基于BERT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 200093)

摘 要:学术合作者推荐是学术大数据的一个有效应用。但是现存的方法忽略了学术研究者和研究主题间的上下文关系,因此不能推荐合适的合作者。该文提出了基于BERT的合作者推荐(BACR),旨在推荐高潜力的合作者以达到研究者的要求。为此,设计了一个新的推荐框架,它有两个基本组成部分:BERT(bidirectional encoder representations from transformers)预训练语言模型和逻辑回归模型(LR)。其中,BERT将研究者和研究主题联合表示得到句子层面的具有上下文关系的特征向量表示。LR将BERT输出的特征向量作为输入得到该样本为正类的概率,最后输出概率最大的前K个合作者信息。通过与基于Network Embedding的SDNE和TSE算法的对比实验,结果表明充分考虑了研究者和研究主题间的上下文关系的BERT模型得到了更好的特征向量表示,提高了合作者推荐的准确率。

关键词:BERT模型;合作者推荐;逻辑回归模型;学术数据挖掘;Network Embedding
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2021)03-0045-07
doi:10.3969/j.issn.1673-629X.2021.03.008

Research on BERT-based Academic Collaborator Recommendation

ZHOU Yi-min,HUANG Jun
(School of Optical-electrical & Computer Engineering,University of Shanghai for Science & Technology,Shanghai 200093,China)

Abstract:Academic collaborator recommendation is an effective application of academic big data. However,existing methods ignore the contextual relationship between academic researchers and research topics,therefore they cannot recommend suitable collaborators. We propose the BERT-based collaborator recommendation (BACR),which aims to recommend high-potential collaborators to meet the requirements of researchers. To this end,we design a new recommendation framework,which consists of two basic components:BERT (bidirectional encoder representations from transformers) pre-trained language model and logistic regression model (LR). In particular,BERT jointly represents the researcher and the research topic to obtain a context-dependent feature vector representation on sentence level. LR takes the feature vector output by BERT as input to obtain the probability that the sample is positive,and finally outputs the information of the top K collaborators with the largest probability. The comparative experiments with Network Embedding-based SDNE and TSE algorithms show that the BERT model that fully takes into account the contextual relationship between the researcher and the research topic gets a better feature vector representation,which improves the accuracy of collaborator recommendation.

Key words:BERT model;collaborator recommendation;logistic regression model;academic data mining;Network Embedding

0 引 言

当今,已经形成一些学术搜索引擎,例如微软学术搜索、谷歌学术搜索和AMiner等,这使得探索诸如科学文献和研究者概况这类海量的数字学术资料更方便。数据量和种类的快速增长需要更先进的工具来帮助学术数据的研究,并且已经付出巨大的努力来开发各种高效的应用。长期以来,学术合作者推荐被认为是开发学术数据的一种有效应用,其目的是为给定的研究者找到潜在的合作者。在过去的几年中,已经提出了一些方法[1-3]来解决这样的问题。尽管取得了进步,但现有的技术只能推荐不考虑上下文关系的合作者。例如,目前的工作不能推荐合适的候选人给一个不仅包含“机器学习[4]”而且包含“推荐算法”主题的研究者。一般来说,在寻找合作者之前,研究者会先确定他要研究的主题。因此,有必要使用上下文关系为学术合作者提供推荐。

收稿日期:2020-03-31 **修回日期**:2020-07-31
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,副教授,硕导,硕士,研究方向为嵌入式系统、计算机系统结构、网络应用与智能设备等;通信作者:黄 俊(1995-),男,硕士,研究方向为推荐算法、数据挖掘。

1 相关工作

1.1 问题定义

研究者:一名研究者与他发表的文献有关,这些文献揭示了他的研究兴趣和与他人的学术合作。

研究主题:研究主题是从特定文段(如标题或关键词列表中的一个词)中提取的关键词或短语;同时,一个文献包含一个或多个研究主题,这些研究主题共同反映了其潜在的范畴。

上下文:合作上下文指的是研究人员在其合作文献中共同研究的主题集。

基于以上的初步研究,对基于 BERT 的学术合作者推荐研究的定义如下:

BACR 给定研究者 r_0 和主题 T_0,从所有候选者 R 中找到合作者 r,这些合作者将在 T_0 上与 r_0 一起工作,具有最高的可靠性。

1.2 数据预处理

该文所研究的学术合作者推荐主要用到两组数据:研究者和研究主题。研究者每篇学术文献都已给出,需要进行预处理的数据是研究主题,主要使用以下两个方法:

(1)词干提取:去除词缀得到词根的过程(得到单词最一般的写法)。对于一个词的形态词根,词干不需要完全相同;相关的词映射到同一个词干一般就能得到满意的结果,即使该词干不是词的有效根。

(2)停用词去除:因为在文献的标题和摘要中通常会有一些高频但无实际意义的词,如:“this”,“of”,“is”,“at”等,该文将此类词语加入停用词表过滤掉。

基于以上两个方法,从文献的标题和摘要中获取到一些词组配合文献已有的关键词生成真正的关键词组。

2 BERT 模型

BERT[5],即是 bidirectional encoder representations from transformers,顾名思义,BERT 模型重要部分是基于双向的 Transformer 编码器来实现的,其模型结构如图 1 所示。

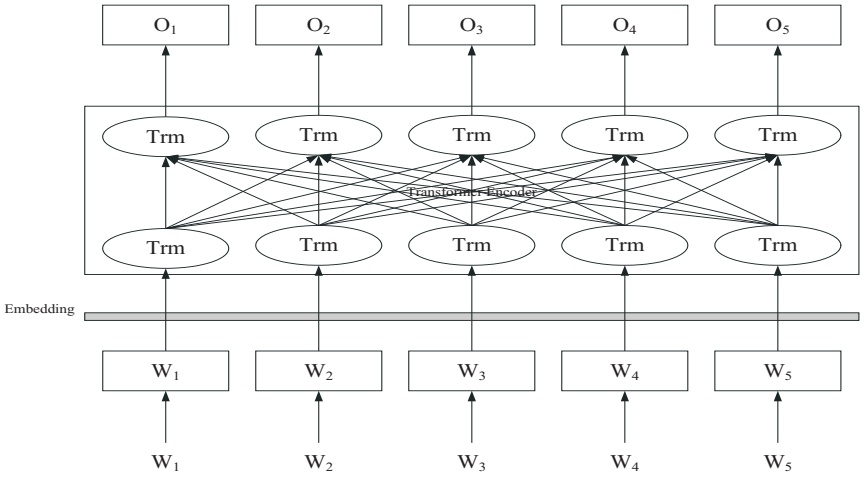


图 1 BERT 模型结构

图 1 中的 $w_1, w_2, \cdots, w_5$ 表示字的文本输入,经过双向的 Transformer 编码器,就可以得到文本的向量化表示,即文本的向量化表示主要是通过 Transformer 编码器实现的。Transformer 是由文献[6]提出,是一个基于 Self-attention 的 Seq2seq 模型,也就是 Encoder 将一个可变长度的输入序列变成固定长度的向量,而 Decoder 将这个固定长度的向量解码成为可变长度的输出序列。通常 Seq2seq 模型中使用 RNN 来实现 Encoder-Decoder 的序列转换,但是 RNN 存在无法并行、运行慢的缺点,为了改进它的不足,Transformer 使用 Self-attention 来替代 RNN。Transformer 模型的 encoder 结构如图 2 所示。

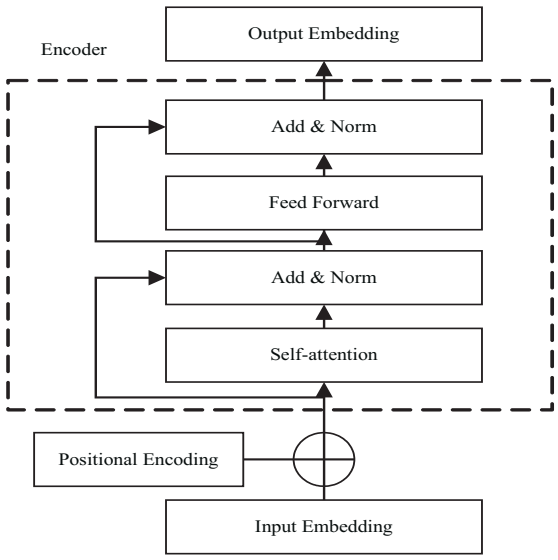


图 2 Transformer Encoder 结构

从图 2 中可以看出,Encoder 的输入是一句话的字嵌入表示,并且加上该句话中每个字的位置信息,再经过 Self-attention 层,使 Encoder 在编码每个字的时候

可以查看该字的前后字的信息。它的输出会经过一层 Add & Norm 层,Add 表示将 Self-attention 层的输入和输出进行相加,Norm 表示将相加过的输出进行归一化处理,使得 Self-attention 层的输出有固定的均值和标准差,其中均值为 0,标准差为 1。归一化后的向量列表再传入一层全连接的前馈神经网络,同样的,Feed Forward 层也会由相应的 Add & Norm 层处理,然后输出全新的归一化后的词向量列表。

图 1 中的 Embedding 包含三个嵌入层分别是 Token Embeddings、Segment Embeddings 和 Position Embeddings,如图 3 所示。

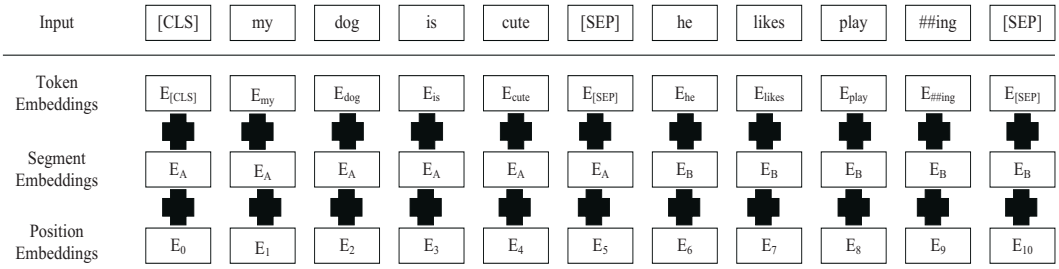


图 3　BERT 的输入表示

Token Embeddings:Token Embeddings 层是要将各个词转换成固定维度的向量。输入文本在送入 Token Embeddings 层之前要先进行 tokenization 处理。此外,两个特殊的 token 会被插入到 tokenization 的结果的开头([CLS])和结尾([SEP])。

Segment Embeddings:Segment Embeddings 层标记输入的句子对的每个句子,只有两种向量表示。前一个向量是把 0 赋给第一个句子中的各个 token,后一个向量是把 1 赋给第二个句子中的各个 token。如果输入仅仅只有一个句子,那么它的 segment embedding 就是全 0。

Position Embeddings:Position Embeddings 层标识序列的顺序信息,最长序列长度为 512。Position Embeddings layer 实际上就是一个 lookup 表,表的第一行代表第一个序列的第一个位置,第二行代表序列的第二个位置,以此类推。

BERT 模型使用两个新的无监督预测任务来对 BERT 进行预训练,分别是 Masked LM 和 Next Sentence Prediction:

MLM:给定一句话,随机掩盖部分输入词,然后根据剩余的词对那些被掩盖的词进行预测。这个任务在业界被称为 Cloze task(完型填空任务),它是为了让 BERT 模型能够实现深度的双向表示,不仅需要某个词左侧的语言信息,也需要它右侧的语言信息,具体做法是:针对训练样本中的每个句子随机抹去其中 15% 的词汇用于预测,例如:“加油武汉,加油中国”,被抹去的词是“中”,对于被抹去的词,进一步采取以下策略:(1)80% 的概率真的用[MASK]去替代被抹去的词:“加油武汉加油中国”->“加油武汉,加油[MASK]国”;(2)10% 的概率用一个随机词去替代它:“加油武汉,加油中国”->“加油武汉,加油大国”;(3)10% 的概率保持不变:“加油武汉,加油中国”->“加油武汉,加油中国”。这样做的主要原因是:在后续微调任务中语句中并不会出现[MASK]标记,若总是使用[MASK]来替代被抹去的词,就会导致模型的预训练与后续的微调不一致。这样做的优点是:采用上面的策略后,Transformer encoder 就不知道会让它预测哪个单词,换言之它不知道哪个单词会被随机单词给替换掉,那么它就不得不保持每个输入 token 的一个上下文的表征分布。也就是说如果模型学习到了要预测的单词是什么,那么就会丢失对上下文信息的学习,而如果模型训练过程中无法学习到哪个单词会被预测,那么就必须通过学习上下文的信息来判断出需要预测的单词,这样的模型才具有对句子的特征表示能力。另外,由于随机替换相对句子中所有 tokens 的发生概率只有 1.5%(即 15% 的 10%),所以并不会影响到模型的语言理解能力。

NSP:给定一篇文章中的两句话,判断第二句话在文章中是否紧跟在第一句话之后。许多重要的自然语言处理下游任务,如问答(QA)和自然语言推理(NLI)都是基于理解两个句子之间的关系,因此这个任务是为了让 BERT 模型学习到两个句子之间的关系。具体做法是:从文本语料库中随机选择 50% 正确语句对和 50% 错误语句对,即若选择 A 和 B 作为训练样本时,B 有 50% 的概率是 A 的下一个句子(标记为 IsNext),也有 50% 的概率是来自语料库中随机选择的句子(标记为 NotNext),本质上是在训练一个二分类模型,判断句子之间的正确关系。在实际训练中,NSP 任务与 MLM 任务相结合,让模型能够更准确地刻画语句乃至篇章层面的语义信息。

BERT 模型的输出有两种形式,一种是字符级别的向量,即输入短文本的每个字符对应的有一个向量表示;另外一种是句子级别的向量,即 BERT 模型输出最左边[CLS]特殊符号的向量,它认为这个向量可以

代表整个句子的语义,如图 4 所示。

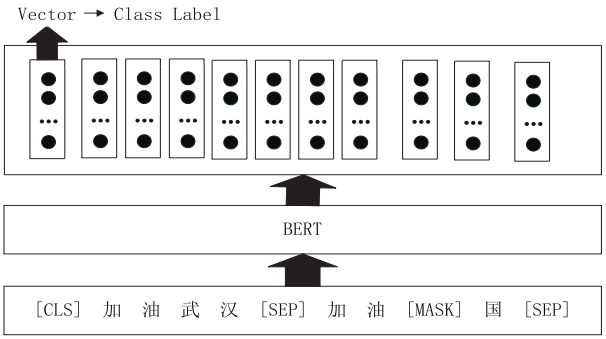


图 4 BERT 模型输出

图 4 中,最底端中的[CLS]和[SEP]是 BERT 模型自动添加的句子开头和结尾的表示符号,可以看到输入字符串中每个字符经过 BERT 模型后都有相应的向量表示,当想要得到一个句子的向量时,BERT 模型输出最左边[CLS]特殊符号的向量,该文应用的就是 BERT 模型的这种输出。

3 逻辑回归模型

经过上节的处理后,有了研究者和研究主题的向量表示,该文要做的是推荐学术合作者,故此巧妙设置二分类判断输入样本是正类的概率,输出此概率,最后按照概率的大小做出推荐。在此引入逻辑回归模型[7](logistic regression),它属于广义线性模型。

假设有训练样本集$\{(x^1,y^1),(x^2,y^2),\cdots,(x^m,y^m)\}$,其中$x^i \in R^n$,表示第 i 个训练样本对应的某篇学术文献里的研究者研究主题向量,维度为 n,共 m 个训练样本,$y^i \in \{0,1\}$ 表示第 i 个训练样本是否是正类。假设预测函数为:

$$h_\theta(x) = g(\theta^{\mathrm{T}}x) \tag{1}$$

其中,x 表示特征向量,g 表示一个常用的 Logistic 函数(Sigmoid 函数):

$$g(z) = 1/(1 + e^{(-z)}) \tag{2}$$

其中,e 是欧拉常数,z 表示曲线陡度。

结合以上两式,构造的预测函数为:

$$h_\theta(x) = g(\theta^{\mathrm{T}}x) = 1/(1 + e^{(-\theta^{\mathrm{T}}x)}) \tag{3}$$

由于 $g(z)$ 函数的特性,它输出的结果不是预测结果,而是一个预测为正类的概率的值,预测为负例的概率就是 $1-g(z)$,函数表示形式如下:

$$h_\theta(x) = \begin{cases} P(y=0 \mid x,\theta) = 1 - g(\theta^{\mathrm{T}}x) \\ P(y=1 \mid x,\theta) = g(\theta^{\mathrm{T}}x) \end{cases} \tag{4}$$

由式(4)可知,$h_\theta(x)$ 预测正确的概率为:

$$P(\text{正确}) = ((g(x^i,\theta))^{(y^i)} * (1 - g(x^i,\theta))^{(1-y^i)} \tag{5}$$

其中,y^i 为某一条样本的预测值,取值范围为 0 或者 1。一般进行到这里就应该选择判别的阈值,由于该文是做出推荐,实际上是输出正类概率,最后筛选出前 k 个即为推荐的合作者,故不需要设定阈值。

此时想要找到一组 θ,使预测出的结果全部正确的概率最大,而根据最大似然估计[8],就是所有样本预测正确的概率相乘得到的 P(正确)最大,似然函数如下:

$$L(\theta) = \prod_{i=1}^{m} (h_\theta(x^i))^{(y^i)} (1 - h_\theta(x^i))^{(1-y^i)} \tag{6}$$

上述似然函数最大时,公式中的 θ 就是所要的最好的 θ。由于连乘函数不好计算,因此对公式两边求对数得到对数似然函数:

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{m} [y^i \log h_\theta(x^i) + (1 - y^i)\log(1 - h_\theta(x^i))] \tag{7}$$

得到的这个函数越大,证明得到的 θ 越好,所以对求 $l(\theta)$ 的最大值来求得参数 θ 的值,由于在函数最优化的时候习惯让一个函数越小越好,故此将式(7)做了以下改变得到逻辑回归的代价函数:

$$J(\theta) = -\frac{1}{m}\sum_{i-1}^{m} [y^i \log h_\theta(x^i) + (1 - y^i)\log(1 - h_\theta(x^i))] \tag{8}$$

对于以上所求得的代价函数,采用梯度下降的方法来求得最优参数 θ。梯度下降过程如下:

Repeat{

$$\theta_j := \theta_j - \alpha \frac{\Delta J(\theta)}{\Delta \theta_j}$$

}

其中 α 为学习率,也就是每一次的步长;$\frac{\Delta J(\theta)}{\Delta \theta_j}$ $(j=1,2,\cdots,n)$ 为梯度。接下来对梯度进行求解也即是对代价函数求偏导:

$$\frac{\theta_j(\theta)}{\partial \theta_j} = -\frac{1}{m}\sum_{i-1}^{m} \frac{\partial}{\partial \theta_j}[y^i \log h_\theta(x^i) + (1 - y^i)\log(1 - h_\theta(x^i))] \tag{9}$$

其中:

$$\frac{\partial}{\partial \theta_j}[y^i \log h_\theta(x^i) + (1 - y^i)\log(1 - h_\theta(x^i))] =$$

$$\frac{\partial}{\partial \theta_j}[y^i \log h_\theta(x^i) + (1 - y^i)\log(1 - g(\theta^{\mathrm{T}}x^i))] =$$

$$y^i \cdot \frac{1}{g(\theta^{\mathrm{T}}x^i)} \cdot \frac{\partial}{\partial \theta_j} g(\theta^{\mathrm{T}}x^i) + (1 - y^i) \cdot \frac{-1}{1 - g(\theta^{\mathrm{T}}x^i)} \cdot \frac{\partial}{\partial \theta_j} g(\theta^{\mathrm{T}}x^i) =$$

$$\left[y^i \cdot \frac{1}{g(\theta^{\mathrm{T}}x^i)} - (1 - y^i) \cdot \frac{1}{1 - g(\theta^{\mathrm{T}}x^i)}\right] \cdot \frac{\partial}{\partial \theta_j} g(\theta^{\mathrm{T}}x^i)$$

而又因为:

$$\frac{\partial}{\partial z}g(z)=g(z)\cdot(1-g(z))$$

则:

$$\frac{\partial}{\partial\theta_j}g(\theta^{\mathrm{T}}x^i)=g(\theta^{\mathrm{T}}x^i)\cdot(1-g(\theta^{\mathrm{T}}x^i))\cdot\frac{\partial}{\partial\theta_j}(\theta^{\mathrm{T}}x^i)$$

因此:

$$\frac{\partial}{\partial\theta_j}[y^i\log h_\theta(x^i)+(1-y^i)\log(1-g(\theta^{\mathrm{T}}x^i))]=$$

$$[y^i\frac{1}{g(\theta^{\mathrm{T}}x^i)}-(1-y^i)\cdot\frac{1}{1-g(\theta^{\mathrm{T}}x^i)}]\cdot g(\theta^{\mathrm{T}}x^i)\cdot$$

$$(1-g(\theta^{\mathrm{T}}x^i))\cdot\frac{\partial}{\partial\theta_j}(\theta^{\mathrm{T}}x^i)=$$

$$[y^i\cdot(1-g(\theta^{\mathrm{T}}x^i))-(1-y^i)\cdot g(\theta^{\mathrm{T}}x^i)]\cdot$$

$$\frac{\partial}{\partial\theta_j}(\theta^{\mathrm{T}}x^i)=[y^i-g(\theta^{\mathrm{T}}x^i)]\cdot x_j^i$$

故:

$$\frac{\theta}{\partial\theta_j^i}j(\theta)=-\frac{1}{m}\sum_{i=1}^{m}[y^i-g(\theta^{\mathrm{T}}x^i)]\cdot x_j^i=$$

$$\frac{1}{m}\sum_{i=1}^{m}[h_\theta(x^i)-y^i]\cdot x_j^i$$

由以上分析可以得到梯度下降过程如下:

Repeat{

$$\theta_j:=\theta_j-\alpha\frac{1}{m}\sum_{i=1}^{m}[h_\theta(x^i)-y^i]\cdot x_j^i$$

}

其中,$i=1,2,\cdots,m$ 表示样本数,$j=1,2,\cdots,n$ 表示特征数。由此方法求得 θ,得到预测函数 $h_\theta(x)$,即可对新输入的数据输出为正类的概率。

4　基于 BERT 的学术合作者推荐算法

综上 2,3,该文提出基于 BERT 的学术合作者推荐算法,其具体流程可以描述如下:

算法 1:基于 BERT 的学术合作者推荐算法。

输入:初始研究者研究主题训练集 $T=\{(x^1,y^1),(x^2,y^2),\cdots,(x^n,y^n)\}$,其中 x^i 为每条研究者研究主题,y^i 表示每条训练样本是否为正类,$i=1,2,\cdots,N$;

输出:学术合作者推荐模型 M。

步骤 1:使用第 1 节中的方法对训练集 T 进行预处理,得到预处理后的训练集 $T=\{(x^1,y^1),(x^2,y^2),\cdots,(x^n,y^n)\}$,其中 x^i 为预处理后的每条研究者研究主题,y^i 表示预处理后的每条训练样本是否为正类,$i=1,2,\cdots,N$;

步骤 2:使用第 2 节中介绍的 BERT 预处理语言模型在训练集 T 上进行微调,采用如图 4 所示的 BERT 模型输出,得到训练集 T 对应的特征表示为 $V=(v^1,v^2,\cdots,v^N)$,其中 v^i 是每条研究者研究主题 x^i 对应的句子级别的特征向量,$i=1,2,\cdots,N$;

步骤 3:将步骤 2 中得到的特征表示 V 输入第 3 节中介绍的逻辑回归模型进行训练,输出学术合作者推荐模型 M。

5　实验与评价

5.1　实验设置

5.1.1　数　据

在该实验中,采用在文献[9]中的 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。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。

5.1.2　评价目标

针对以下目标进行实验:

主题受限的合作者推荐:评价 BERT 在确定特定主题下推荐合作者的有效性。该文是基于二分类做出的推荐,分类问题最常用的评价指标包括精确率 P、召回率 R 以及 F1 值,它们的计算需要用到混淆矩阵,混淆矩阵[10]如表 1 所示。

表 1　分类结果的混淆矩阵

	0	1
0	预测 negative 正确 TN	预测 positive 错误 FP
1	预测 negative 错误 FN	预测 positive 正确 TP

其中行代表真实值;列代表预测值;0 表示 negative;1 表示 positive。如果预测的是 0,真实值是 0,就为 TN;如果预测为 1,真实值为 0,就为 FP;预测为 0,真实值为 1,就为 FN,预测为 1,真实值为 1,就为 TP。

(1)精确率 P 是指分类器预测为正类且预测正确的样本占所有预测为正类的样本的比例,计算公式如下:

$$P=\frac{\mathrm{TP}}{\mathrm{TP+FP}}\tag{10}$$

(2)召回率 R 是指分类器预测为正类且预测正确的样本占所有真实为正类的样本的比例,计算公式如下:

$$R=\frac{\mathrm{TP}}{\mathrm{TP+FN}}\tag{11}$$

(3)F1 值是兼顾 P 和 R 的一个指标,一般计算公式[11]如下:

$$\mathrm{F1}=\frac{(\alpha^2+1)P*R}{\alpha^2(P+R)}\tag{12}$$

此时 F 是 P 和 R 的加权调和平均,$\alpha>0$ 度量了 R 对 P 的相对重要性,通常取 $\alpha=1$,此时是最常见的 F1,也即:

$$F1 = \frac{2P * R}{P + R} \tag{13}$$

其中，$0 \leqslant F1 \leqslant 1$。当 $P = 1$ 且 $R = 1$ 时，F1 达到最大值为 1，此时精确率 P 和召回率 R 均达到 100%，这种情况是完美状态，而由文献[11]知实际中很难达到，因为 P 和 R 是一对矛盾的变量，当 P 较高时，R 往往会偏低；当 R 较高时，P 又往往偏低。因此，在使用 F1 值评估性能时，其值越接近 1，说明分类器的性能越好。由于 F1 值是对 P 和 R 两个评价指标的综合考虑，可以更加全面地反映性能，因此它是评价实验效果的主要评价指标。

5.1.3 评价方法

为了评估 BERT 在主题限制的情况下推荐合作者的性能，采用以下基于 Network Embedding 的具有代表性的方法进行比较。

(1)深层网络结构嵌入(SDNE)。SDNE[12]代表为编码实体及其关系的结构信息而设计的方法[13]。当应用于 BACR 问题时，研究者被视为实体，在特定主题中的合作被视为上下文关系。由于主题的组合，BACR 中实际上存在无限数量的上下文关系，因此不能直接采用像[13]这样的常规方法。

(2)特定任务嵌入(TSE)。针对作者识别问题，在文献[14]中提出了 TSE。简单来说，TSE 由三个层次构成：第一个层次用嵌入学习方法表示上下文关系的每个来源(如关键词、场所)，如文献[15-16]；在实验中，这些嵌入是为研究者和研究主题独立学习的。在第二层，对所有源的提取嵌入进行不同权重的集成；最后，在集成嵌入的前提下，第三层学习特定分类任务的模型参数。

在以上比较方法中共享嵌入维数 Dim。在该实验中，Dim 的范围为{10，20，30，**40**，50，60}。粗体表示 Dim 有效性的比较的默认值，其他用于评估参数灵敏度。而该文使用的 BERT 预训练模型是 Google 提供的 BERT-Base 模型，Transformer 层数 12 层；隐藏层 768 维；采用了 12 头模式；共有 110 M 个参数；其他的训练参数如表 2 所示。该文使用逻辑回归用于所有的比较方法，评价指标主要采用 F1 值。

表 2 BERT 模型训练参数

BERT 模型参数	BERT 模型参数取值
Dropout 随机失活率	0.1
Epoch 模型迭代轮次	2，3，4
Learning Rate 学习率	1e-4，1e-5，1e-6
Batch Size 每批训练集数据大小	24

5.2 主题受限的合作预测

5.2.1 BERT 的有效性

所有比较方法的性能总结为表 3，其中维数 Dim 设置为 20、40 和 60；而推荐人数 K 设置为 5、10、15 和 20。根据所呈现的结果，BERT 证明了其在预测主题受限的合作者方面的优势，因为所产生的 F1 明显高于其他人。

表 3 主题受限的合作者预测表现

方法	Dim-20				Dim-40				Dim-60			
	F1@K				F1@K				F1l@K			
	5	10	15	20	5	10	15	20	5	10	15	20
SDNE	0.1374	0.1736	0.2021	0.2127	0.1403	0.1811	0.2093	0.2235	0.1591	0.1866	0.2038	0.2202
TSE	0.1439	0.1867	0.2117	0.2294	0.1534	0.2093	0.2285	0.2483	0.1616	0.1921	0.2143	0.2339
BERT	0.1633	0.2137	0.2376	0.2557	0.1633	0.2137	0.2376	0.2557	0.633	0.2137	0.2376	0.2557

给定不同主题研究者倾向于与不同研究员合作，使得有必要使推荐算法考虑研究者主题依赖关系。通过对实验结果的进一步分析，可以得出以下三点：

首先，与 SDNE 相比，BERT 和 TSE 都具有更好的性能。这些方法中最显著的区别在于，BERT 和 TSE 都在训练推荐模型的同时提取上下文关系。因此，将上下文关系引入到推荐模型中是十分必要的。

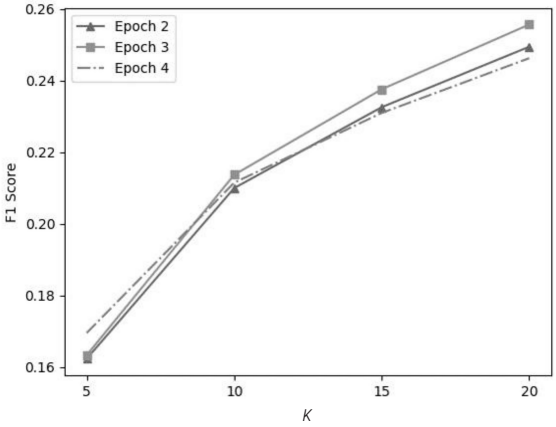
其次，通过观察，BERT 的表现甚至比 TSE 更好。与研究者和研究主题的特征是独立训练的 TSE 不同，特征训练在 BERT 中是一起训练的。通过这种方式，研究者和研究主题的上下文关系自然得以保留，从而有助于更准确的推荐。

最后，在所有的比较方法中，SDNE 的性能最低。很明显的原因是 SDNE 没有考虑到研究者和研究主题之间存在的上下文关系，不同的研究主题下，研究者倾向于与不同的研究者合作。如研究者 A 倾向于与 B 在“推荐算法”上合作而不是 C，但却倾向于与 C 在“机器学习[4]”上合作而不是 B。

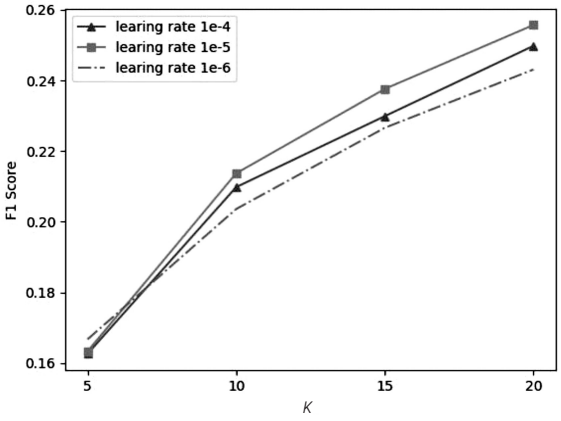
5.2.2 超参数对推荐算法的影响

图 5(a)和(b)分别演示了超参数 Epoch 和 Learning Rate 的影响。根据给出的结果，当 Epoch = 3 和 Learning Rate = 1e-5 时，获得了优越的性能。此外，很显然，Learning Rate 对 BERT 的性能影响更大，因为生成的结果因 Learning Rate 不同而差异很大。较大

的 Learning Rate 使梯度下降的速率更快,但也可能导致错过全局最优点;因此,需要通过网格搜索仔细选择适当的值。



(a)Epoch 对 F1 值的影响



(b)Learning Rate 对 F1 值的影响

图 5　超参数的影响

5.3　实验结论

综上可知,基于 BERT 的学术合作者推荐充分考虑了研究者和研究主题间的上下文关系,对比以往的方法显著提高了性能,相较于 TSE 最高提高达到了 6.45%,最低提高 2.10%;而相较于 SDNE 最高提高则高达 18.00%,最低也提高了 13.52%,这也充分展示了 BERT 的优越性能。

6　结束语

在解决学术合作者推荐的问题中,使用 BERT 模型进行研究者和研究主题的向量表示,提出了一种基于 BERT 模型的学术合作者推荐算法,并与 SDNE、TSE 两个模型进行对比,实验结果表明 BERT 模型在研究者研究主题向量的表示上能达到很好的效果,在一定程度上提升了推荐算法的准确性。

参考文献:

[1] ZHONG L I, HAN H, GUANGYIN W U, et al. Academic collaboration recommendation based on sparse distributed representation[J]. DEStech Transactions on Social Science, Education and Human Science, 2019, 2(1): 24-29.

[2] XIA F, WANG W, BEKELE T M, et al. Big scholarly data: a survey[J]. IEEE Transactions on Big Data, 2017, 3(1): 18-35.

[3] ZHOU X, DING L, LI Z, et al. Collaborator recommendation in heterogeneous bibliographic networks using random walks [J]. Information Retrieval Journal, 2017, 20(4): 317-337.

[4] MURPHY K P. Machine learning: a probabilistic perspective [M]. Massachusetts: MIT Press, 2012.

[5] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]//North American chapter of the association for computational linguistics. Minneapolis, MN, USA: [s. n.], 2019: 4171-4186.

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Neural information processing systems. Long Beach, CA, USA: [s. n.], 2017: 5998-6008.

[7] MICHAEL P LAVALLEY. Logistic regression[J]. Circulation, 2008, 117(18): 2395-2399.

[8] WHITE H. Maximum likelihood estimation of misspecified models[J]. Econometrica, 1982, 50(1): 1-25.

[9] ADOMAVICIUS G, TUZHILIN A. Context-aware recommender systems [M]//Recommender systems handbook. Boston, MA: Springer, 2011: 217-253.

[10] VISA S, RAMSAY B, RALESCU A L, et al. Confusion matrix-based feature selection[C]//Proceedings of the 22nd midwest artificial intelligence and cognitive science conference 2011. Cincinnati, Ohio, USA: [s. n.], 2011.

[11] ZHOU Zhihua. Machine learning[M]. Beijing: Tsinghua University Press, 2016: 30-32.

[12] WANG D, CUI P, ZHU W, et al. Structural deep network embedding[C]//Knowledge discovery and data mining. San Francisco, CA, USA: [s. n.], 2016: 1225-1234.

[13] BORDES A, USUNIER N, GARCIADURAN A, et al. Translating embeddings for modeling multi-relational data[C]// Neural information processing systems. Long Beach, CA, USA: [s. n.], 2013: 2787-2795.

[14] LE Q V, MIKOLOV T. Distributed representations of sentences and documents[C]//International conference on machine learning. Detroit, MI, USA: [s. n.], 2014: 1188-1196.

[15] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]//Neural information processing systems. Long Beach, CA, USA: [s. n.], 2013: 3111-3119.

[16] PEROZZI B, ALRFOU R, SKIENA S, et al. DeepWalk: online learning of social representations[C]//Knowledge discovery and data mining. New York City, USA: [s. n.], 2014: 701-710.