doi:10.3772/j.issn.2095-915x.2022.03.006

融合网络表示学习与文本信息的学术文献推荐方法

祝婷

西安工业大学图书馆 西安 710021

摘要：[目的 / 意义] 为了从引文网络、文献内容、标签等多角度挖掘文献间的深层次关系，进而提高学术文献推荐的效果，提出一种融合网络表示学习与文本信息的学术文献推荐方法。[方法 / 过程] 首先，使用网络表示学习、BERT、标签针对文献库分别生成基于引文网络的特征向量表示、基于长文本内容的特征向量表示以及基于短文本标签的特征向量表示；其次，针对网络表示学习及 BERT 生成的向量进行一次特征融合，采用余弦相似性算法分别计算特征融合及标签对应的文献相似度矩阵，并对其进行二次相似度矩阵融合，获取文献综合相似度矩阵；最后，按照相似度大小对待推荐的文献进行排序，实现 Top-N 推荐。[结果 / 结论] 在 CiteUlike 数据集上进行实验验证，相比于对比方法在准确率、召回率和 F 值上平均提升了 31.05%、28.51% 和 29.70%，结果表明本文方法较于单一推荐方法可以有效提高学术文献推荐的质量。

关键词：网络表示学习；Node2vec；标签；BERT；推荐；学术文献

中图分类号：G35；G254

An Academic Paper Recommendation Method by Network Representation Learning and Text Information

ZHU Ting

Library, Xi'an Technological University, Xi'an 710021, China

Abstract: [Objective/Significance] In order to dig out the deep-level relationship between academic papers from multiple perspectives such as citation network, content, and tags, and improve the effect of academic paper recommendation, an academic paper recommendation method by network representation learning and text information is proposed. [Methods/Process] Firstly, use network representation learning, bert algorithm, and tags to generate feature vectors based on citation network, long text content, and short text tags for the paper dataset; Secondly, according to the vectors generated by the network representation learning and bert, the first feature fusion vector is calculated, and based on the first feature fusion vector and the vector calculated by paper labels, the secondary similarity fusion matrix is carried out by the cosine similarity algorithm; Finally, sort

基金项目 陕西省科学技术情报学会项目“网络表示学习在图书馆文献资源推荐系统中的应用研究”（2022KTF-06）。

作者简介 祝婷（1990-），硕士，助理馆员，研究方向为个性化推荐，E-mail: zhting0708@163.com。

引用格式 祝婷 . 融合网络表示学习与文本信息的学术文献推荐方法 [J]. 情报工程 , 2022, 8(3): 81-92.

doi:10.3772/j.issn.2095-915x.2022.03.006

the similarities of papers and realize Top-N recommendation. [Results/Conclusions] The proposed method is validated in the CiteUlike dataset, the precision, recall and F-measure of this method increased by 31.05%, 28.51% and 29.70% on average compared with the comparison method, and the results show that this method can effectively improve the quality of academic paper recommendation compared with single recommendation method.
Keywords: Network representation learning; Node2vec; tag; BERT; recommendation; academic paper

引言

学术文献作为学者在科学研究过程中必不可少的知识源，对于辅助学术研究具有重要意义。然而随着互联网与科学技术的快速发展，学术科研领域中的文献数量与日俱增，一方面，面对海量的学术文献，学者很难从中获取自己所需的文献；另一方面，部分学术文献被淹没，没有展现给学者的机会，造成了资源浪费的现象。在这种场景下，如何快速有效的帮助学者获取自己感兴趣的文献已成为目前广泛关注的研究课题。而学术文献推荐方法可以缓解这种“信息过载”问题，它是从海量的学术文献中挖掘学者可能感兴趣的文献，然后展现给学者，进而实现学术文献个性化推荐的过程。

常见的文献推荐方法有基于内容的文献推荐，基于协同过滤的文献推荐以及混合文献推荐。基于内容的文献推荐是指分别获取用户兴趣和文献内容的向量空间模型，通过匹配两者的相似度，向用户推荐相似度较高的文献。陈长华等[1]利用 Word2Vec 方法对论文进行向量化表示，同时引入论文发表时间权重计算论文间相似性进行推荐。耿立校等[2]使用余弦值 r 与匹配度值 Sim 相结合的方法对传统的基于内容的推荐进行改进。基于内容的推荐方法具备较强的直观性和可解释性，但是该方法只能推荐与用户历史阅读文献相似的文献，缺乏多样性，并且没有考虑用户对文献的评价等信息。基于协同过滤的文献推荐是目前应用最为广泛且成功的推荐方法，它是通过计算用户之间的相似性获取近邻用户，将近邻用户感兴趣的文献推荐给目标用户。陈浩[3]在计算用户相似性时融合了用户点击和搜索词的相似度，进一步改进了基于用户的协同过滤方法。顾明星等[4]对用户属性进行聚类，然后将时间因素引入至评分相似性中，同时将新人误差引入至信任关系中计算用户相似性，提出了一种改进的协同过滤推荐。基于协同过滤的文献推荐可以在一定程度上缓解基于内容推荐的单一性，但是该方法仍旧存在一些问题，例如稀疏性和冷启动问题。为了弥补上述两种推荐方法的缺点同时结合其优点，进而形成了混合推荐方法。王妍等[5]将基于内容的推荐和协同过滤推荐相结合，提出了一种混合论文推荐方法，有效的解决了冷启动问题。王永贵等[6]针对基于内容的推荐和协同过滤算法中存在的问题，提出了一种融合内容与协同矩阵分解技术的混合推荐方法。混合推荐算法框架较为复杂，且推荐效果受单一推荐方法选择的影响。

综上所述，以上方法分别从不同角度对学术文献进行了推荐，并产生了良好的推荐效果，但是忽略了文献间引用关系在推荐过程中的重

要性，进而影响推荐的准确性。在学术文献推荐过程中，最直接的方法是对文献的文本信息进行挖掘从而进行推荐，文本信息包含长文本内容和短文本标签。内容信息是以非结构化的长文本形式描述文献的内容，如文献的摘要、正文等，具有直观性、具体性等特点。标签信息是以结构化的短文本形式描述文献的特征，可以准确的反映用户的喜好，具有规范化、易处理等特点。但是仅仅使用文本信息的推荐方法仍具有局限性，如信息单一，未考虑文献间的交互关系等。除了文献自身的文本信息外，从文献间的引用关系图中可以挖掘更深层次的语义信息，但是由于爆炸式增长的文献数量使得引用关系图中的节点和边往往非常庞大，进而导致文献向量表示出现高维稀疏的问题，网络表示学习方法可以将复杂网络图中的节点表示为低维稠密的向量表示，同时保留原有的网络结构。因此，本文提出一种融合网络表示学习与文本信息的学术文献推荐方法，分别从文献引用网络、长文本内容和短文本标签这三个方面对学术文献进行特征表示，在此基础上计算文献间的综合相似性，进而提高学术文献推荐的质量。

1 相关工作

1.1 网络表示学习

在互联网时代，爆炸式增长的信息资源之间构成了复杂的信息网络，如何将复杂信息网络进行准确的网络表示是目前科学研究的重要过程。网络表示学习（Network Representation Learning, NRL），又称网络嵌入（Network Embedding, NE）或者图嵌入（Graph Embedding, GE），它可以将复杂信息网络中的节点表示为低维、稠密、实值的向量表示[7]，从而解决传统的网络表示使用稀疏高维的向量需要花费大量计算空间和运行时间的问题。

网络表示学习方法一般分为基于矩阵分解的方法、基于随机游走的方法以及基于深度神经网络的方法[8]。基于矩阵分解的方法是针对高维的原始矩阵进行特征值分解、奇异值分解等操作获取节点的低维向量表示。由于该方法的时间复杂度和空间复杂度较高，在大规模的数据应用上并不理想。基于随机游走的方法思想来源于 Word2vec[9-10] 模型，它将节点类比为单词，随机序列类比为句子，进而获取网络嵌入，代表性的算法有 DeepWalk[11]、LINE[12]（Large-scale Information Network Embedding）、Node-2vec[13] 等。DeepWalk 使用网络中节点间的共现关系来学习节点的向量表示，首先采用随机游走算法获取网络中的节点序列，这些节点序列可以看作自然语言处理中的语句，节点序列中的节点可以看作自然语言处理中的单词。其次通过 Word2vec 中的 Skip-Gram 模型对随机游走中的节点进行概率建模，最大化随机游走序列的似然概率。最后使用随机梯度下降法获取节点的向量表示。该方法避免了邻接矩阵需要将所有信息存储在内存而影响到计算效率的问题。其中随机游走策略采用的是一种可重复访问已访问节点的深度优先遍历算法（Depth First Search, DFS）。LINE 是一种基于节点与邻居间关系的大规模信息网络表示学习算法，通过结合一阶相似性和二阶相似性来保存网络结构信息，获取节点嵌入。其中一阶相似性是指节

doi:10.3772/j.issn.2095-915x.2022.03.006

点与直接相连的相邻节点间的相似性，由于一阶相似性不能代表节点的全局网络结构，因此引入具有共同邻居节点的节点间的相似性，即二阶相似性，通过最小化一阶和二阶相似的损失函数获取网络中节点的向量表示。该方法采用了广度优先遍历算法（Breadth First Search, BFS）的思想。Node2vec 对 DeepWalk 进行了改进，同时考虑了广度优先遍历算法和深度优先遍历算法，形成了有偏的随机游走，按照广度优先遍历算法进行游走趋向于节点周围采样序列，按照深度优先遍历算法进行游走趋向于朝更远方向采样序列。因此将两者结合可以获得反应网络全局信息及局部信息的节点序列，然后使用Skip-Gram模型输出节点的向量表示，同时保证了网络的同质性与结构性。基于深度神经网络的方法是利用深层神经网络模型对网络中节点的非线性结构进行建模，进而获取网络节点表示。以上网络表示学习方法在复杂网络上的成功应用，对于学术文献推荐具有重要启示作用。

1.2 文本向量化表示模型

文本向量化表示是将自然语言转化为实数向量，即计算机可以处理的格式。常见的文本向量表示模型有词袋模型（bag of words）、Word2vec 和 Doc2vec 等。词袋模型仅考虑了词频，忽略了语序和语法信息，且易造成词向量的稀疏性和高维性。Word2vec 的基本思想是使用上下文词语预测当前词语或者使用当前词语预测上下文词语，分别对应 Word2vec 中的 CBOW 和 Skip-gram 模型，使用 Word2vec 进行文本向量表示是在 Word2vec 模型生成词向量的基础上，对文本包含的词向量进行加权平均等操作，该方法可以有效解决稀疏问题和维度灾难，但是同样忽略了语序信息。Doc2vec 是 Le 等 [14] 于 2014 年在 Word2vec 的基础上提出来的，区别在于增加了一个与词向量维数相同的段落向量，该模型包含 PV-DM（Distributed Memory version of Paragraph Vector）和 PV-DBOW（Distributed Bag of Words version of Paragraph Vector）。PV-DM 模型与 Word2vec 中的 CBOW 模型相对应，是通过上下文的词向量和段落向量来预测目标词语，PV-DBOW 模型与 Word2vec 中的 Skip-gram 模型相对应，是以段落向量作为输入，输出其段落中词向量的概况分布。与 Word2vec 相比，Doc2vec 不仅考虑了语序信息，而且可以直接将文本向量化，训练过程方便简单。以上文本向量化本质上是一种静态表示方法，不能表达自然语言中的一词多义，也不能获取全局文本语义信息。针对上述问题，Devlin 等 [15] 提出了 BERT 模型，该模型以 Transformer 编码器为主要框架，通过大量通用语料库对预训练获取通用语义信息，并针对专业语料库进行微调，进而更好的实现文本特征表示。

2 研究思路与方法

本文首先根据学术文献库中的引用关系构建学术文献引用网络，使用网络表示学习模型 Node2vec 获取学术文献的向量表示，同时利用 Bert 模型获取学术文献的向量表示。其次对网络表示学习与 Bert 模型生成的向量进行一次特征融合，采用余弦相似性算法分别获取特征融合后向量及基于标签对应的学术文献相似度矩

阵，并对其进行二次相似度矩阵融合，获取文献综合相似度矩阵。最后在文献综合相似度矩阵的基础上，根据其相似性大小实现学术文献推荐。整个推荐流程如图 1 所示。

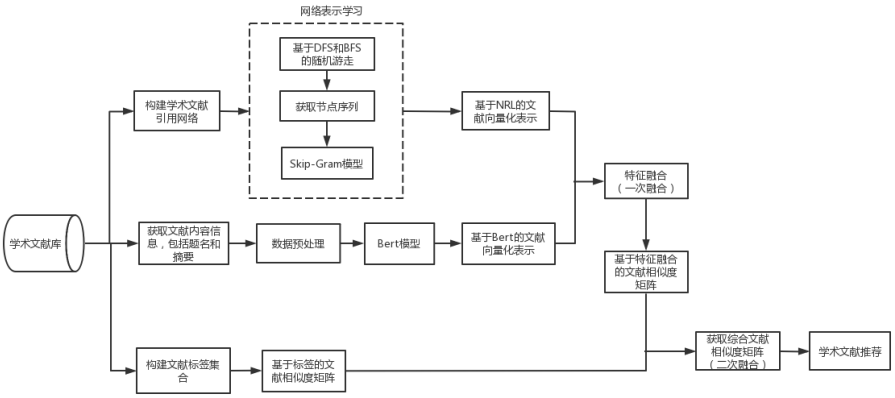

图 1 融合网络表示学习与文本信息的学术文献推荐流程图

2.1 基于网络表示学习的学术文献向量表示

在学术文献库中，文献并不是单独存在，一篇文献通常包含多个引文文献，这些引文文献与该文献在研究内容上有着密切的关系，而引文文献也有下一层的引文文献，这样层层关联便组成了引文网络。网络表示学习方法可以将引文网络中的节点通过非线性模型转化为更高层次的低维稠密的文献向量表示，主要分为以下两个步骤：

（1）构建学术文献引文网络

学术文献引文网络反映了文献间的引用关系，可将其表示为 G=（D,E），其中 D 表示顶点集合，即学术文献集合 $D=\{d_1,d_2,d_3,d_4,d_5,\cdots\}$，集合中每一个顶点 d_i 对应不同的文献；E 表示边集合，即文献间的引用关系集合 $E=\{e_{12},e_{13},e_{23},e_{24},e_{25},\cdots\}$，集合中每一条边 e_{ij} 表示文献 d_i 和文献 d_j 存在引用关系。若一篇文献的引文列表中包含另一篇文献，则两者构成一条边，通过上述规则可构建学术文献引文网络。

（2）基于网络表示学习的学术文献向量表示

基于学术文献引用关系构建的学术文献引用网络，通过对比 DeepWalk、LINE 以及 Node-2vec 等网络表示学习模型。本文选择 Node2vec 对文献引用网络进行训练，以文献引用网络作为输入，通过调整相关参数，获取每个文献对应的低维向量表示。

Node2vec 是在 DeepWalk 网络表示学习模型的基础上，综合广度优先搜索和深度优先搜索思想，通过引入参数 p、q 进行有偏的随机游走，获取随机游走序列，实现通过广度优先搜索获取文献在数据集的微观局部信息以及深度优先搜索获取文献在数据集的宏观全局信息。

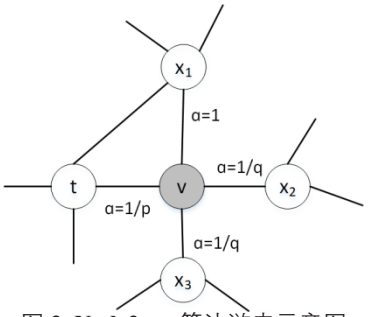

图 2 Node2vec 算法游走示意图

doi:10.3772/j.issn.2095-915x.2022.03.006

如图 2[13] 所示，根据 Node2vec 网络表示学习模型，若游走路径为（t,v），定义 p 为游走至前一文献邻居的概率，则 p 越大，已游走过的文献被再次游走到的概率越低；定义 q 为游走至前一文献非邻居的概率，则当 q>1 时，随机游走将局限于文献 t 附近，反之当 q<1 时，随机游走将远离文献 t；定义 d 为从文献 t 到文献 x 的最短路径，则按照 Node2vec 游走思想，从文献 t 到文献 x 的概率计算方法如公式 1[13] 所示：

$$\alpha_{pq}(t,x)=\begin{cases}\dfrac{1}{p}, & if\ d_{tx}=0\\ 1, & if\ d_{tx}=1\\ \dfrac{1}{q}, & if\ d_{tx}=2\end{cases} \quad (1)$$

基于上述算法获取随机游走序列，选用 Skip-gram 模型对游走序列建模，实现随机游走似然概率最大化，利用随机梯度下降方法获取文献的向量表示。

2.2 基于内容的学术文献向量表示

文献内容是个性化推荐过程中需要关注的重点文本信息，如文献的标题、摘要等。因此，通过文献内容获取特征向量表示，然后在此基础上进行相似性计算是实现学术文献推荐的重要环节。基于内容的学术文献向量表示主要分为以下两个步骤：

（1）数据预处理。主要包括分句、分词和去停用词，对学术文献库中的文献进行预处理的主要目的是生成结构化的文本数据，并且去除无意义的词语，这些是对文献进行特征提取和表示的基础。对于由 m 个文献组成的文献数据集合 $D=\{d_1,d_2,d_3,\cdots,d_m\}$，首先将文献的题目和摘要进行合并，选取 Stanford Tokenizer 英文分词系统将合并后的文本切分为单独且具有语义的词，其次剔除没有实际含义的词，如介词、语气词等，最后将所有文献转化为这些词的集合 $d_i=\{w_1,w_2,w_3,\cdots,w_n\}$。

（2）学术文献向量化表示。Bert 模型以多层 Transformer 编码器为主要框架，基于其注意力机制获取词的表征信息，该表征信息包含了该词本身语义和该词与文本其他词的关系，进而获得该词的上下文语义信息。另外，Bert 模型利用掩蔽语言模型（Masked Language Model, MLM）和下一句预测（Next Sentence Prediction, NSP）技术进行预训练，并针对具体任务进行微调，从而获得较好的特征提取和分类表现。对于学术文献推荐领域，由于文献标题、摘要等信息从整体上实现了对文献的描述，因此计算文献标题、摘要等文献内容的特征向量有助于实现文献推荐。本文数据集为英文数据集，且不区分大小写，故选用 BERT-Base-Uncased 生成学术文献向量表示。

2.3 基于标签的学术文献相似性计算

除了文献中以长文本形式存在的文献内容外，标签也是描述一篇学术文献的重要信息，它以不同粒度反应了文献的主题特征，同时也包含了文献中可能未提及的重要关键词或词组，是学术文献推荐方法中重要的数据源之一，被广泛应用于推荐系统中。标签数据一方面可以规范且直观的揭示文献的特征，另一方面具备易抽取性和语义性，因此将标签作为学术文献的特征表示来计算其相似性，对于提高学术文献推荐的准确性具有重要意义。

由于标签通常是由简短的词语或词组组成，不需要像处理长文本那样进行语义分析。

因此本章节选择 Jaccard 系数即文献对应标签集合间的共现关系来计算文献间的相似性，设定文献 d_1 由 m 个标签组成，文献 d_2 由 n 个标签组成，文献 d_1 和文献 d_2 的标签集合分别表示为 $Tag_{d_1}=\{a_1, a_2, \cdots, a_m\}$ 和 $Tag_{d_2}=\{b_1, b_2, \cdots, b_n\}$。在此基础上获取基于标签的文献相似度矩阵，其中文献 d_i 和文献 d_j 间相似性计算方法如公式 2 所示：

$$Sim\_tag(d_1,d_2)=\frac{|Tag_{d_1}\cap Tag_{d_2}|}{|Tag_{d_1}\cup Tag_{d_2}|} \quad (2)$$

2.4 融合网络表示学习与文本信息的学术文献推荐

通过上述方法获取两种学术文献向量表示与基于标签的学术文献相似度矩阵后，如何融合这些信息成为学术文献推荐的关键。基于网络表示学习的学术文献向量体现了文献引用网络结构中的语义信息，基于内容的学术文献向量体现了文献描述的内容信息，将两者进行特征融合可以充分挖掘文本信息，同时保留文献间的引用关系。因此本文首先使用向量拼接的方法对基于网络表示学习及内容的学术文献向量进行一次特征融合，然后计算特征融合后向量的学术文献相似度矩阵，常用的相似性算法包含余弦相似性算法、Jaccard 系数与欧氏距离相似度等。本文选择余弦相似性算法获取文献相似度矩阵，其中文献 d_i 和文献 d_j 间相似性计算方法如公式 3 所示，其中 k 代表文献向量的维数。

$$Sim\_Node2vec\_Bert(d_1,d_2)=\frac{\sum_{i=1}^{k} d_{1,i}\times d_{2,i}}{\sqrt{\sum_{i=1}^{k}(d_{1,i})^2}\times\sqrt{\sum_{i=1}^{k}(d_{2,i})^2}} \quad (3)$$

基于网络表示学习及 Bert 模型的学术文献向量进行特征融合，并计算出相似度矩阵后，考虑到不同方法计算出的相似度矩阵代表了不同的意义，其重要程度不同。因此，本文将基于特征融合的学术文献相似度矩阵与基于标签的学术文献相似度矩阵以一定的权值加权求和获取文献的综合相似度矩阵，其中文献 d_i 和文献 d_j 间综合相似性计算方法如公式 4 所示。

$$Sim(d_1,d_2)=\alpha * Sim\_Node2vec\_Bert(d_1,d_2)+(1-\alpha)*Sim\_tag(d_1,d_2) \quad (4)$$

在获取文献的综合相似度矩阵后，接下来将待推荐的文献与用户喜好或已发表文献进行匹配，按照相似性大小对其排序，最终选取相似性最高的 Top-N 个文献推荐给用户。

3 实验设计与评价

3.1 数据集

本文选择 CiteUlike 数据集作为实验数据源，该数据集是由施普林格出版社 (Springer) 提供的免费的在线科研平台，科研人员可以在该平台上发现、存储、组织和管理学术文献等，最终形成个人文献库，数据集的具体内容信息及统计数据如表 1、表 2 所示。为了保证实验数据的准确性，并将其转化为计算机可识别的数据，需要对其进行预处理操作，具体步骤如表 3 所示。

表 1 CiteUlike 数据集

序号	数据文件名	文件内容
1	users.dat	存储了用户浏览的文献编号
2	raw-data.csv	存储了文献信息，包括编号、题名和摘要
3	citations.dat	存储了文献之间的引用关系
4	tags.dat	存储了标签编号及信息
5	item-tag.dat	存储了每篇文献对应的标签编号

doi:10.3772/j.issn.2095-915x.2022.03.006

表 2 CiteUlike 数据集统计

序号	实体或关系名称	数量
1	用户数	5551
2	文献数	16980
3	标签数	46391
4	文献间引用关系	44709
5	用户与文献间关系	204987

表 3 数据预处理

步骤	预处理内容	说明
1	文献引用关系预处理	删除文献引用关系中引用数小于10的文献
2	文献标签预处理	删除文献标签数据里标签为0的文献
3	文献索引预处理	删除剩余数据集中由1和2引起的关联文献，调整文献索引编号，形成新的引用关系
4	文献内容预处理	对文献题名和摘要进行拼接，并去停用词

3.2 对比实验

为了客观分析融合网络表示学习与文本信息的学术文献推荐方法的效果，本文选择以下模型产生的推荐方法作为对比实验进行评价。

（1）DeepWalk：在文献引用网络中，采用随机游走算法获取网络中的节点序列，通过 Skip-Gram 模型学习节点的向量表示。

（2）Line：利用一阶相似性和二阶相似性来保存网络结构信息，学习节点的向量表示。

（3）Node2vec：对 DeepWalk 随机游走的方式进行改进，综合考虑基于深度优先搜索和广度优先搜索的随机游走策略，进而获取节点的网络嵌入表示。

（4）Tag：使用 Jaccard 系数即文献对应标签集合间的共现关系来计算文献间的相似性。

（5）Bert：使用 Bert 模型中 BERT-Base-Uncased 版本对文献内容进行向量表示。

在以上模型的基础上，获取文献相似度矩阵，将与目标文献相似的前 N 个文献推荐给用户。

3.3 评价标准

为了评价融合网络表示学习与文本信息的学术文献推荐的效果，本文选择准确率 (Precision)、召回率 (Recall) 和 F-measure 值作为评价标准。

（1）Precision

准确率可以衡量推荐文献的精准性，它是指在推荐列表中用户真实喜好的文献所占的比例，计算方法如公式 5 所示，针对所有用户推荐的准确率求平均值可以获取整体准确率。

$$Precision = \frac{|R(u) \cap T(u)|}{|R(u)|} \tag{5}$$

其中，R(u) 为给用户 u 推荐的文献集合，T(u) 为测试集中用户 u 喜好的文献集合。

（2）Recall

召回率可以衡量推荐文献的全面性，它是指用户真实喜好的文献被推荐的概率，即推荐列表中用户真实喜好的文献与测试集中用户所有喜欢的文献比例，计算方法如公式 6 所示，针对所有用户推荐的召回率求平均值可以获取整体召回率。

$$Recall = \frac{|R(u) \cap T(u)|}{|T(u)|} \tag{6}$$

（3）F-measure

随着准确率的增加，而召回率会减小，两者是相互矛盾又统一的指标。F-measure 值综合考虑了两者，对准确率和召回率进行加权调和平均，可以综合体现推荐结果的准确性和全面性，计算方法如公式 7 所示。

$$F\text{-}measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (7)$$

3.4 实验结果与讨论

3.4.1 实验分析

在学术文献推荐的过程中，由于针对每篇目标文献的推荐个数 n、综合相似性权值 α 以及不同融合方式对推荐结果有着重要影响。因此本节将讨论 n 和 α 这两个参数在不同数值设置下以及不同融合方式对实验的影响。

（1）推荐个数对实验结果的影响

为了分析推荐个数对学术文献推荐效果的影响，在保持其他参数不变的情况下，分别设置推荐个数为 120、140、160、180、200，计算对应的 Precision、Recall 和 F-measure 值，实验结果如图 3—图 5 所示。从图 3—图 5 可以看出，总体上，随着推荐个数的增加，大部分推荐方法的 Precision 值呈略微下降趋势，Recall 及 F-measure 值呈上升趋势。这是因为随着推荐文献数的增多，排名靠后的文献与用户的偏好相差较大，但是可以提升推荐文献的全面性。当推荐个数分别等于 120、140、160、180、200 时，本文推荐方法的 Precision、Recall 和 F-measure 值均高于其他对比方法；当推荐个数等于 120 时，各方法的 Precision 达到最大值，但是与其他推荐个数对应的 Precision 值差别不大；当推荐个数等于 200 时，Recall 和 F-measure 值达到最大值，与其他推荐个数对应的 Recall 和 F-measure 值差别较大，同时考虑到 F-measure 值可以综合体现推荐结果的准确性和全面性，因此本文选取 n=200 为最优推荐个数。

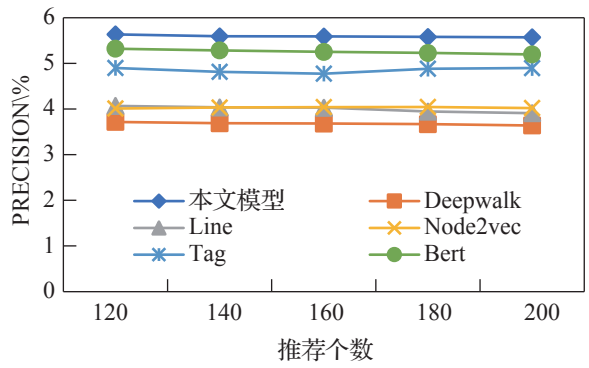

图 3 不同推荐个数下各推荐方法的准确率对比

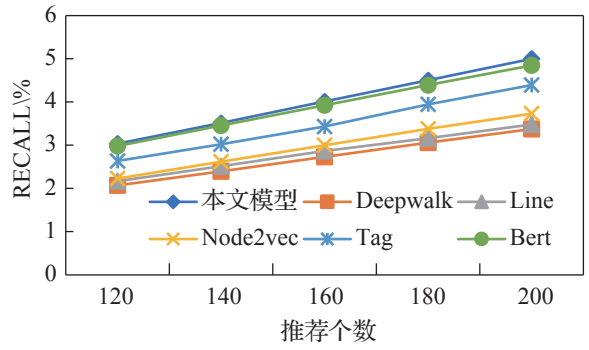

图 4 不同推荐个数下各推荐方法的召回率对比

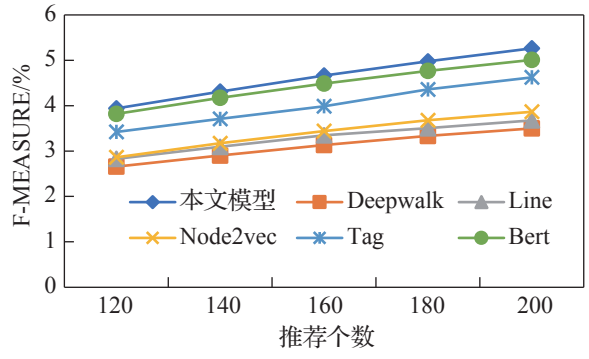

图 5 不同推荐个数下各推荐方法的 **F-measure** 对比

（2）综合相似性权重对实验结果的影响

针对本文提出的推荐方法，在计算综合相似度矩阵时，为了分析网络表示学习、文本内容、标签这三个角度对应的相似度权重分配对推荐结果的影响，在保持其他参数不变的情况下，分别设置 α 值为 0、0.2、0.4、0.6、0.8、1，计算对应的 Precision、Recall 和 F-measure 值，

doi:10.3772/j.issn.2095-915x.2022.03.006

实验结果如图 6 所示。从图 6 可以看出，随着 α 值的上升，Precision、Recall 和 F-measure 值均呈现先上升后下降的变化趋势；当 α=0 或 1 时，表示未对相似度矩阵进行融合，Precision、Recall 和 F-measure 值达到最低和次低，说明将文献引用关系、文本内容和标签进行融合可以提高推荐的效果；当 α=0.4 时，本文推荐方法取得最优值，因此本文选取 α=0.4 为最优权值。

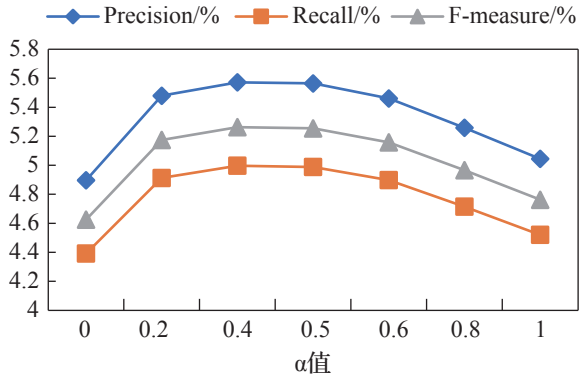

图 6 不同权值下本文推荐方法的准确率、召回率及 **F-measure** 对比

（3）不同融合方式对实验结果的影响

为了进一步分析文献引用关系与文本信息融合过程中，不同融合方式对文献推荐效果的影响，本文分别计算以下两种融合方法对应的 Precision、Recall 和 F-measure 值，结果如图 7 所示。

①特征与相似度融合。将 node2vec 与 bert 模型生成的文献特征向量拼接获取融合后的特征向量，计算其相似度矩阵，然后与基于标签的文献相似度矩阵进行加权二次融合。

②相似度融合。将 node2vec、tag 与 bert 模型对应的三种文献相似度矩阵以一定的权值加权求和获取文献的综合相似度矩阵。

通过图 7 可以看出，特征与相似度融合方法对应的 Precision、Recall 和 F-measure 值较高，因此本文选择先进行特征融合、后进行相似度融合的方法进行学术文献推荐。

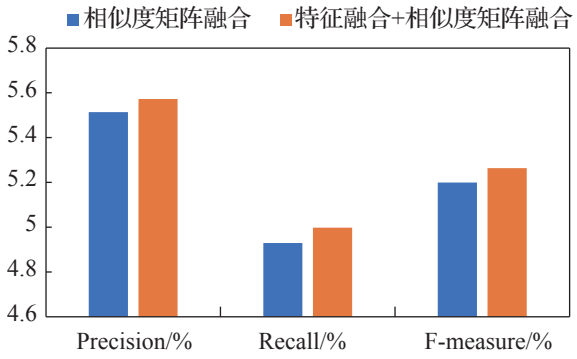

图 7 不同融合方式下准确率、召回率及 **F-measure** 对比

3.4.2 实验结果

根据上述实验分析，本文最终选取推荐个数 n=200, 综合相似性权重 α=0.4，以及特征与相似度融合方法进行实验，得到本文推荐方法与对比方法对应的 Precision、Recall、F-measure 以及相比对比方法本文推荐方法的提升率，实验结果如表 4 所示。从表 4 可以看出，本文推荐方法的 Precision、Recall 和 F-measure 均高于其他 5 种对比方法，且平均提升率分别为 31.05%、28.51% 和 29.70%，说明本文提出的融合网络表示学习与文本信息的学术文献推荐方法具有更好的推荐效果。除此之外，在网络表示学习的推荐方法中，基于 Node2vec 的方法优于基于 DeepWalk、Line 的方法，这是因为 Node2vec 综合考虑了广度优先遍历算法和深度优先遍历算法，可以同时保证文献在数据集上的局部信息和全局信息。在基于本文信息的推荐方法中，基于 Bert 的方法优于基于 Tag 的方法，说明使用 Bert 对标题摘要进行特征提取比文献的标签更能准确表示文献内容信息。综上

所述，本文使用 Node2vec、Bert 模型、Tag 从文献间的引用关系、内容信息和标签三个角度进行融合，进而实现学术文献推荐，对于提高推荐方法的质量具备一定的优势。

表 4 本文方法与对比方法推荐结果对比

模型	Precision（%）	Precision 提升率	Recall（%）	Recall 提升率	F-measure（%）	F-measure 提升率
本文模型	**5.572345**	/	**4.997320**	/	**5.263339**	/
DeepWalk	3.639570	53.10%	3.375470	48.05%	3.499181	50.42%
Line	3.910421	42.50%	3.479470	43.62%	3.678384	43.09%
Node2vec	4.020847	38.59%	3.730234	33.97%	3.866411	36.13%
Tag	4.896827	13.80%	4.391799	13.79%	4.625471	13.79%
Bert	5.194888	7.27%	4.846277	3.12%	5.010099	5.05%
平均提升率	/	**31.05%**	/	**28.51%**	/	**29.70%**

4 结语

针对传统的学术文献推荐忽略了文献间引用关系的重要性，以及文献向量表示维数过大进而影响推荐效果的问题，本文提出了融合文献引用网络、长文本内容和短文本标签的学术文献推荐方法。首先，分别利用 Node2vec、Bert 模型生成文献向量表示，并对其进行特征融合，计算特征融合和标签对应的文献相似度矩阵；其次，加权两种文献相似度矩阵获取文献综合相似度矩阵，根据与目标文献的相似性大小实现学术文献推荐；最后，在 CiteUlike 数据集上进行实验验证，结果表明本文方法在 Precision、Recall 和 F-measure 上均有一定的提升，验证了网络表示学习融入至基于文本信息的推荐方法中的有效性。由于本文仅在单一数据集上进行了验证，因此具有一定局限性。除了文献间引用关系外，用户间的社交关系、文献包含的多种特征信息以及用户与文献间的评分关系也是学术文献推荐过程中需要关注的重点信息，如何将这些信息引入至文献引用关系网络中进行推荐将是本文下一步的研究重点。

参 考 文 献

[1] 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.

[2] 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.

[3] 陈浩 . 基于协同过滤算法的论文推荐系统研究与设计 [D]. 武汉 : 武汉科技大学 , 2018.

[4] 顾明星 , 黄伟建 , 黄远 , 等 . 结合用户聚类与改进用户相似性的协同过滤推荐 [J]. 计算机工程与应用 , 2020, 56(22):185-190.

[5] 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.

[6] 王永贵 , 陈玉伟 . 融合内容与矩阵分解的混合推荐算法 [J]. 计算机应用研究 , 2020, 37(5):1359-1363.

[7] 丁钰 , 魏浩 , 潘志松 , 等 . 网络表示学习算法综述 [J]. 计算机科学 , 2020, 47(9):52-59.

[8] 周晓旭 , 刘迎风 , 付英男 , 等 . 网络顶点表示学习方法 [J]. 华东师范大学学报 (自然科学版), 2020(5):83-94.

[9] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]. Proceedings of Advances in Neural Information Processing Systems. Lake Tahoe.

doi:10.3772/j.issn.2095-915x.2022.03.006

2013:3111-3119.

[10] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013:1-12.

[11] Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online learning of social representations[C]. Proceedings of the 20th ACM SIG - KDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press. 2014: 701-710.

[12] Tang J, Qu M, WANG M Z, et al. LINE: Large-scale information network embedding[C]. Proceedings of the 24th International Conference on World Wide Web. Switzerland: International World Wide Web Conferences Steering Committee. 2015: 1067-1077.

[13] Grover A, Leskovec J. Node2vec: Scalable feature learning for networks[C]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press. 2016: 855-864.

[14] Le Q V, Mikolov T. Distributed representations of sentences and documents[C]. Proceedings of the 31st International Conference on Machine Learning. 2014.

[15] Devlin J, Chang M W, Lee K, et al. BERT: Bidirectional Encoder Representations from Transformers for language Understanding[J]. Computation and Language, 2018, 23(2):3-19.