\left\{-\frac{\sum_{d=1}^{M} \ln p(w_d)}{\sum_{d=1}^{M} N_d}\right\}, \tag{1}$$

式中,M 表示文档的数量,w_d 表示文档 d 中的单词,N_d 表示文档 d 中的单词数量,$P(w_d)$表示文档中词 w_d 产生的概率.

2 文献质量评估

文献质量的衡量取决于文献老化率、期刊影响因子及作者权威度 3 个因素,综合 3 个因素对文献质量进行定量评估.

2.1 文献老化

随着科学技术的不断演进发展,文献随之发生新陈代谢、老化淘汰.对于读者而言,较新的论文能够捕捉某学科的研究热点及其理论技术发展前沿,论文质量、研究价值相对较高.文献老化是科学计量学与文献计量学的重要课题,衡量文献老化速度和程度的主要度量指标有半衰期和普赖斯指数[20].半衰期是指在利用的全部文献中较新的一半是在多长时间内发表的.普赖斯指数(P_r)是指在某一个知识领域内,年限不超过 5 a 的被引文献数量与引文文献总量的比例,计算公式为:

$$P_r = \frac{\text{出版年限不超过 5 a 的被引文献数量}}{\text{被引文献总量}} \times 100\%. \tag{2}$$

文献老化经典数学模型利用引文共时数据分析法,反映文献引用频率与时间(以 10 a 为单位)之间的函数关系,揭示某些特定学科领域文献的老化规律.参考文献[21],定义 Age(简记为 A_g)表示文献老化率,用于进一步区分每篇论文的老化程度,计算公式为:

$$A_g = \begin{cases} 1, & T < t, \\ \mathrm{e}^{T-t}, & T \geq t, \end{cases} \tag{3}$$

式中,T 为半衰期,根据文献[21],图书情报类文献的半衰期 T 值为 6 a.t 为文献自发表之日起至推荐时间所间隔的时长,计算方式为以 d 为单位再换算为以 a 为单位.

2.2 期刊影响因子

论文的质量与期刊影响因子密切相关,期刊影响因子(Impact Factor,IF)是指期刊中论文的平均应用率,等于期刊近两年被引用量与发文量之比,IF 直观反映期刊整体的论文质量,利用 IF 表示同一期刊中每篇候选论文的通用质量.根据 CNKI 期刊数据统计,2021 年图书情报类期刊的 IF 值区间范围为[0.811,7.343],为抑制 IF 值过大对整体文献质量的影响,利用离差标准化(Min-Max)方法对特征做归一化处理,IF

（简记为 I）计算公式为：

$$I=\frac{I-I_{\min}}{I_{\max}-I_{\min}}. \tag{4}$$

2.3　文献影响力

在同一研究领域，某篇文献被引次数较高，表明该论文更受学者们的青睐和认可，其学术影响力较高，其中可能包括历久不衰的经典文献.定义 I' 表示文献影响力，取 100 作为被引次数的阈值，超过 100 可以当作高影响力文献，计算公式为：

$$I'=\begin{cases}1, & \text{被引次数}\geq 100,\\ \dfrac{\text{被引次数}}{100}, & \text{被引次数}<100.\end{cases} \tag{5}$$

定义 Q_a 表示候选论文的文献质量权重，该指标综合期刊影响因子、文献老化率以及文献被引次数（I'）3 个因素，并取其平均值作为 Q_a 指标值，计算公式为：

$$Q_a=\frac{1}{3}(I+A_g+I'). \tag{6}$$

3　推荐模型框架

推荐模型架构如图 3 所示，本文以 CNKI 期刊论文的摘要文本作为语料库，融合两种 Doc2Vec、LDA 模型进行建模，用于训练语料库，利用优势互补来挖掘文本局部上下文语义、词序信息以及隐藏的全局主题信息，扩充了所提取的论文文本的特征丰度和细粒度，为后续的 NLP 任务处理提升识别及预测能力.大量的文本相似度计算会增加计算复杂度，造成算法运行缓慢，为此通过聚类进行了优化处理，将候选论文集进行聚类，划分出多个类群，然后在类群范围内寻找文献质量加权相似度较高的候选论文集.

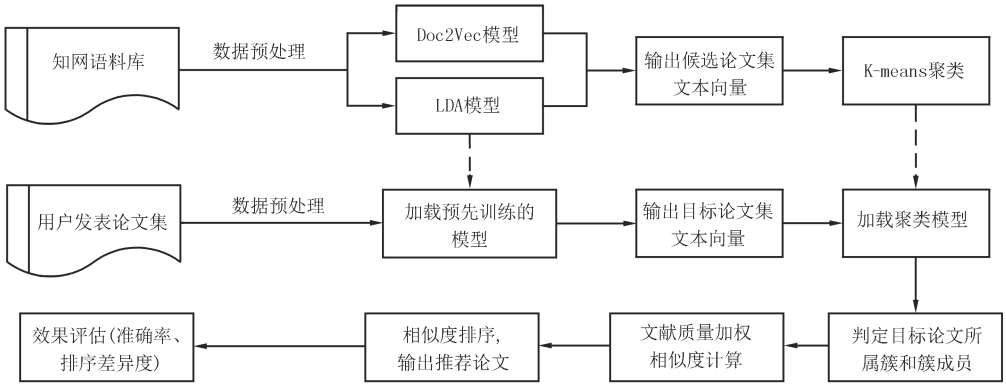

图3　论文推荐模型架构

Fig.3　Paper recommendation model architecture

假定语料库是由一系列文档（论文摘要文本）组成的集合 $D=\{d_1,d_2,\cdots,d_n\}$，文档 d 的词集合 $W=\{w_1,w_2,\cdots,w_n\}$，LDA 模型训练文档 D 后得到多个隐含主题集合 $T=\{t_1,t_2,\cdots,t_n\}$. 推荐模型的推荐结果处理流程如下.

步骤 1　对语料库进行分词、去除停用词等数据预处理，利用 Doc2Vec 模型训练语料库，得到所有词向量和文本向量，假定某篇论文文档 d 的向量，记作 $v(d)$，$v(d)=[w_{d1}\quad w_{d2}\quad w_{d3}\quad\cdots\quad w_{dm}]$，$w_{di}$ 表示文档 d 的第 i 个特征值.

步骤 2　利用 LDA 模型训练语料库，得到每篇论文文档的主题概率分布，即文档的主题向量，记作 $v(d)^t$，$v(d)^t=[t_{d1}\quad t_{d2}\quad t_{d3}\quad\cdots\quad t_{dm}]$，$t_{di}$ 表示文档 d 在主题 t_i 上的概率分布.

步骤 3　利用 Doc2Vec 和 LDA 模型，融合语料库的上下文语义信息及全局语义信息，进一步扩充提取文档的特征.做法是将步骤 1 得到的文档向量 $v(d)$ 和步骤 2 得到文档的主题分布向量 $v(d)^t$ 横向拼接，从而

得到改进的文档向量 $v(d)'$,$v(d)'=[v(d),v(d)^t]$.

$$v(d)'=[w_{d1}\quad w_{d2}\quad w_{d3}\quad \cdots\quad w_{dm}\quad t_{d1}\quad t_{d2}\quad t_{d3}\quad \cdots\quad t_{dm}].$$

步骤 4　采用 K-Means 经典聚类算法对所有文档进行聚类并保存模型,两篇文档的距离采用余弦相似度进行度量,相似度越高则距离越接近.通过多次迭代计算簇中心,直至簇中心收敛,不再改变位置,最终确定多个聚类簇.余弦相似度计算公式如下:

$$\mathrm{sim}(d,d')=\cos(d,d')=\frac{d\times d'}{\|d\|\times\|d\|}. \tag{7}$$

步骤 5　将用户发表的论文集作为目标论文集,经过分词、去除停用词等数据预处理后,通过步骤 1、2 训练保存的模型得到文档向量.接着通过步骤 4 保存聚类模型,计算出每篇目标论文距离最近的簇中心,进而判定目标论文所属簇及簇内的类群成员(即待推荐论文集).

步骤 6　修正余弦相似度公式(式 8),为其赋予文献质量权重,接着对相似度进行排序,排序后的 TOP-N 篇候选论文作为输出推荐结果.聚类示意图如图 4 所示,文献质量(Q_a)加权相似度计算公式如下:

$$\mathrm{sim}'(d,d')=Q_a\times\mathrm{sim}(d,d'). \tag{8}$$

4　实证分析

4.1　数据准备及预处理

通过编写 Python 程序,从 CNKI 爬取图书情报类期刊论文作为候选论文集,同时将论文摘要作为语料库.期刊论文的检索条件为文献分类:图书情报与数字图书馆,时间范围为 2014—2021 年(8 年),来源类别为北大中文核心期刊和 CSSCI 来源期刊,清除选题指南、名人专访等无效文献,最终采集总计 42 072 篇论文.目标论文集选取中国矿业大学 10 位馆员近 5 年发表的论文.数据预处理环节采用结巴分词工具进行分词,停用词采用哈工大停用词词典追加部分自定义停用词,比如论文中经常出现的“目的”、“意义”、“过程”等无关词.图书情报学领域具有很多专业术语,为了使得分词更加精准,通过 BICOMB 工具提取文献关键词,进而构建包含 1 724 个词的自定义词典.

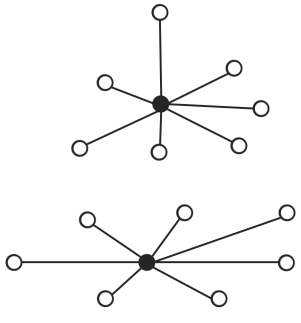

图4　目标论文(黑圆圈)与簇中心邻近点(白圆圈)示意图

Tab.4　Schematic diagram of target paper(black circle) and cluster center adjacent points(white circle)

4.2　混合语义模型训练

采用 Python 版本的 Gensim 软件包训练 Doc2Vec、LDA 模型,Doc2Vec 模型的参数设置为:window(窗口大小)设为 5,min_count(最小词频阈值)设为 5,Dm(模型类别)设为 1,即 PV-DM 模型,Size(段落向量维度)设为 100,epochs(迭代次数)设为 200.LDA 模型的参数设置为:超参数 α 设为 0.05,超参数 β 设为 0.01,iterations(迭代次数)设为 200,dictionary(字典)过滤词频小于 5 的词.

前文提到,LDA 模型采用困惑度指标评估最优主题数,本文通过绘制困惑度与主题数间的曲线,结果表明:当主题数为 20 时,困惑度出现拐点,下降趋势逐渐平缓,另外从 PyLDAvis 库生成的主题可视化结果(图 5)来看,各个圆圈代表一个主题,各个主题之间重合性不高,因此最佳主题数取值为 20.根据 20 个主题对应的高频次出现的词语分布,提取主题所代表的含义标识,其中前 8 个热门主题分别为:“图书馆服务模式创新”“图书馆建设发展”“用户行为分析”“知识组织与知识服务”“文献资源建设与保护”“科技文献发展”“阅读推广”“高校学科服务”等.图 5 中圆圈的大小代表每个主题相关的文献数量,圆圈较大的主题即为热门主题.由图 5 可见,图书情报类论文研究的主题分布整体较为分散,少数主题关系紧密,如主题 1 与主题 2 和主题 5 有部分重合,三者均与图书馆建设发展关联密切.

4.3　推荐实例

时间是一种重要的上下文信息,用户的研究兴趣会随着时间上下文的推移而发生迁移,本文假定读者的

研究兴趣 5 年内不会衰减，以中国矿业大学图书馆某馆员为例，表 1 是该馆员近 5 年的发表论文汇总（共 5 篇），采用 K-Means 算法对语料库聚类，每篇论文所在簇的成员数侧面反映出该研究主题的发文量及研究热度.

以该馆员序号 1 的发表论文《学科分析中科研合作网络分析方法研究》为例，模型的部分推荐论文如表 2 所示，序号 1～6 是按照初始的余弦相似度进行排序.可见，推荐论文的主题与发表论文研究主题均与科研合作网络相关，契合度十分高.在引入文献老化率、期刊影响因子及文献影响力3 个因素后，对相似度进行加权修正计算后，重新排序的次序为序号 2、序号 5、序号 6、序号 3、序号 4、序号 1.序号 1 论文初始排名第 1，但是因为其老化率较低，使其最终的推荐结果产生改变，排名转为第 6.序号 6 论文因为老化率和被引次数较高，最终排名转为第 3.由此可见，加权后的论文推荐结果不仅保证了内容的相关性，而且在论文质量上得到了很好的保证.

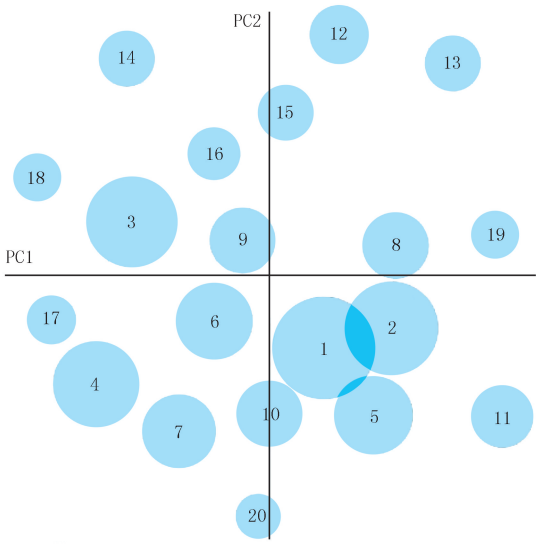

图5　主题可视化

Fig.5　Theme visualization

表 1　某馆员近 5 年发表论文列表

Tab. 1　List of papers published by a librarian in the past 5 years

序号	题名	研究主题	期刊	发表日期	所在簇	簇成员数
1	学科分析中科研合作网络分析方法研究	科研合作网络	农业图书情报学报	2019-12-24	2	641
2	基于用户画像的图书馆主动信息服务	基于用户画像的推荐服务	科技视界	2019-11-15	17	3 366
3	基于多数据源的机构知识可视化研究与应用	机构知识库可视化	现代情报	2019-02-01	18	940
4	开放模式下电子书在线阅读平台设计	资源建设	数字图书馆论坛	2018-02-25	10	831
5	基于位置感知的图书馆主动信息服务系统设计	基于位置的推荐服务	数据分析与知识发现	2016-02-25	1	1 708

表 2　某馆员部分推荐论文列表

Tab. 2　List of some recommended papers by a librarian

序号	题名	期刊	发表日期	影响因子	老化率	被引次数	相似度	修正后相似度
1	基于合著网络和被引网络的科研合作网络分析	情报理论与实践	2014-10-30	3.419	0.308	29	0.945	0.314
2	科研合作网络多数据源加权模型研究	情报理论与实践	2016-08-31	3.419	1	4	0.924	0.443
3	科研合作关系网络数据源分布研究	图书馆杂志	2017-05-15	1.683	1	3	0.915	0.354
4	基于社会网络分析的机构科研合作关系研究	图书情报知识	2014-03-10	4.038	0.162	38	0.913	0.315
5	基于学科内容的科研人员隐性合作关系研究	情报理论与实践	2017-07-11	3.419	1	6	0.902	0.421
6	基于 Pajek 的科研领域合作关系网络特征分析	图书馆	2016-07-11	2.296	1	10	0.896	0.396

4.4　推荐效果评估

本文采用精确率评估本文推荐模型的推荐精确度，方法是采用本文模型与 TF-IDF、Word2Vec、LDA 3 种模型，分别向 10 位馆员推荐相似度较高的 TOP-N 论文，N 分别取值 10、15、20、25、30，并对推荐论文进

行满意度(满意或不满意)评价,精确率(Precision,P)的计算公式如下.

$$P=\frac{P_T}{P_T+P_N}, \tag{9}$$

式中,P_T 为用户满意的推荐论文数,P_N 为用户不满意的推荐论文数.

实验结果表明:随着 N 的增加,精确率逐渐提高,当 $N=20$ 时,论文推荐的精确率最高($P@20=0.729$),随后又发生明显降低.4 种模型推荐精确率如图 6 所示,由图 6 可见,本文模型与其他模型对比,精确率最高,其次为 Word2Vec、LDA 及 TF-IDF.分析其原因是:TF-IDF 模型仅考虑论文的关键词权重,而且文本向量数据稀疏,导致精确率最低.Word2Vec 模型考虑上下文语境信息,并且文本向量低维稠密,因此精确率高于 TF-IDF,但是文本向量是简单地通过词向量取均值表示,会丢失部分上下文信息.LDA 模型是基于论文的词分布提取论文的主题信息,缺乏上下文语义信息,与 Word2Vec 模型相比,精确率较低.本文模型综合论文的上下文语义、全局语义以及词序信息,更能准确表达论文的内容主旨,因此精确率最高.

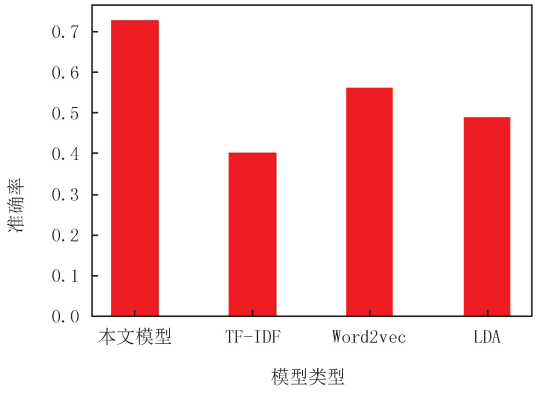

图6 4种模型推荐精确率计算结果对比

Fig.6 Comparison of calculation results of recommended precision of four models

为了验证本文文献质量加权计算方法的有效性,即推荐论文的排序效果,需计算推荐论文的输出排序与用户真实排序的差异,该差异值越小,则越符合用户期望的排序结果.定义 p 为排序差异值,计算公式为:

$$p=\frac{\sum_{i=1}^{20}|O_i-O_i'|}{20}, \tag{10}$$

其中,O_i 和 O_i' 分别为论文在模型推荐排序与用户真实排序的排名,对 20(最佳推荐数)篇推荐论文取平均排序差额,作为排序差异度.10 位馆员的排序差异度平均为 1.7,效果尚佳,较为贴近用户的真实排序.

5 结束语

本文采用单机对设计的推荐模型进行实证,语料库使用的图书情报类期刊论文文本,并取得良好的实验效果.同时,本文研究存在一定的局限性,下一步考虑从以下 3 个方面进行优化改进:

1)处理性能提升方面,在实际场景中,语料库会涵盖各个学科的期刊论文文本,超大规模的语料库在提升模型精度的同时,会造成训练收敛缓慢,并且数以万计的论文相似度计算存在性能瓶颈.采用 Spark 分布式计算框架作为推荐系统的计算平台,区别于 Hadoop 的 MapReduce 框架,Spark 将数据集缓存在内存中,避免计算过程中频繁的磁盘 I/O 操作,从而有效提升推荐系统的处理性能[22].

2)目标论文集数据来源方面,图书馆的机构知识库存储展示学者的学术科研成果,用户发表论文的标题、摘要、作者等元数据可以通过机构知识库定期采集获取,随着用户发表论文量的增多,推荐结果同步发生改变,保证了推荐系统的实时性.

3)论文和图书的资源整合方面,用户发表的论文及图书借阅信息都揭示了用户的阅读兴趣,可以将两者有机结合起来,更精准地提取用户兴趣特征,向用户推荐优质论文.反之亦然,可以为平常更关注论文的学术用户推荐优质书籍,扩充可供学习参考的文献范围.

参 考 文 献

[1] 蒲姗姗,何燕.个性化学术资源推荐研究:现状、特点及展望[J].图书馆学研究,2019(16):9-17.
PU S S,HE Y.Personalized academic resource recommendation:research status,features,and prospects[J].Research on Library Science,2019(16):9-17.

[2] 王仁武，张文慧.学术用户画像的行为与兴趣标签构建与应用[J].现代情报，2019，39(9)：54-63.
WANG R W，ZHANG W H.Behavior and interest labeling construction and application of academic user portraits[J].Journal of Modern Information，2019，39(9)：54-63.

[3] 刘岩.基于机器学习算法的图书馆书目协同推荐系统[J].现代电子技术，2020，43(14)：180-182.
LIU Y.Library bibliographic collaborative recommendation system based on machine learning algorithm[J].Modern Electronics Technique，2020，43(14)：180-182.

[4] 李澎林，洪之渊，李伟.基于兴趣度与类型因子的高校图书推荐算法[J].浙江工业大学学报，2019，47(4)：425-429.
LI P L，HONG Z Y，LI W.Book recommendation algorithm base on the interest and type factor for university[J].Journal of Zhejiang University of Technology，2019，47(4)：425-429.

[5] 张戈一，胡博然，常力恒，等.基于大数据分析挖掘的地质文献推荐方法研究[J].中国矿业，2017，26(9)：92-97.
ZHANG G Y，HU B R，CHANG L H，et al.Basics big date analysis analytic excavation geology reference recommendation method research[J].China Mining Magazine，2017，26(9)：92-97.

[6] 阮光册，谢凡，涂世文.基于 Word2Vec 的图书馆推荐系统多样性问题应用研究[J].图书馆杂志，2020，39(3)：124-132.
RUAN G C，XIE F，TU S W.Application research based on Word2Vec diversity in library recommender system[J].Library Journal，2020，39(3)：124-132.

[7] 耿立校，晋高杰，李亚函，等.基于改进内容过滤算法的高校图书馆文献资源个性化推荐研究[J].图书情报工作，2018，62(21)：112-117.
GENG L X，JIN G J，LI Y H，et al.Research on personalized recommendation of university library literature resources based on improved content-based filtering algorithm[J].Library and Information Service，2018，62(21)：112-117.

[8] 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.
XIONG H X，MENG X，YE J X.Recommendation of academic papers based on keyword semantic type and literature obsolescence[J].Journal of Modern Information，2021，41(1)：13-23.

[9] 陈长华，李小涛，邹小筑，等.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.
CHEN C H，LI X T，ZOU X Z，et al.A new Word2Vec algorithm of academic paper recommendation[J].Library Tribune，2019，39(5)：110-117.

[10] 杜永萍，杜晓燕，姚长青.基于主题效能的学术文献推荐算法[J].北京工业大学学报，2015，41(2)：215-222.
DU Y P，DU X Y，YAO C Q.Recommendation algorithm based on topic utility for academic papers[J].Journal of Beijing University of Technology，2015，41(2)：215-222.[11]

[11] 张卫卫，胡亚琦，翟广宇，等.基于 LDA 模型和 Doc2vec 的学术摘要聚类方法[J].计算机工程与应用，2020，56(6)：180-185.
ZHANG W W，HU Y Q，ZHAI G Y，et al.Academic abstract clustering method based on LDA model and Doc2vec[J].Computer Engineering and Applications，2020，56(6)：180-185.

[12] WANG Z Y，LIU Y，YANG J J，et al.A personalization-oriented academic literature recommendation method[J].Data Science Journal，2015，14：17.

[13] KANAKIA A，SHEN Z H，EIDE D，et al.A scalable hybrid research paper recommender system for microsoft academic[C]//WWW'19：The World Wide Web Conference.New York：ACM，2019：2893-2899.

[14] 黄泳航，汤庸，李春英，等.基于社区划分的学术论文推荐模型[J].计算机应用，2016，36(5)：1279-1283.
HUANG Y H，TANG Y，LI C Y，et al.Academic paper recommendation model based on community partition[J].Journal of Computer Applications，2016，36(5)：1279-1283.

[15] 贺超波，沈玉利，余建辉，等.基于学术社区的科技论文推荐方法[J].华南师范大学学报(自然科学版)，2012，44(3)：55-58.
HE C B，SHEN Y L，YU J H，et al.Method for scientific paper recommendation based on academic community[J].Journal of South China Normal University(Natural Science Edition)，2012，44(3)：55-58.

[16] 王庆，赵发珍.基于"用户画像"的图书馆资源推荐模式设计与分析[J].现代情报，2018，38(3)：105-109.
WANG Q，ZHAO F Z.Design and analysis of library resource recommendation model based on user profile[J].Journal of Modern Information，2018，38(3)：105-109.

[17] 黄立威，江碧涛，吕守业，等.基于深度学习的推荐系统研究综述[J].计算机学报，2018，41(7)：1619-1647.
HUANG L W，JIANG B T，LV S Y，et al.Survey on deep learning based recommender systems[J].Chinese Journal of Computers，2018，41(7)：1619-1647.

[18] 王妍，唐杰.基于深度学习的论文个性化推荐算法[J].中文信息学报，2018，32(4)：114-119.
WANG Y，TANG J.Deep learning-based personalized paper recommender system[J].Journal of Chinese Information Processing，2018，32(4)：114-119.

[19] 赵凯，王鸿源.LDA 最优主题数选取方法研究：以 CNKI 文献为例[J].统计与决策，2020，36(16)：175-179.
ZHAO K，WANG H Y.Research on the selection method of LDA's optimal topic number：taking CNKI literature as an example[J].Statistics & Decision，2020，36(16)：175-179.

[20] 邱均平.文献计量学[M].2 版.北京:科学出版社,2019:67-70.
[21] 刘伙玉.基于 CNKI 的图书、情报学与档案学学科文献半衰期分析[J].图书与情报,2015(1):106-111.
LIU H Y.Analysis of half-life of literatures on libraries,information and archives based on CNKI[J].Library & Information,2015(1):106-111.
[22] 何胜,熊太纯,柳益君,等.基于 Spark 的高校图书馆文献推荐方案及实证研究[J].图书情报工作,2017,61(23):129-137.
HE S,XIONG T C,LIU Y J,et al.A spark-based scheme of university library literature recommendation and its empirical study[J].Library and Information Service,2017,61(23):129-137.

Research on academic paper recommendation based on Doc2Vec and LDA model and literature quality

Wang Dafu, Deng Zhiwen, Jia Zhiyong, Wang Jing

(Library, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: The content-based recommendation algorithm is used to recommend academic papers with adaptive content and high quality for readers, so as to solve the problems of "information overload" caused by massive electronic resources. The context, word order & global topic information of the thesis text are taken into consideration. Firstly, Doc2Vec and LDA hybrid semantic model are used to train the summary corpus of candidate thesis sets, and the text vector of each thesis is learned. Then, the candidate thesis sets are clustered by K-means algorithm, and then the cluster members of the target papers are searched as the papers to be recommended, Finally, the similarity is calculated and sorted by fusing the literature quality weight, so as to obtain the TOP-*N* nearest neighbor recommendation results. Taking CNKI library & information journal pap as the corpus, an empirical analysis is conducted. Word2Vec & LDA models, the hybrid model adopted in this paper, compared with the traditional TF-IDF, the hybrid model adopted in this paper has higher accuracy and lower ranking difference, and achieves good recommendation results.

Keywords: academic papers; mixed semantic model; literature quality; recommend

[责任编校 陈留院 赵晓华]