文章编号：2095-2163(2020)07-0113-04 中图分类号：TP301 文献标志码：A

基于卷积神经网络的学术合作者推荐研究

周亦敏，黄 俊
（上海理工大学光电信息与计算机工程学院，上海 200000）

摘 要：随着论文数量和种类的快速增长，越来越需要更先进的工具来帮助学术数据的探索，关于合作者的推荐问题成了近几年研究的重点。为了解决这一问题，本文提出了一种基于卷积神经网络的学术合作者推荐算法。通过卷积神经网络(CNN)去学习论文摘要的情境特征，并使用 PMF 去学习研究员-主题矩阵的隐层特征，使用孪生网络来比较两个研究员间的特征相关度，根据特征间的相似度高低进行推荐。实验证明，该模型在合作者推荐方面的推荐精度优于其对比模型。

关键词：孪生网络；合作者推荐；卷积神经网络

Research on CNN-based academic collaborator recommendation

ZHOU Yimin, HUANG Jun
(School of Optical-Electrical & Computer Engineering, University of Shanghai for Science & Technology, Shanghai 200000, China)

【Abstract】 With the rapid growth of the number and types of papers, more and more advanced tools are needed to help the exploration of academic data, and the recommendation of collaborators has become the focus of research in recent years. In order to solve this problem, this paper proposes an academic collaborator recommendation algorithm based on convolutional neural network. Convolutional neural network (CNN) is used to learn the situation features of the abstract, and the hidden layer features of researcher topic matrix is learned through Probabilistic Matrix Factorization(PMF). Siamese Network is used to compare the feature correlation between the two researchers, and the recommendation is based on the similarity between the features. Finally, experiments show that the accuracy of this model is better than that of its comparative model.

【Key words】 Siamese Network; Collaborator Recommendation; Convolutional Neural Networks; Matrix Decomposition; GloVe

0 引 言

许多学术搜索引擎，如：微软学术搜索、谷歌学者、ArnetMiner 和 CiteSeerX 的出现，方便了人们探索大量的数字学术资料[1]。随着论文数据量和种类的快速增长，需要更先进的工具来帮助探索学术数据，许多研究人员投入了巨大的努力来促进各种类型的数据进行有意义的应用[2]。但是学术合作者推荐长期以来，一直得不到有效的发展，其原因是研究员的合作者潜在的特征不容易被挖掘，或是挖掘出的特征不够准确[3]。

在近几年，许多学者提出了这一问题的解决方法。Chen 和 Gou 等人，利用论文的主题，将论文库按照主题转变嵌入为研究员-主题矩阵，通过矩阵分解来学习其中的隐层特征[4]；Sun 和 Barber 等人将论文的主题与研究者映射到网络结构中，通过学习网络中节点的属性，以及聚合多个节点的属性，进行链路预测，从而进行合作者的推荐[5]。实验证明这些研究对合作者推荐的推荐精度有所提高，但是它们都只是单纯的考虑了主题与研究员之间的影响。在现实生活中，学术合作本质上是“情境依赖”的，不能单纯的考虑论文主题[6]。

本文提出了一种基于卷积神经网络的学术合作者推荐算法 TCCR（Themes and Context - aware Collaborator Recommendation）。通过卷积神经网络(CNN)去学习论文摘要的情境特征，使用 PMF 学习研究员-主题矩阵的隐层特征，最后使用孪生网络来比较两个研究员间的特征相关度，相关度高的研究员即为推荐合作对象。

1 卷积神经网络与孪生网络相关介绍

1.1 卷积神经网络

卷积神经网络是一种用于图像分类、相似性聚类和场景内目标识别的 DNN，其结构如图 1 所示。通过一个称为滤波器的移动函数对输入图像进行分析，过滤器从图像中提取特征。卷积层输出的激活图是下采样的，池化层通常执行此任务。全连接层将输出分类为类，卷积神经网络有不同的结构。可

基金项目：上海市科委科研计划项目(17511107203)。

作者简介：周亦敏(1962-)，男，硕士，副教授，主要研究方向：嵌入式系统、计算机系统结构、网络应用与智能设备等；黄 俊(1995-)，男，硕士研究生，主要研究方向：推荐系统、数据挖掘。

通讯作者：黄 俊 Email：1615753305@qq.com

收稿日期：2020-04-13

以在深度上寻求更大的模型,即每层的层数或神经元数目,但可能会面临以下问题:第一,由于训练不足,容易出现过度拟合;第二,大量的参数指向系统需要更多的计算能力。

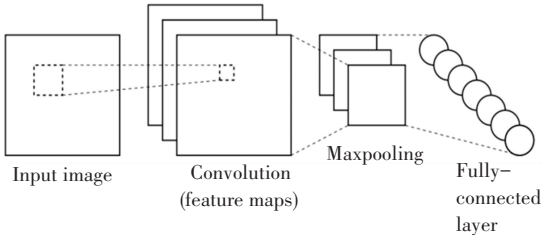

图 1 卷积神经网络结构图

Fig. 1 Structure of CNN

1.2 孪生网络框架

孪生神经网络的概念最早是由 LeCun 等人提部分为应用于人脸识别。使用孪生模型进行二值散列的主要动机是能够学习图像的相似性或相异性,检索系统所需要的特征。相似的图像必须有相似的特征嵌入,而不同的图像必须在特征空间中离得更远。孪生模型试图将相似的特征向量聚合在一起,不同的特征向量被推到一个边界之外。

孪生神经网络由两个前馈分支组成,但权值在两个分支之间共享。本文实验中使用的孪生神经网络的结构如图 2 所示,由三个卷积层组成。每一个卷积层后接校正线性单元(ReLU)、非线性和最大池化运算。孪生模型的每个分支都充当特征提取器,有两个完全连接的层,完全连接层包含可变数目的节点 n,生成的二进制代码中的位数。这个完全连接层的输出通过 sigmoid 层将每个节点的输出压缩为 0 或 1。

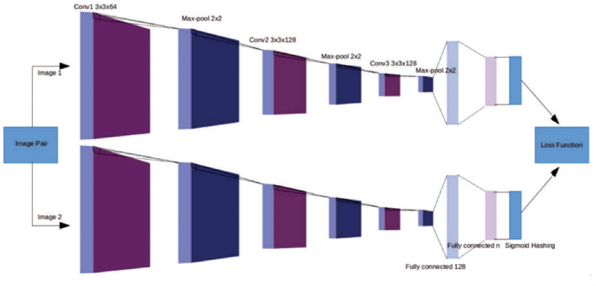

图 2 孪生神经网络结构图

Fig. 2 Structure of Siamese Network

损失函数模块尝试计算从网络的两个分支提取的特征向量之间的距离 D。网络试图最小化的损失函数定义为式(1):

$$L = \frac{1}{2N}\sum_{n=1}^{N} yd^2 + (1-y)\max(\text{margin} - d, 0)^2. \tag{1}$$

其中,$d = \| a_n - b_n \|_2$,代表两个样本特征的欧氏距离;y 为两个样本是否匹配的标签,$y = 1$ 代表两个样本相似或者匹配;$y = 0$ 则代表不匹配;$margin$ 为设定的阈值。

2 TCCR 推荐算法模型

本文提出的推荐模型 TCCR 分为二个部分,第一部分为研究员特征的提取,通过 CNN 去学习论文摘要的隐层情境特征,通过 PMF 去学习研究员-主题矩阵的隐层特征,并使二类特征有效的结合起来;第二部分为特征对比阶段,本文使用孪生网络框架进行特征相似度对比。TCCR 模型结构如图 3 所示。

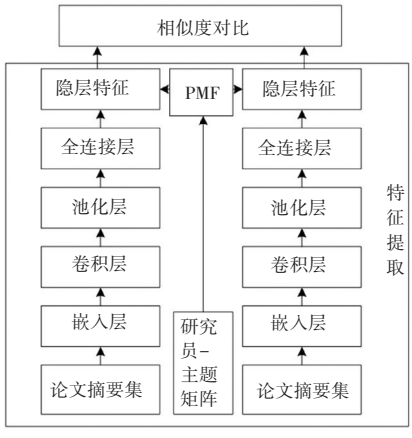

图 3 TCCR 模型结构图

Fig. 3 Structure of TCCR

2.1 基于 CNN 的摘要特征提取

这部分模型主要是用来提取研究员对论文摘要的情境特征模型。通过 CNN 来处理摘要文本集,分析研究员研究方向的情境特征。

输入层,输入的内容为论文摘要集。因为摘要文本通常是长文本且语义情感上下文关联性强,本文采用 GloVe 来获得词语映射成为词向量。通过 GloVe 将文本集中的每个词映射成为词向量,并且按照文本的顺序把这些词向量组成成为摘要文本矩阵。

卷积层,这一层通过卷积运算,提取出摘要文本中的语义信息。通过高度为 h 的卷积核在摘要文本矩阵上移动,与重合部分做卷积操作,来生成摘要文本的特征矩阵,可用公式(2)表示。

$$c_i = f(W \cdot s_{i:i+h-1} + b). \tag{2}$$

其中,f 表示 ReLU 激活函数;b 为偏置;W 表示卷积核。在卷积核中,由序列 $s_{i:i+h-1}$ 卷积后的语义特征可以表示为 c_i。

为了尽可能多的去学习摘要文本中的不同特征,本文使用了三组卷积核。在某组卷积核中,将捕

获到的摘要评论文本的特征向量表示为式(3)：

$$c_k = [c_1, c_2, c_3, \cdots, c_{n-h+1}]. \quad (3)$$

其中，c_k 为第 k 组滤波器提取出的摘要文本特征。

池化层，本文采用最大池化的方法来进行池化操作。把每组卷积核中抽取到的特征向量取最大的特征值，按照一定的顺序连接起来。为了防止卷积神经网络在训练时出现过拟合情况，本文模型在训练时会在全连接层之前加入 dropout。

全连接层，池化层的输出经过 flatten 处理为 N×1 的特征向量。

输出层，通过 softmax 函数，计算出摘要的特征向量在每一种分类上的分布概率，式(4)。

$$h = soft\ \max(W \cdot P + b). \quad (4)$$

其中，W 和 b 表示为预测概率的矩阵和偏置；P 为全连接层的 $N \times 1$ 的特征向量，也即是研究员在摘要上的特征向量；h 为研究员所属的研究方向的概率。

2.2　基于 PMF 的研究员—主题矩阵

为了更全的考虑推荐合作者的适合性，本文使用 PMF 学习到的特征去修正模型。使用 one-hot 编码将研究员与研究主题映射成为研究员-主题矩阵。PMF 模型可以将研究员-主题矩阵拆分为研究员特征和主题特征，公式(5)。

$$E(U,V) = \frac{1}{2}\sum_{i,j}(R_{ij} - U_i^{\mathrm{T}} \cdot V_j)^2 + \frac{1}{2}\lambda^U \| U_i \|^2 + \frac{1}{2}\lambda^V \| V_j \|^2. \quad (5)$$

其中，U、V 分别表示研究员特征和主题特征；R_{ij} 表示研究员 i 研究主题 j 的概率；λ^U 和 λ^V 分别为 U、V 的正则化系数。采用梯度下降法得到 U、V。

合作者推荐一般是基于研究员的，故本文只使用研究员特征 U。为了全面的考虑研究员的主题偏好和情境问题，本文将两类特征使用直接拼接的方法结合起来，式(6)。

$$Q = P + U. \quad (6)$$

其中，Q 为结合后研究员特征向量。

2.3　相似度比较

本文使用孪生网络框架来进行对比，分为二个部分。第一部分为特征提取部分；第二部分为特征相似度比较部分。在特征提取时，本文使用二个 CNN 同时学习不同研究员的摘要文本集，所有的参数使用一致的参数，利于对比特征。第二部分，本文使用的是欧氏距离来进行对比，公式(7)。

$$o_{i,j} = \frac{1}{1 + |Q_i - Q_j|}, \quad (7)$$

其中，$o_{i,j}$ 即为研究员 i 与研究员 j 的特征相似度。数据集中全部作者的特征均学习后，对某个研究员的合作者推荐，应该推荐与其特征相似度排名前 K 的所有研究员。

2.4　模型训练

为了加快训练速度，应该将 PMF 模型提前单独训练，提高隐层特征的获取精度，再与 CNN 模型联合训练。本文使用的损失函数为式(8)：

$$Loss = \sum_{(i,j)\in R}(\hat{o}_{i,j} - o_{i,j})^2. \quad (8)$$

其中，$\hat{o}_{i,j}$ 表示为研究员 i 与研究员 j 的预测相似度，$o_{i,j}$ 表示为研究员 i 与研究员 j 的真实相似度。对于神经网络的优化算法，采用 *Adam* 算法以提升训练速度。

3　实验与评价

3.1　实验环境与数据集

本文采用的电脑配置为 Intel Core i7-8700K 的 CPU、GTX2080 显卡、32G 内存，软件平台为 tensorflow 1.11.0 版本的深度学习框架。

在本文的实验中，采用 Citation 数据集进行数据分析，该数据集共包含 629 814 篇学术文献和130 745 名来自数据库和信息系统相关社区的研究者。通过预处理后共获得 13 379 个关键字，每个关键字都被视为一个独特的主题。因为论文摘要有很多无关词语，对本文进行了必要的清理和整合，如去除停用词等，再者摘要有长有短，在 GloVe 提取词向量时，难以操作。本文使用选取一个能覆盖大部分摘要文本长度的最大文本词覆盖长度，来预处理摘要文本集，使每个摘要长度都在最大覆盖长度以下。本文选取其中的600 000篇学术文献，按照 8∶1∶1 的比例进行训练集、验证集以及测试集的划分。

3.2　评价指标与超参数的选取

为了对比模型的性能，本文使用平均绝对误差(MAE)来衡量预测评分的准确性，公式(9)。平均绝对误差的值越低，则模型性能越好。

$$MAE = \frac{1}{N}|\hat{o}_{i,j} - o_{i,j}|. \quad (9)$$

其中，N 表示测试数据的数目。

选用的三组卷积核高度分别为 3、4、5，卷积核的数量为 150。设置的挖掘潜在特征数为 9，学习率为 0.01，dropout 设置为 0.5，使用的 batch size 为 32。

为了提高摘要文本挖掘语义情感特征的准确

度,本文使用实验数据集中的全部摘要文本作为训练文本集,通过 GloVe 训练出词向量,词向量的维度设置为 200。

3.3 对比模型

为了验证本文提出的推荐算法的性能,将本文模型与以下几种推荐算法进行比较:

Logistic 回归(LR)。LR 是最流行的有监督学习方法之一,在推荐系统中也得到了广泛的应用。在实验中,LR 的输入特征是研究者的嵌入和话题嵌入的平均值的级联。

深层网络结构嵌入(SDNE)。SDNE 代表为编码实体及其关系的结构信息而设计的方法。当应用于学术合作者推荐问题时,研究员被视为实体,在特定主题中的合作被视为情境关系。由于主题的组合,实际上存在无限数量的情境关系,因此不能直接采用像这样的常规方法[7]。

卷积矩阵因子分解模型(ConvMF):ConvMF 使用 CNN 挖掘摘要文本的上下文特征,将挖掘到的特征输入到 PMF 中进行评分预测。

3.4 实验结果与分析

本文实验分为二个部分:第一部分为与其他模型对比推荐 Top-K 性能;第二部分为进一步研究融合 PMF 模型对推荐精度的影响。

不同模型基于四个数据集的总体推荐见表 1。

表 1 推荐算法性能对比

Tab. 1 Performance of recommended algorithm

	Top-K@3	Top-K@6	Top-K@9	Top-K@12
LR	0.971	1.023	1.011	1.083
SDNE	0.842	0.996	0.914	0.869
ConvMF	0.710	0.782	0.757	0.734
TCCR	0.687	0.734	0.727	0.694

由数据可以发现,TCCR 模型在数据集上的推荐性能都优于其他对比模型。LR 仅考虑了研究员和论文主题的相似度给出的推荐,没有考虑他们间的情境特征,故 MAE 最高;SDNE 使用 Network Embedding 方法,考虑了研究员和研究主题间的上下文关系,但没有应用到研究员整个学术生涯的兴趣取向,不符合实际,因此推荐效果也欠佳;TCCR 模型虽与 ConvMF 模型都在算法中都使用了 CNN,但是在实验中前者的推荐效果明显优于后者,反映出合理的融合 PMF 模型可以提高推荐精度。本文为了进一步验证这一结果,设计了实验 2。

在实验中,不使用 PMF 模型的做法是将 CNN 提取出的特征直接对比相似度,其余操作不变,在数据集上与有 PMF 的 TCCR 模型对比性能见表 2。有 PMF 时比 PMF 时推荐结果更为精确。结合实验 1 中的数据,证明合理的融入 PMF 可以使推荐效果得到提升。对于此结果,本文认为合理的解释为良好的建模有助于挖掘数据的特征,使模型学习到的特征更为精准,本文不仅考虑了论文的情境问题还考虑了主题相关性的问题,使学习模型更接近现实生活,从而提升推荐精度。

表 2 PMF 对 TCCR 推荐性能的影响

Tab. 2 Impact of PMF on TCCR

	Top-K@3	Top-K@6	Top-K@9	Top-K@12
TCCR(无 PMF)	0.895	0.944	0.925	0.908
TCCR	0.687	0.734	0.727	0.694

4 结束语

本文基于卷积神经网络的学术合作者推荐算法 TCCR,利用 CNN 学习基于摘要评论文本的隐层特征;利用 PMF 学习基于研究员-主题矩阵的特征;最后有效的将二种特征结合起来,通过孪生网络框架对比特征相似度。通过在数据集上的对比实验,证明了 TCCR 模型较好的推荐性能。

参考文献

[1] HE Q, PEI J, KIFER D, et al. Context-aware citation recommendation[C]//The web conference, 2010: 421-430.

[2] HUANG W, KATARIA S, CARAGEA C, et al. Recommending citations: Translating papers into references [C]// Acm International Conference on Information & Knowledge Management. ACM, 2012:1910-1914.

[3] REN X, LIU J, YU X, et al. ClusCite: effective citation recommendation by information network-based clustering[C]// Knowledge discovery and data mining, 2014: 821-830.

[4] CHEN H H, GOU L, ZHANG X L, et al. CollabSeer: A Search Engine for Collaboration Discovery[C]// Proceedings of the 2011 Joint International Conference on Digital Libraries, JCDL 2011, Ottawa, ON, Canada, June 13-17, 2011. ACM, 2011:231-240.

[5] SUN Y, BARBER R, GUPTA M, et al. Co-author Relationship Prediction in Heterogeneous Bibliographic Networks [C]// International Conference on Advances in Social Networks Analysis & Mining. IEEE, 2011:6-9.

[6] TANG J, WU S, SUN J, et al. Cross-domain collaboration recommendation [C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. 2012: 1285-1293.

[7] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[C]//Advances in neural information processing systems. 2013: 2787-2795.